An International ICT R&D Journal Sponsored by ZTE Corporation

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

ZTE COMMUNICATIONS

tech.zte.com.cn

June 2017, Vol. 15 No. S1

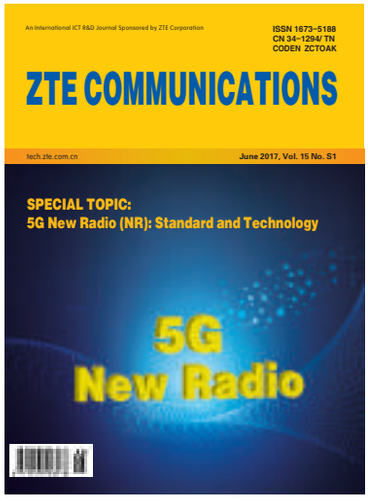
SPECIAL TOPIC:
5G New Radio (NR): Standard and Technology

ZTE中兴 ZTE COMMUNICATIONS

VOLUME 15 NUMBER S1 JUNE 2017

ZTE Communications Editorial Board

CONTENTS



Special Topic: 5G New Radio (NR): Standard and Technology

CONTENTS

Review

ZTE COMMUNICATIONS
Vol. 15 No. S1 (Issue 57)
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, 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:
June 25, 2017

Publication Licenses:

ISSN 1673-5188
CN 34-1294/ TN

Annual Subscription:
RMB 80

Introduction to *ZTE Communications*

ZTE Communications is a quarterly, peer-reviewed international technical journal (ISSN 1673-5188 and CODEN ZCTOAK) sponsored by ZTE Corporation, a major international provider of telecommunications, enterprise and consumer technology solutions for the Mobile Internet. The journal publishes original academic papers and research findings on the whole 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. *ZTE Communications* was founded in 2003 and has a readership of 5500. The English version is distributed to universities, colleges, and research institutes in more than 140 countries. It is listed in Inspec, Cambridge Scientific Abstracts (CSA), Index of Copernicus (IC), Ulrich's Periodicals Directory, Abstract Journal, Chinese Journal Fulltext Databases, Wanfang Data — Digital Periodicals, and China Science and Technology Journal Database. Each issue of *ZTE Communications* is based around a Special Topic, and past issues have attracted contributions from leading international experts in their fields.

Guest Editorial

Fa-Long Luo

5G New Radio (NR): Standard and Technology

▶ Fa-Long Luo

Dr. Fa-Long Luo is an IEEE Fellow and the Chief Scientist of Micron Technology Inc., USA. He is now also an Affiliate Full Professor of EE Department at University of Washington, Seattle, USA, the chairman of IEEE Industry DSP Standing Committee and a technical board member of IEEE Signal Processing Society. He has served as an associate editor of *IEEE Access* and *IEEE Internet of Things Journal*. He has 33 years of research and industrial experience in multimedia, communications and broadcasting with real-time implementation, applications and standardizations areas with receiving worldwide attention and recognition. Including his well-received Wiley-IEEE book "*Signal Processing for 5G*", Dr. Luo has published 5 books, more than 100 technical papers and 30 patents in these and closely related fields. He was awarded the Fellowship by the Alexander von Humboldt Foundation of Germany.

Led by 3GPP, industry and research communities are now investing tremendous efforts to develop advanced technologies in order to meet all related requirements of ITU "IMT 2020 and Beyond" for 5G wireless communications in terms of enhanced mobile broadband (eMBB), massive machine-type communications (mMTC) and ultra-reliable and low latency communications (URLLC). 3GPP has defined a new name for its planned standard proposal of 5G as "New Radio" (NR), which will be submitted to ITU for the official international 5G standard. The related technologies for 5G NR can mainly be categorized into four areas, namely, 1) new modulation and coding algorithms including multi-user superposition and shared access, enhanced waveform generation and advanced error-correction coding; 2) new system and network architectures including network slicing, device to device (D2D) communications, cloud radio access network (C-RAN), and ultra-dense network (UDN); 3) new spatial-domain processing such as massive multiple-input multiple-output (massive-MIMO), adaptive 3D beamforming and multi-antenna diversity; 4) new spectrum opportunities including millimeter-wave band and license assisted access (LAA).

According to the 3GPP plan, the 5G NR standard has two phases. Phase 1 (Release 15) will be completed in 2018, mainly addressing a more urgent subset of the commercial needs, which could make it possible to deploy the first 5G network by around 2020. Phase 2 (Release 16) will be completed for the IMT-2020 submission in March of 2020 by addressing all identified use cases and requirements. More importantly, 5G NR design should be forward compatible at its core so that features can be added in later releases in an optimal way. It can be seen that the development of 5G NR standard in each stage will be very challenging tasks and require the huge efforts of the related industry, research, alliances and regulatory authorities so as to bring the success in a high quality and a timely manner.

From practice, technology and standard points of view, this special issue aims to provide a unique and timely forum to address all technology issues related to 5G NR standardization. The call-for-papers of this special issue has brought excellent submissions in both quality and quantity. After two-round reviews, seven excellent papers have been selected for publication in this special issue which is organized into the following two category groups.

Consisting of four papers, the first group of this special issue mainly addresses technology aspects of the physical layer for 5G NR by covering new modulation, channel coding, waveform generation, multi-user superposition and shared access, multiple antennas, frame structures, numerology, hybrid automatic repeat request (HARQ) and duplex. As its title "5G New Radio: Physical Layer Overview" exactly means, the first paper presents a comprehensive introduction to all the above key components of the physical layer of 5G NR by addressing basic principles, mathematical models, step-by-step algorithms, implementation complexities, schematic processing flow and the corresponding application scenarios. Moreover, a more detailed roadmap and timeline of the 3GPP 5G NR standard development is presented in this paper.

Guest Editorial

Fa-Long Luo

Orthogonal frequency division multiplexing (OFDM) based waveform generation has become the dominant technology in many existing wireless and broadcasting standards and is actually also considered as one of key technologies in 5G radio access network (RAN). Titled as "Enhanced OFDM for 5G RAN", the second paper provides an overview on the various improved versions of OFDM for 5G NR in terms of waveform characteristics and parameters, out of band leakage, peak to average power, cyclic prefix, pilot signals, inter-carrier interference (ICI), inter-symbol interference (ISI), multipath distortion, the orthogonality and the related effect of frequency offset and phase noises, synchronization requirement in both the time domain and frequency domain, latency, mobility, signaling, compatibility, co-existence and integration with other processing such as massive MIMO.

The title of the third paper is "An Overview of Non-Orthogonal Multiple Access". Being considered as a novel and promising multiple access scheme, NOMA has been proposed and proved by NTT DOCOMO to be able to contribute to a significant improvement of the compromise among system capacity, spectrum efficiency and user fairness by introducing power-domain user multiplexing on the basis of channel difference among users. This paper provides an excellent overview on various technical aspects of NOMA by covering concept, design, performance, combination with MIMO and comparisons over orthogonal multiple access. More importantly, the status and open issue of 5G NR standardization related to NOMA in 3GPP are provided in this paper with emphasis on the multi-user superposition transmission (MUST).

As pointed out in the fourth paper titled as "Uplink Multiple Access Schemes for 5G: A Survey", signal transmitter and receiver are jointly optimized in a non-orthogonal multiple access (NMA) system, which suggests that multiple layers of data from more than one user can be simultaneously delivered in the same resource. As a matter of fact, uplink NMA is becoming a key candidate technology for 5G NR standard and a number of uplink NMA schemes from different companies have been proposed in recent 3GPP meetings. This paper reviews the key features, characteristics and standardization process status of these proposed NMA systems for 3GPP 5G NR by classifying them into three major categories in terms of their basic technique principles, namely: scrambling based NMA schemes, interleaving based NMA schemes and spreading based NMA schemes.

Organized into the second group, the fifth paper through the seventh paper included in this special issue deal with technical aspects related to massive MIMO in different frequency bands, D2D based cooperative relaying and a practical use of URLLC, respectively. "Massive MIMO 5G Cellular Networks: mm-Wave vs. μ-Wave Frequencies" is the fifth paper, which provides a comprehensive comparison between massive MIMO systems at mm-waves and at μ-waves (frequencies in the sub-6 GHz range) due to different propagation mechanisms in urban scenarios. All key differences between mm-waves and μ-waves are given in this paper in terms of channel modeling, antenna size, beam-steering , channel matrix rank, channel estimation, pre-coding design, pilot contamination and antenna diversity. It is believed that the comparisons and analyses given in this paper are very useful to the corresponding 3GPP working groups in addressing the massive MIMO part of the 5G NR standard.

"Novel MAC Layer Proposal for URLLC in Industrial Wireless Sensor Networks" is the title of the sixth paper. Taking industrial wireless sensor network (IWSN) as a representative case of the URLLC, this paper proposes a new hybrid multi-channel scheme so as to deliver a better performance in terms of enhanced throughput, increased reliability and reduced latencies. With the utilization of the multiple frequency channels, the proposed scheme defines a special purpose frequency channel that facilitates failed communications by retransmissions where the retransmission slots are allocated according to the priority level of failed communications of different nodes. The presented results show the accurateness of the theoretical analyses and the effectiveness of the proposed scheme.

Featured into the "Review" section of this special issue, the seventh paper is titled as "Device-to-Device Based Cooperative Relaying for 5G Network: A Comparative Review". As the author of this paper pointed out, the concept of cooperative communications opens a possibility of using multiple terminals to cooperatively achieve spatial diversity and hence the utilization of D2D-based cooperative relaying is promising in the era of 5G. By first proposing a new opportunistic space-time coding scheme, this paper then presents a comparative study on several cooperative multi-relay schemes in the presence of imperfect channel state information. The numerical results prove that the proposed scheme is the best up-to-date cooperative solution from the perspective of multiplexing-diversity trade-off.

As we conclude the introduction to this special issue and the contents of seven papers, we would like to thank all authors for their valuable contributions. We also express our sincere gratitude to all the reviewers for their timely and insightful comments on all submitted papers. It is hoped that the contents in this special issue are informative and useful from various aspects related to 5G NR standardization and technology development.

DOI: 10.3969/j. issn. 1673-5188. 2017. S1. 001
http://kns.cnki.net/kcms/detail/34.1294.TN.20170614.0712.002.html, published online June 14, 2017

5G New Radio: Physical Layer Overview

YUAN Yifei and WANG Xinhui
(ZTE Corporation, Shenzhen 518057, China)

Abstract

This paper provides an overview of the physical layer of 5G new radio (NR) system. A general framework of 5G NR is first described, which glues together various key components, all of them helping to fulfill the requirements of three major deployment scenarios: enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC) and massive machine type communications (mMTC). Then, several key components of the 5G NR physical layer are discussed in more detail that include multiple access, channel coding, multiple antennas, frame structures, and initial access. The two-phase approach of NR is also discussed and the key technologies expected to be specified in Phase 1 and Phase 2 are listed.

Keywords

5G; IMT-2020; new radio (NR)

1 Introduction

After years of research on 5G by various promotion groups around the world, such as METIS, IMT-2020, and 5G Forum, 5G finally arrived in 3GPP, the most influential standard development body on cellular communications in the world. In September 2015, the first 5G workshop was held in Phoenix of USA, which marks the start of 5G in 3GPP. The channel model work for 5G was then started, suitable for spectrum bands up to 100 GHz. This is the foundation of technology study of 5G. At the same time, 3GPP kicked off the study on deployment scenarios for radio access network (RAN) and key performance indicators (KPIs) of 5G. This study provides the guidance on potential technologies that may fulfill the performance requirements and applications expected for 5G. In January 2016, the actual work on the 5G channel model was moved to 3GPP WG1 whose mission is to specify the physical layer features of 5G, LTE, and Universal Mobile Telecommunications System (UMTS)/high-speed packet access (HSPA) system. Significant progress was achieved over the three working group level meetings till March 2016 and a stochastic model was chosen as mandatory and a hybrid model as optional. Both the models were captured in the technical report for channel modeling [1]. By March 2016, most of the work on requirements [2] had been finished at the RAN level. The three most important deployment scenarios for the requirements were identified: enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC) and massive machine type communications (mMTC). Performance metrics were also agreed, which include peak data rates, user experienced throughput, cell average and edge spectral efficiency, control and user plane latency, reliability, connection density and efficiency, and etc. The actual numbers of KPIs for 5G are still in discussion in 3GPP.

In March 2016, a study item was approved by 3GPP to investigate the potential technologies for 5G new radio (NR) [3]. The reason of naming it "new radio" is that this air interface will not be backward compatible with any 4G standards such as LTE-Advanced-pro. This new radio interface can be deployed stand-alone, i.e., an independent network not relying on other networks. NR can also be deployed in non-stand-alone fashion, i.e., along with LTE-Advanced-pro, since LTE evolution is also considered part of 5G. At the physical layer, potential technical areas of NR include multiple access, waveform, channel coding, frame structure, multiple-input multiple-output (MIMO), device-to-device (D2D), vehicle-to-vehicle (V2V), unlicensed spectrum, etc. This study item (Phase 1) of NR will be finished by March 2017, followed by the work item stage (Phase 2). Phase 1 aims to specify basic functionalities of NR that can partially fulfill the performance requirements of 5G. Its specification work will be completed by June 2018 and this would make it possible to deploy the first 5G network by around 2020. It became clear that not all the technologies can be specified in Phase 1, because some of features are not considered as so urgent until 2020. The RAN plenary decided in September 2016 to postpone the specification of some technologies or deployment scenarios, such as D2D, V2V, carrier frequency over 40 GHz, unlicensed spectrum, and mMTC, so that

the work groups of RAN and Services & System Aspects (SA) could focus on the specification of frame structure, channel coding, MIMO, etc. in Phase 1.

The rest of this paper is organized as follows. In section 2, we discuss the framework of NR physical layer. In section 3, key components of the NR physical layer are described. The phased approach of NR specification is discussed in Section 4. Some conclusions are drawn in Section 5.

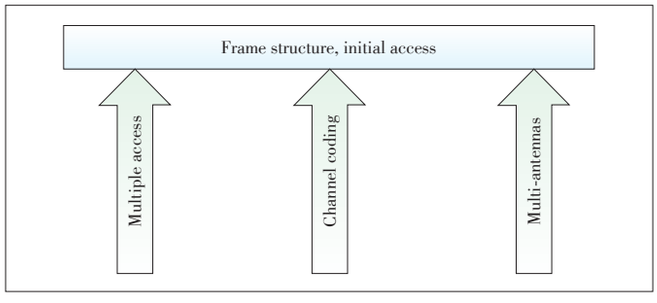
2 Framework of NR Physical Layer Technologies

NR is a new air interface that is not backward compatible with LTE, LTE-Advanced or LTE-Advanced-pro. NR networks can be independently deployed without relying on any 2G, 3G or 4G networks. This means that NR should have a complete set of RAN functionalities to be able to work alone. Similar to previous generations of cellular networks, NR should have basic frame structure, numerology, initial access procedures, and scheduling for operation. In some sense, these basic components for 5G NR system may inherit lots of design of previous generations, in particular 4G standards. It is apparent that multiple access schemes for NR would be at least based on orthogonal frequency division multiple access (OFDMA). Moreover, the fundamental waveform of NR is cyclic prefix-orthogonal frequency division modulation (CP-OFDM), while the frame structure and numerology would share some characteristics of LTE.

Innovations in areas of multiple access, channel coding and MIMO will play important roles in achieving 5G performance requirements. They serve as the main technical drivers to propel the work on basic functionalities listed above such as numerology, frame structures, initial access, and scheduling.

Fig. 1 shows the framework of NR physical layer technologies. To be more specific, new multiple access schemes such as those based on non-orthogonal properties would introduce a "scheduling-light" and/or "light initial access" mechanism to significantly reduce the control overhead and access latency in order to efficiently support mMTC.

New channel coding schemes such as low-density parity check (LDPC) and polar codes can significantly reduce the decoding complexity for scenarios with large block sizes and/or high coding rates, and improve the performance for those with short block sizes. It is known that the decoding of channel codes would consume a lot of processing power of a receiver. Codes with efficient decoding algorithms can shrink the timing budget for processing at the receiver. With these advancements, the frame structure can be made more "self-contained". Fast decoding also facilitates the design and specification for URLLC. With the super-wide bandwidth expected for 5G, i.e., up to 1 GHz, channel codes efficient for large block sizes become an imperative, rather than an option. Without it, frame structure of NR would be merely a hollow place holder.



▲**Figure 1. Framework of NR physical layer technologies.**

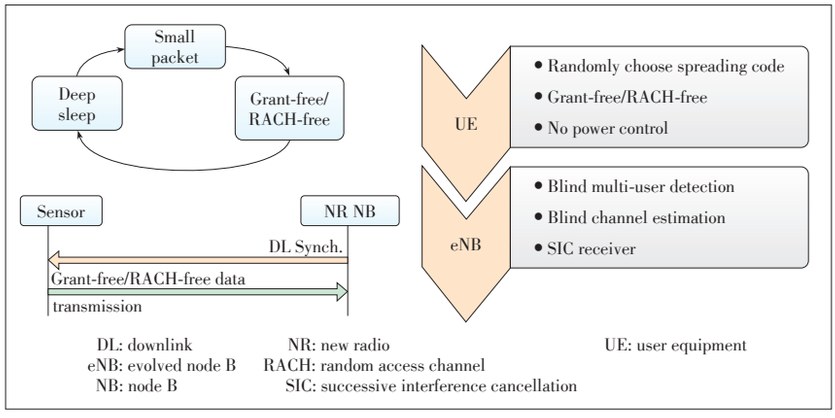
MIMO techniques help to differentiate NR from 4G in many aspects. NR would operate up to 100 GHz bands. At such high carrier frequency, MIMO or beamforming is a must-have feature. Otherwise, the severe path loss and penetration would render NR useless, even indoor. Hence, beamformed transmission would widely be employed, not only for traffic channels, but also for control signaling, random access signal, synchronization signal, and broadcast channels carrying system information. Beamforming is also used at the receiver. Because of these, initial access and frame structure of NR would have many footprint of beamforming, which is seldom seen in LTE.

3 Key Components of NR Physical Layer

3.1 Multiple Access and Waveform

Multiple access has become a landmark technology for each generation of cellular communications; 1G is frequency division multiple access (FDMA), 2G is time division multiple access (TDMA), 3G is code division multiple access (CDMA), and 4G is OFDMA. The extensive study so far has shown that OFDMA can deliver reasonably high system throughput for eMBB both for downlink and uplink. It is also well recognized that orthogonal transmission in general requires less sophisticated receivers. Therefore, OFDMA is mandatory at least for eMBB. On the other hand, non-orthogonal multiple access will be an important complementary feature to further enhance the system throughput for eMBB services in NR, similar to multi-user superposition transmission (MUST) feature for LTE.

A more important use of non-orthogonal transmission in NR is mMTC uplink. 5G should support up to 1 million devices per square kilometer. With such dense connections, the system has to handle huge volume of users' data simultaneously. This puts significant constrain on control signaling where using traditional grant-based transmission becomes highly inefficient, simply due to the excessive overhead of signaling. To solve this issue, grant-free transmission is proposed where no explicit dynamic grant is signaled to terminal devices. Non-orthogonal multiple access is often grant-free so that multiple devices can share resources at the same time and frequency. Grant-free non-orthogonal multiple access also reduces the latency. As illustrated in **Fig. 2**, a device can immediately start the transmission when it has data to send, without the need to wait until the



▲**Figure 2. Grant-free and RACH-free transmission.**

full-blown random access is completed, nor limited to radio-resource-control (RRC)-connected state only. For mMTC services that are characterized by small packets of infrequent arrivals, significant power saving is expected at the terminal devices.

Grant-free transmission is autonomous and contention-based. Different users can use different signatures to facilitate the detection and decoding at the receiver. One option would be that each device randomly selects the signatures, resulting in certain signature collisions. The other option would be that the signatures of users are pre-defined. Multiple access signatures can be sequence, code, interleaver pattern, demodulation reference signal, preamble, etc. They can roughly be categorized as follows:

1) Type 1 signatures: short codes and sequences that have more structures
   - Type 1a: short codes and sequences that have low cross correlation or better inter-distance property
   - Type 1b: short codes and sequences that have low density property
2) Type 2: long codes, sequences, or interleaver patterns that have less structure

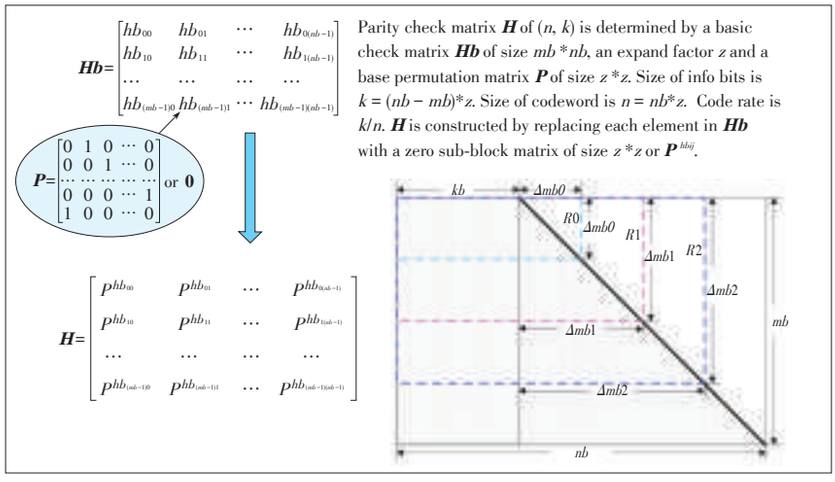
Here, long or short is of relative sense. In general, if a code or sequence length is much less than the code block length, it is considered as "short". The low cross correlation property of Type 1a signatures facilitates the successive interference cancellation (SIC) at the receiver. The low density property of Type 1b signatures helps to reduce the receiver complexity of message passing algorithm (MPA), a type of parallel interference cancellation (PIC) scheme. In 3GPP, so far about 15 grant-free schemes have been proposed. Each of them uses one or two types of signatures listed above. For example, Multi-User Shared Access (MUSA) [4] and Non-Orthogonal Coded Access (NOCA) [5] use Type 1a, Sparse Code Multiple Access (SCMA) and Pattern Defined Multiple Access (PDMA) [5] use Type 1b, and Resource Spread Multiple Access (RSMA) and Interleave Division Multiple Access (IDMA) [5] use Type 2.

Several waveforms have been proposed, such as windowed-OFDM (W-OFDM), filtered-OFDM (F-OFDM), filter-bank-OFDM (FB-OFDM), and universal OFDM (UFDM). These waveforms are able to make the spectrum more localized, i.e., less out-of-band emission, which is important to the scenarios where different numerologies are frequency division multiplexed (FDM). An extensive simulation campaign has been done, and it was decided that the study on NR waveform was to be moved to 3GPP WG4 whose mission is to set the radio frequency requirements. The rationale is that all the above mentioned waveforms are based on CP-OFDM and the receiver do not need to know the exact filtering process at the transmitter. To improve the link budget, discrete-Fourier-transform-spread OFDM (DFT-S-OFDM) is also supported for uplink with single-stream transmission, at <40 GHz.

3.2 Channel Coding and Modulation

In NR, four major types of channel coding are being studied: low density parity check (LDPC) codes, turbo codes, polar codes, and tail-biting convolutional codes (TBCC). LDPC was originally proposed by Gallager in early 1960s [6]. After experiencing over 30 years of obscurity, its potential performance and uniqueness of parallel decoding process were re-discovered by the channel coding society in late 1990s. LDPC was adopted as the optional feature in IEEE standards. However, it lost the battle to turbo codes for the channel codes of LTE, mainly because that the block size and code rate were not high enough for LDPC to be mandatory. The situation in NR becomes more favorable to LDPC: the code block size can be greater than 10,000 and the code rate can reach 8/9, so that the advantage of parallel processing of LDPC is more pronounced. Since the adoption in IEEE standards, quite a lot of studies have been carried out to make LDPC more flexible to different block sizes and to more efficiently support hybrid automatic repeat request (HARQ). There have been many implementations of LDPC in WiMAX and proprietary microwave backhaul from which plenty of experience has been accumulated in the hardware and firmware development. In another word, LDPC has been tested and now reaches certain a level of maturity in the wireless industry. Main stream LDPC schemes have the quasi-cyclic structure outlined in **Fig. 3**. Basically, all the low density parity check matrices (***H***) are derived from the proto-matrix (***Hb***), by the process of either "lifting", or "shortening". Lifting process helps to construct different block sizes of LDPC. Permutation matrix (***P***) is used during the "lifting". "Shortening" process can tune the LDPC to different code rate. The HARQ can be achieved by fetching the system bits and parity bits from circular buffer. In November RAN1 meeting of 2016, LDPC was chosen as the channel code for eMBB data channel.

Turbo codes have been widely used in 3G and 4G standards,



▲Figure 3. Basic concept of quasi-cyclic LDPC.

due to their good performance and flexibility to different block sizes. Apparently turbo codes are the most mature of all the four types mentioned above. However, the core of turbo decoding is the extrinsic information exchange between the two constituent soft-input and soft-out decoders [7]. Each runs maximum a posterior (MAP) based BCJR algorithm that is fundamentally a serial process. While windowing operation can be employed at the decoder to improve the decoder's throughput, it comes at the cost of performance degradation and hardware complexity. This issue becomes a show-stopper when the code block size reaches 6000 and the code rate is higher than 2/3 where turbo codes can no longer compete against LDPC. Certain enhancement of LTE turbo has been proposed, e.g., to extend the quadratic permutation polynomials (QPP) to around 8000. However, above that, turbo is still inferior to LDPC. Another enhancement of LTE turbo includes mother code rate < 1/3, for example, 1/5 turbo, so that the performance can be improved at low SNR operating points.

The polar code was first proposed by Erdal Arikan in 2009 [8]. The advent of polar code is seminal in the sense that it is the first channel code that constructively builds and reaches channel capacity. The original channel is binary symmetric channel (BSC) which can be considered as the simplest form of "polarized channel". This "polarization" would be extended to more sophisticated channels, so that channels are categorized into "good" channels and "bad" channels. Good channels have the maximal capacity and can carry the information bits, while bad channels have the least capacity and can carry the known bits. The basic decoding algorithm of polar is successive cancellation (SC) which is fundamentally serial. To improve the performance, list decoding is normally needed whose decoding complexity linearly increases with the list size (L). L can be 4, 8, 16 or 32. The polar code is relatively new and the most of study to date has still been in academia. Limited implementation has been conducted for polar. Nevertheless, the polar code show better performance than turbo and LDPC for short block sizes. In November RAN1 meeting of 2016, the polar code was chosen as the channel code for eMBB control channel, except for very short length.

Convolutional codes have been used for over 60 years. Normally, tail bits are reserved to bring the trellis state back to zero. In tail-biting convolutional codes (TBCC), the tail bits are saved to reduce the overhead, with a little more processing at the decoder. Even with maximal likelihood (ML) decoding such as Viterbi algorithm, the decoder of TBCC is the simplest among the four major channel codes. Due to the weak code structure, the performance of TBCC is the worst when the block size becomes larger, i.e., > 100 bits. Hence, TBCC is suitable for control channels, or small packets, low latency and low cost scenarios. Recently, list-decoding was proposed to improve the TBCC's performance by utilizing CRC bits, although the corresponding decoding complexity could be increased.

Conventional Quadrature Phase-Shift Keying (QPSK) and Quadrature Amplitude Modulation (QAM) constellation (with Gray mapping property) will be used for NR. Other low-Peak to Average Power Ratio (PAPR) modulation schemes are to be studied.

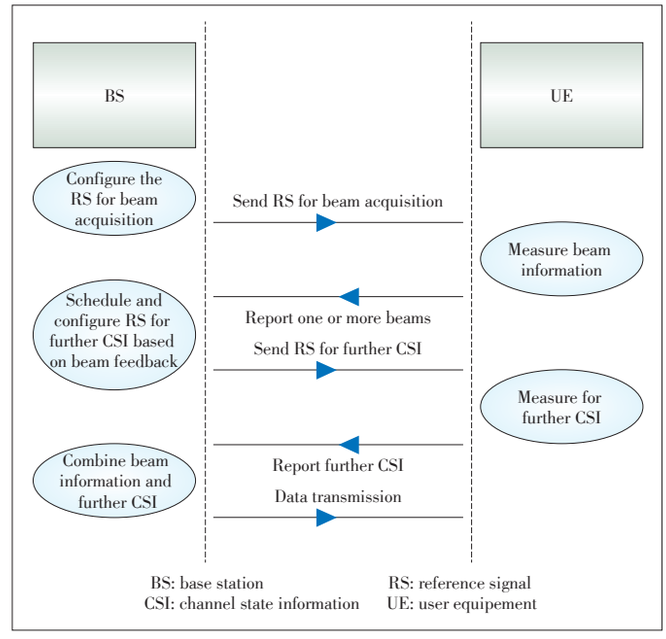
3.3 MIMO Technologies

MIMO in NR has to work for high frequency where beamforming would be widely used to compensate for the adverse propagation environment like severe path loss over the air and more attenuation when penetrating the building. Beamforming becomes more feasible for high frequency due to the shorter wavelength, and therefore the size of antenna array can be kept small even with a large number of elements. At high frequency, a high sampling rate and various imperfections at RF makes digital processing more expensive. Analog beamforming is more cost-effective. On the other hand, digital beamforming is preferred at low frequency bands due to its flexibility of resource scheduling and better performance. Hence, unified design is preferred so that common architecture is used for both low frequency bands (< 6 GHz) and high frequency bands (> 6 GHz), so that the systems can adaptively support analog/hybrid/digital beamforming, as well as the dynamic switching between various transmission schemes.

Similar to coordinated multi-point processing (CoMP) in LTE, group-based transmit/receive (Tx/Rx) beam selection can be considered where multiple links for MIMO is supported in flexible manner. It would follow a semi-transparent architecture.

Reference signaling and control signaling can be made more flexible. For example, multi-level or multi-shot designs are potential features. More specifically, in LTE, quite a number of

transmission modes were introduced to individually optimize the MIMO performance for each deployment scenario. The switching between those transmission modes is semi-static and cannot dynamically adapt to the changing environment. In NR, it is preferred to allow dynamic adaptation of transmission schemes. The experience from LTE MIMO study and specification has proved that channel state information (CSI) feedback is vital for MIMO to deliver its promised throughput gain. One of the reasons that multiple transmission modes are defined in LTE MIMO is the challenging task of accurate CSI feedback. Advanced CSI framework is currently considered to solve this issue. For instance, the aperiodic reference signal is to be triggered for channel measurement and interference measurement, only when needed. The feedback CSI is based on linear combination of selected beams. Orthogonal basis based beam selection can be considered when both quantized amplitude and phase coefficients are used to combine the selected beam. CSI would be fed back in multiple levels, as illustrated in **Fig. 4**.



▲**Figure 4. CSI feedback.**

3.4 Numerology, Frame Structures, HARQ and Duplex

CP - OFDM would be the fundamental waveform for NR. Hence, the LTE numerology can largely be reused. In August 2016, it was decided that the reference numerology of NR is LTE, i.e., 14 OFDM symbols in 1 ms subframe (SF), 15 kHz for each subcarrier and 12 subcarriers in a physical resource block (PRB). The scaling of subcarrier spacing is to the power of 2, for example, 3.75 kHz, 7.5 kHz, 30 kHz, 60 kHz. Consequently, the symbol duration is shrunk to 1/2, and 1/4 of reference symbol duration. Subframe duration is fixed to 1 ms, and symbol level alignment should be enforced, in the case of same overhead of cyclic prefix. The symbol structures of 3.75 kHz, 7.5 kHz, 15 kHz, 30 kHz and 60 kHz for normal CP family are shown in **Fig. 5**.

It is observed in the cases of 30 kHz and 60 kHz that other than the first OFDM symbol, all the other OFDM symbols within 0.5 ms have the same size.

NR faces the issue of how to multiplex different numerologies. To reduce the complexity and resource fragmentation, nested structure is adopted which can be applied to the cases where physical resource blocks (PRBs) of different numerology are multiplexed in TDM or FDM (**Fig. 6**).

The frame structure of NR should give the networks more flexibility in scheduling, regardless of frequency division duplex (FDD) or time division duplex (TDD) spectrum bands. In

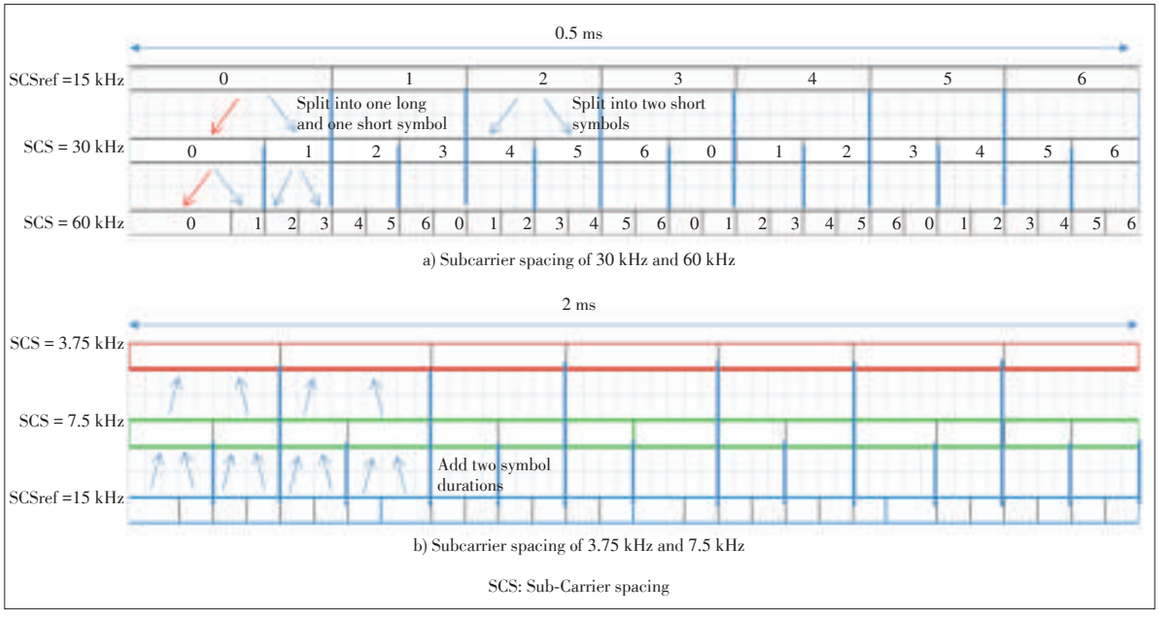
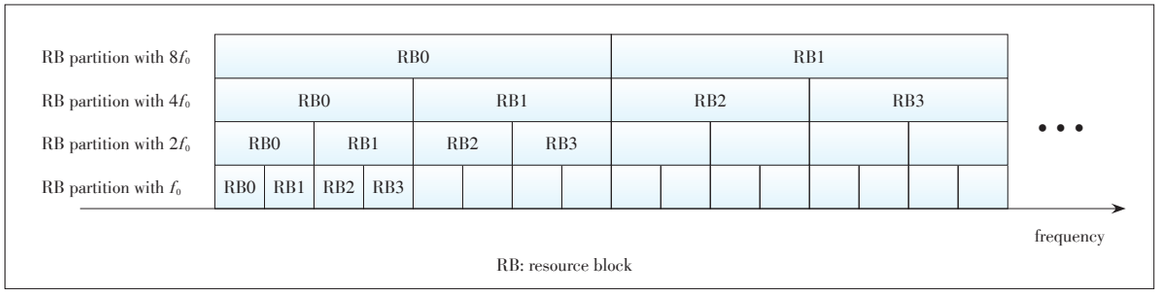


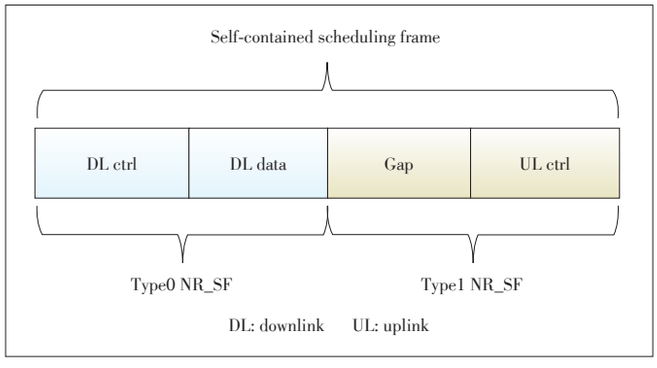
Figure 5.▶ NR frame structure.



◀Figure 6. Nested structure of PRB.

fact, the design principle of NR frame structure would be general for FDD and TDD. This is quite different from the case of LTE where seven subframe configurations are defined for LTE-TDD, leading to very complicated HARQ timing. In particular, the all downlink or all uplink subframes of LTE-TDD cause longer round-trip time for HARQ and severely degrade the user experience. In NR, self-contained subframe structure becomes the prevailing view. This is also made possible because of the fast growing processing power of terminal devices. As illustrated in **Fig. 7**, 1 ms subframe can contain downlink control at the beginning of a subframe, followed by downlink data. This arrangement would give more time budget for the terminal receiver to decode the data, and be able to provide the feedback within this subframe. Reference signal for demodulation can be placed at the beginning of this subframe to facilitate in-time channel estimation. After the gap for Tx/Rx switching, the last symbol can be used for uplink control channel. Note that the structure in Fig. 7 represents the normal coverage situation where downlink and/or uplink control channel can be carried in partial subframe (i.e., one or several OFDM symbols). If the scenario is coverage limited, the entire 1 ms subframe can be used for uplink control, or OFDM symbols of multiple subframes are used for downlink control.

Due to self-contained frame structure, HARQ timing becomes more flexible. For instance, for downlink, NR supports both same-slot scheduling and cross-slot scheduling. The timing relationship between downlink (DL) data and the corresponding acknowledgement in uplink (UL) can be dynamically or semi-statically indicated. Uplink supports asynchronous HARQ. The timing offset between the UL assignment and the corresponding UL data can also be dynamically or semi-statically indicated.



▲Figure 7. Self-contained scheduling frame.

Another key principle of frame structure is to minimize the resource allocation for common channel. Here, it is more reflected as the control channel and reference signals are dedicated to each user. Control channel does not need to span over the entire system bandwidth, and can occupy one or several sub-bands.

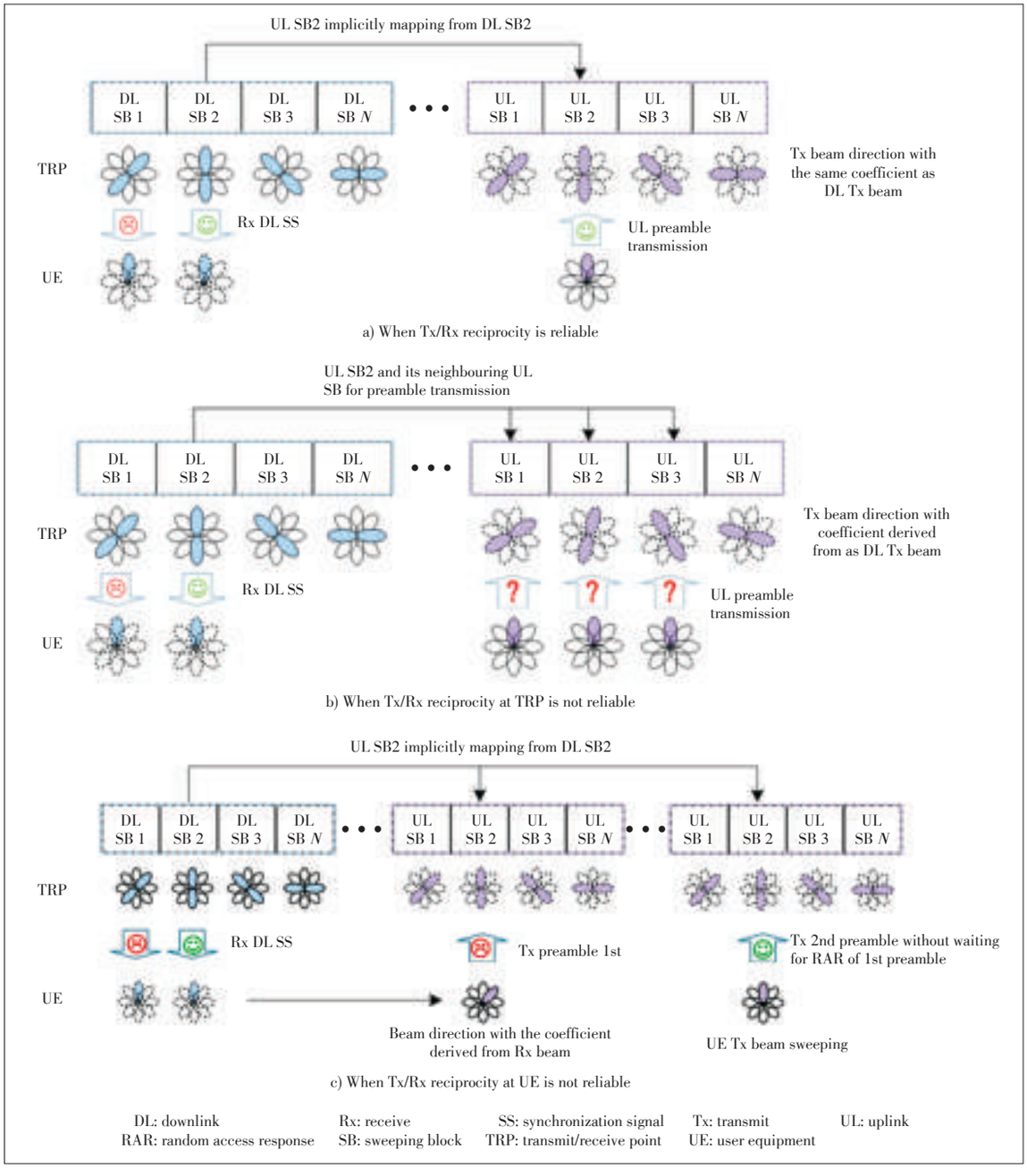
3.5 Initial Access and Mobility

Generally speaking, initial access and mobility support of different generations of cellular communications would share many common functions such as synchronization, system information broadcasting, and random access. Nevertheless, initial access of NR has its own characteristics, for example, extensive use of beamforming for common channels, super wide bandwidth for initial cell searching, minimum usage of resource for common channels, and introduction of a new RRC state.

For high frequency bands, analog beamforming would be more often used. Due to the limited number of transmit and receive unit (TXRU), signals may have to be transmitted or received along only one beam direction in one sweeping block (SB). **Fig. 8a** shows an example of uplink SB determination when Tx/Rx reciprocity is available. The base station (transmit/receive point, TRP) can indicate to the UE in the system information that there is a one-to-one mapping between sweeping blocks in DL sweeping time interval (STI) and UL STI. The UE should send physical random access channel (PRACH) preamble in the UL SB corresponding to the DL SB in which the best or acceptable signal strength is detected.

If the Tx/Rx reciprocity at the TRP is not reliable and TRP Rx beam sweeping is needed, the TRP can request the UE (e.g. in SI) to repeat the preamble transmission on multiple UL SBs, as seen in **Fig. 8b**. When Tx/Rx reciprocity is not available at UE, the UE can send preambles with different UE Tx beam directions in different UL SBs and/or in different UL STIs, as seen in **Fig. 8c**.

The super wide system bandwidth means that multiple numerologies may coexist in FDM fashion. Ideally for numerology, one synchronization signal would be defined. However, this may lead to excessive overhead. Right now, RAN1 is striving for minimizing the number of subcarrier spacings for synchroni-



◀Figure 8. Beam sweeping when Tx/Rx reciprocity is reliable.

zation signals and primary broadcast channels.

4 Phased Approach for NR

4.1 Technologies and Scopes for NR Phase 1

Apart from the basic frame structure, numerology and initial access to be defined for a network to be operable, NR phase 1 should also specify a few other key technologies that hopefully help to fulfill some KPIs of 5G. They are:

- Channel coding scheme for eMBB and URLLC
- MIMO techniques
- Beam management for initial access
- HARQ and scheduling.

It is noted that the links in consideration are primarily between the base station and the UE. The frequency bands are licensed bands and up to 40 GHz.

4.2 Technologies and Scope for NR Phase 2

Technologies and the scope in Phase 2 will extend the use cases of 5G and help to fulfill all KPIs of 5G. They are:

- Waveform above 40 GHz

- mMTC
- Flexible duplex of FDD
- Interworking with non-3GPP systems
- Wireless relay
- Satellite communications
- Air-to-ground and light air craft communications
- Extreme long distance coverage
- Sidelink
- Vehicle-to-Vehicle (V2V) and Vehicle-to-X (V2X)
- Multimedia broadcast/multicast service
- Shared spectrum and unlicensed spectrum
- Location/positioning functionality
- Public warning/emergency alert
- New self-organized network (SON) functionality
- 2-step RACH process
- Uplink based mobility
- Analog CSI feedback and explicit CSI feedback for MIMO.

5 Conclusions

In this paper we provided an overview of new radio (NR) physical layer for 5G, which is currently in the study phase in 3GPP. This survey starts from the framework of NR physical layer, followed by more detailed description of each key technology and functionality. The items to be specified in Phase 1 and Phase 2 are also listed.

References

[1] J. M Meredith, "Study on channel model for frequency spectrum above 6 GHz," 3GPP TR 38.900, Jun. 2016.
[2] J. Krause, "Study on scenarios and requirements for next generation access technology," 3GPP TR 38.913, Sept. 2016.
[3] NTT DoCoMo, "New SID proposal: study on new radio access technology," 3GPP RP-160671, Mar. 2016.
[4] Y. Yuan, Z. Yuan, G. Yu, et. al, "Non-orthogonal transmission technology in LTE evolution," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 68–74, Jul. 2016. D10.1109/MCOM.2016.7509381.
[5] NTT DoCoMo, "Technical Report: study on new radio access technology," 3GPP TR 38.802, Mar. 2017.
[6] R. G. Gallager, *Low Density Parity Check Codes*. Cambridge, USA: MIT Press, 1963.
[7] Y. Yuan, *LTE/LTE-Advanced key technologies and system performance*. Beijing, China: Posts & Telecom Press, Jun. 2013.
[8] E. Arikan, "Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, Vol. 55, no. 7, pp. 3051–3073, Jul. 2009. doi: 10.1109/TIT.2009.2021379.

Manuscript received: 2016-10-26

Biographies

YUAN Yifei (yuan.yifei@zte.com.cn) received his B.S. and M.S. degrees from Tsinghua University, China. He received his Ph.D. from Carnegie Mellon University, USA. From 2000 to 2008, he worked with Alcatel-Lucent on 3G and 4G key technologies. Since 2008, he has worked for ZTE Corporation researching 5G technologies and standards for LTE-Advanced physical layer. His research interests include MIMO, iterative codes, and resource scheduling. He was admitted to the Thousand Talent Plan Program of China in 2010. He has written three books on LTE-A relay, Narrow-band IoT, and LTE-Advanced key technologies and system performance, respectively. He has had more than 40 patents approved.

WANG Xinhui (wangxinhui@zte.com.cn) received his bachelor's and master's degrees from Northeastern University of China. He joined ZTE Corporation in 2000, working on software development and system design of wireless communication system. Since 2006, he has been focusing on advanced radio access technology research and standardization. His research interests include network densification, massive machine communications and high frequency communications. He has been elected as the vice chairman of 3GPP GERAN TSG for three consecutive terms since 2011. He is currently chairing the Radio Access Technology Group of IMT-2020(5G) Promotion Group as the vice chairman. He held over 40 granted patents globally.

DOI: 10.3969/j. issn. 1673-5188. 2017. S1. 002
http://kns.cnki.net/kcms/detail/34.1294.TN.20170629.1752.002.html, published online June 29, 2017

Enhanced OFDM for 5G RAN

Zekeriyya Esat Ankaralı[1], Berker Peköz[1], and Hüseyin Arslan[1,2]
(1. Department of Electrical Engineering, University of South Florida, FL 33620, USA;
2. College of Engineering, Istanbul Medipol University, Istanbul 34810, Turkey)

Abstract

Support of many different services, approximately 1000x increase of current data rates, ultra-low latency and energy/cost efficiency are among the expectations from the upcoming 5G standards. In order to meet these expectations, researchers investigate various potential technologies involving different network layers and discuss their tradeoffs for possible 5G scenarios. As one of the most critical components of communication systems, waveform design plays a vital role here to achieve the aforementioned goals. Basic features of the 5G waveform can be given in a nutshell as more flexibility, support of multiple access, the ability to co-exist with different waveforms, low latency and compatibility with promising future technologies such as massive MIMO and mmWave communications. Orthogonal frequency division multiplexing (OFDM) has been the dominant technology in many existing standards and is still considered as one of the favorites for broadband communications in 5G radio access network (RAN). Considering the current interest of industry and academia on enhancing OFDM, this paper drafts the merits and shortcomings of OFDM for 5G RAN scenarios and discusses the various approaches for its improvement. What is addressed in this paper includes not only enhancing the waveform characteristics, out of band leakage and peak to average power ratio in particular, but also methods to reduce the time and frequency redundancies of OFDM such as cyclic prefix and pilot signals. We present how the requirements of different 5G RAN scenarios reflect on waveform parameters, and explore the motivations behind designing frames that include multiple waveforms with different parameters, referred to as numerologies by the 3GPP community, as well as the problems that arise with such coexistence. In addition, recently proposed OFDM-based signaling schemes will also be discussed along with a brief comparison.

Keywords

5G waveform; 5G RAN; eMBB; multicarrier systems; OFDM

1 Introduction

Exponential growth in the variety and the number of data-hungry applications along with mobile devices leads to an explosion in the need for higher data rates, and this is definitely the main driving factor in 5G [1]. Therefore, a wide range of data rates up to gigabits per second are targeted in 5G technologies which are expected to be deployed around 2020. In order to achieve these goals, academia has been in a great collaboration with industry as obviously seen in European Union projects as 5GNOW [2], METIS [3], MiWaveS [4] and FANTASTIC-5G [5]. Along with those, standardization has been started in 3GPP to deliver the demanded services timely.

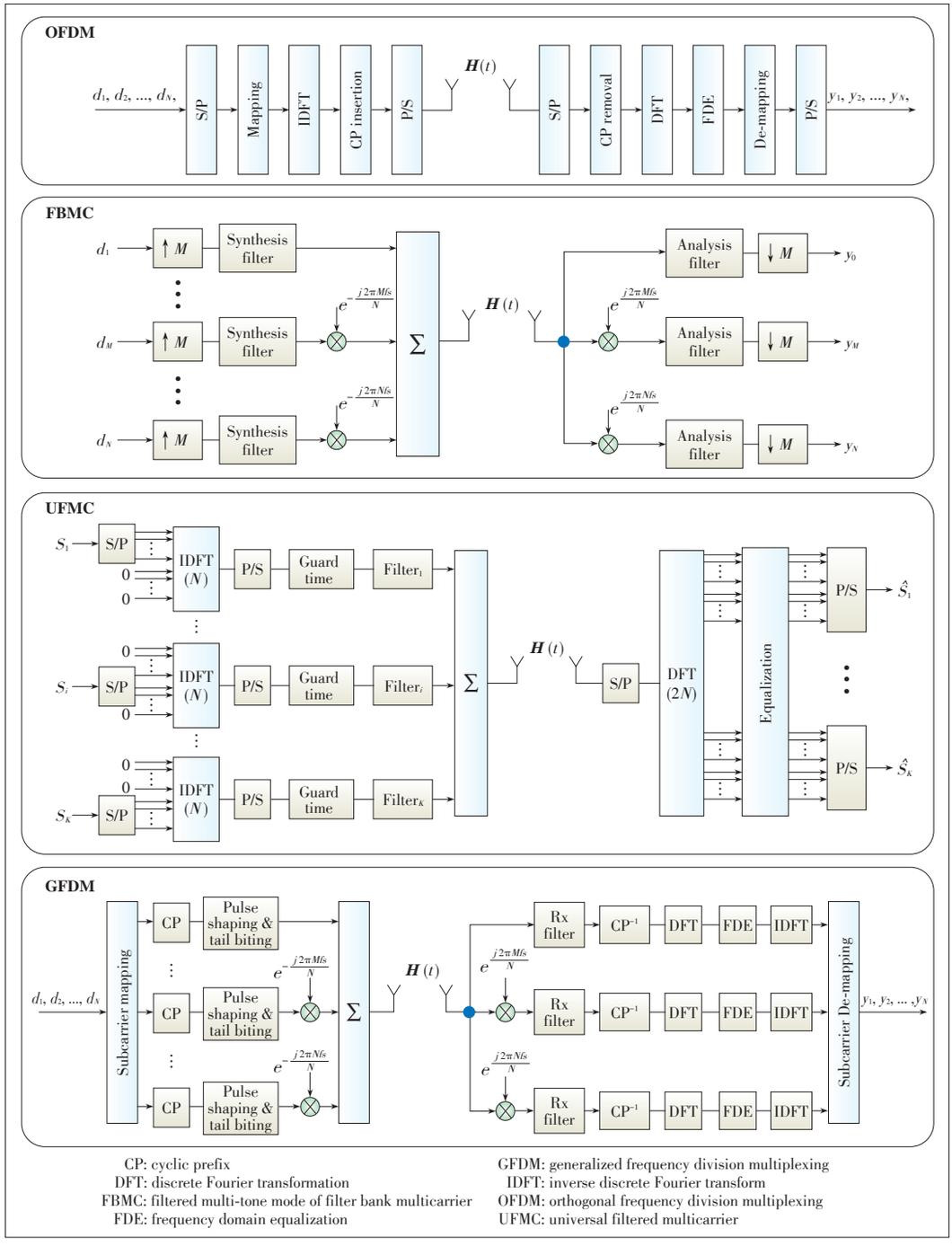
One of the most challenging parts of achieving targeted high data rates is physical scarcity of the spectrum, and researchers have been putting an extensive effort to this challenge. One popular approach is to extend existing spectrum towards virgin higher frequencies up to 100 GHz [6]. Another approach is to increase spectral efficiency for a given spectral resource. Millimeter wave (mmWave) communications and massive multiple-input multiple-output (MIMO) are the representative concepts of these two approaches and very promising technologies for facilitating 5G goals, especially for enhanced mobile broadband (eMBB) services which constitutes one of the main service groups considered for 5G radio access network (RAN).

Even though not considered as the revolutionary part of 5G, one of the most fundamental components of any communication system is the waveform design. Therefore, intensive discussions are being conducted in academia and industry, in order to select the proper waveform meeting 5G requirements. Among all the candidates, multicarrier techniques are prominent especially for broadband wireless communications due to several advantages such as immunity against frequency selectivity, multiuser diversity support and adaptive modulation/coding techniques. Orthogonal frequency division multiplexing (OFDM) has been the dominating technology so far and successfully deployed in many of the current standards such as Long Term Evolution (LTE) and Wi-Fi. In the transition from existing technologies (4G) to the next generation, waveform se-

lection ramifies to two paths for 5G RAN. The first one is reconsidering OFDM based methods by improving its characteristics and handling its drawbacks with proper solutions. The second one, on the other hand, is to implement alternative multicarrier technologies and redesign everything based on a different rationale. **Fig. 1** shows transceiver block diagrams for OFDM and other popular multicarrier schemes including filtered multi-tone mode of filter bank multicarrier (FBMC), universal filtered multicarrier (UFMC) and generalized frequency division multiplexing (GFDM). Let us firstly provide the merits and challenges of the multicarrier technologies considered as an alternative to OFDM in the context of 5G expectations.

1.1 FBMC

FBMC is one of the most well-known multi-carrier modulation formats in wireless communications literature which is also discussed as a 5G waveform in [7]. It offers a great advantage of shaping each subcarrier and facilitating a flexible utili-



CP: cyclic prefix
DFT: discrete Fourier transformation
FBMC: filtered multi-tone mode of filter bank multicarrier
FDE: frequency domain equalization
GFDM: generalized frequency division multiplexing
IDFT: inverse discrete Fourier transform
OFDM: orthogonal frequency division multiplexing
UFMC: universal filtered multicarrier

Figure 1. Block diagrams of popular multicarrier schemes (OFDM, FBMC, UFMC and GFDM) considered for 5G radio access.

zation of spectral resources along with meeting various system requirements, e.g., low latency, multiple access, etc. This is also an advantage for making signal robust against channel effects, i.e., dispersion in time and frequency domains. For example, rectangular filters are preferable for time dispersive channels while raised cosine filters are more robust against frequency dispersion. Many other pulse shaping filters are also investigated to cope with various effects of the channel and provide a reliable system design based on different scenarios [8].

Despite all the advantages of FBMC, the significantly long filter lengths resulting in colossal symbol durations not only become a problem if low latency applications or short bursts of machine type communications are in focus [9], but also introduce an excessive computational complexity for MIMO detection as the channel coherence bandwidth would fall below the subcarrier bandwidth [10], which would mean problems in all main applications of 5G.

1.2 UFMC

UFMC is a generalized version of filtered multicarrier techniques where groups of subcarriers, i.e., sub-bands, are filtered rather than filtering each subcarrier individually [11]. By doing so, interference between neighboring sub-bands is decreased compared to conventional OFDM. Also, sub-band based filtering operation, when compared to the subcarrier filtering operation performed by FBMC, aims to increase the efficiency for short-burst type communications such as IoT scenarios or very low latency packets by reducing the filtered symbol duration and outperforms both cyclic prefix (CP)- OFDM and FBMC for such use cases [9]. A similar scheme is also presented as resource block (RB)-filtered OFDM in [12]. On the other hand, while UFMC aims to solve the problems of FBMC while maintaining its advantages, the increased fast Fourier transformation (FFT) length introduces complexity issues at the transmitter and receiver operations.

1.3 GFDM

GFDM is a block-based multicarrier filtered modulation scheme, designed to address the challenges in the vast usage scenarios of the fifth generation by providing a flexible waveform [9]. GFDM allows reuse of techniques that were originally developed for OFDM, as circular convolution is employed to filter the individual subcarriers, making the GFDM frame self-contained in a block structure. For tactile internet scenarios, GFDM can be distinguished from other multicarrier waveforms by how it achieves robustness over highly mobile channels. This is accomplished via taking the advantage of the transmit diversity provided by the easy generation of impulse responses simply obtained with circularly shifting the single prototype filter in time and frequency. To improve the reliability and latency characteristics even further, the GFDM waveform can be combined with the Walsh-Hadamard transform for increased performance in single-shot transmission scenarios. When combined with offset quadrature amplitude modulation mapping, GFDM avoids self-generated interference if non-orthogonal filters are employed for next generation multiple accessing.

In a different point of view, GFDM can be considered as a highly parameterizable waveform that is flexible across frames rather than a single waveform. By choosing the parameters of the GFDM waveform appropriately, one can obtain different waveforms such as OFDM, single carrier-frequency domain equalization (SC-FDE), FBMC and Faster-Than-Nyquist at the output, as demonstrated in [9]. In spite of these type of interesting flexibilities, GFDM is a computationally extensive scheme mainly because FFT/inverse FFT (IFFT) could not immediately be employed at the GFDM based transceiver [13]. Furthermore, the circular convolution used in the filtering process, referred to as tail-biting in [14], introduces non-orthogonality across subcarriers as explained in [8]. Therefore, a successive interference cancellation at the receiver side is required so as to remove inter-carrier interference (ICI) [15].

Unlike the aforementioned technologies, OFDM has been widely and successfully used in wireless digital communication systems such as LTE and Wi-Fi due to its numerous advantages such as low-complexity implementation with FFT and the robustness against multipath channels with single-tap FDE. However, plain OFDM signals suffer from the distortions due to the non-linear characteristics of power amplifier (PA). At the same time, the block nature of OFDM symbols may result in a high out-of-band (OOB) leakage and cause severe adjacent channel interference. Considering these issues, alternative schemes, GFDM, UFMC and FBMC definitely offer some advantages over OFDM. However, backward compatibility of OFDM with the existing technologies along with the other advantages makes enhancement of OFDM more appealing for the industry rather than going for a new waveform, as far as seen in the current standard discussions [16]–[18]. Therefore, in this paper, we draft various approaches addressing the shortcomings of OFDM and discuss how OFDM fits the envisioned 5G concepts and technologies.

Organized into the following five sections, this paper first presents the methods that enhance the spectral compactness, robustness against nonlinear effects of radio frequency (RF) front-end and spectral efficiency of OFDM. We provide examples of transmitter algorithms that lower peak-to-average power ratio (PAPR) and OOB leakage of OFDM without requiring major modifications at the receiver side. Secondly, considering newly introduced 5G RAN applications such as massive MIMO, we discuss the methods that address reducing the redundancies of OFDM in time (CP duration) and frequency (pilots) in order to increase spectral efficiency. Following that, we describe one of the main goals of 5G, which is to make all the diverse applications requiring different waveforms co-exist in the same frame, as addressed by 3GPP community within numerology contributions. The motivations and challenges of this concept are briefly investigated. Finally, new OFDM-derived ap-

proaches that aim to overcome the drawbacks of OFDM are provided along with the issues they encounter.

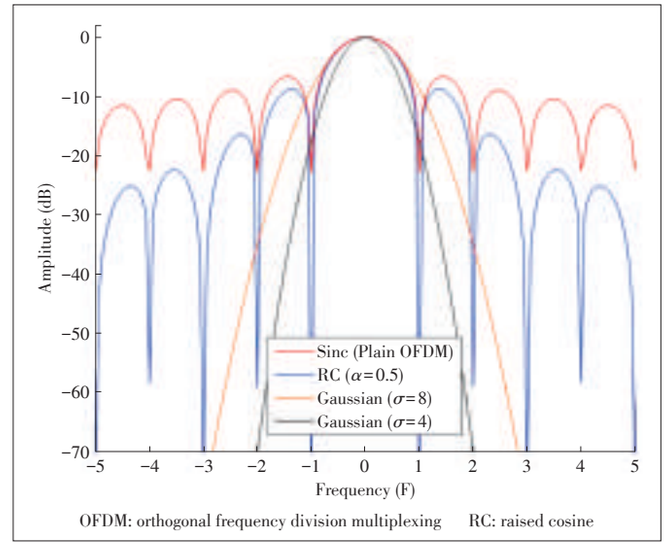
2 Improvements in Waveform Characteristics

In a general sense, the key terms characterizing a basic OFDM waveform are multicarrier modulation and rectangular pulse shape[1], and the majority of the advantages and disadvantages of OFDM are stemming from these features. In this section, we discuss how to improve characteristics of OFDM and make it a more convenient waveform for 5G RAN in terms of high PAPR, OOB leakage and redundancy introduced by CP.

2.1 OOB Leakage Suppression

High OOB leakage is a major issue in OFDM due to the inherent rectangular shape of OFDM symbols. In the frequency domain, subcarriers are shaped by sinc functions and addition of their sidelobes results in a considerable energy leakage on the neighboring channels as shown in **Fig. 2**. Although there are well-known filters emitting less energy on side bands, e.g., raised cosine and Gaussian filters, OFDM does not allow pulse shaping unlike FBMC and GFDM, and therefore, a severe interference might be inevitable for users operating on the neighboring frequencies, especially for asynchronous scenarios. Leaving sufficient guard bands between the users might be considered as a practical solution, but this would not be an efficient way of utilizing spectral resources. In 5G scenarios, as far as envisioned so far, a huge number of asynchronous and data-hungry users should co-exist within a limited spectrum. Therefore, OFDM signals should be more localized in the frequency domain by handling OOB leakage problem in a practical way to adapt OFDM to such scenarios.

For the aforementioned purpose, OOB leakage of OFDM signals has been extensively addressed with numerous techniques in the literature as reviewed and compared in [19]. For instance, a time domain windowing approach is proposed in [20], which can make the transitions between the OFDM symbols smoother and avoid signal components at higher frequencies. Hence, the OOB leakage of OFDM symbols is significantly reduced. This approach became very popular due to its simplicity, effectiveness and requirement of no modification at the receiver side. However, the introduction of an extra redundancy as much as the windowing duration remained a problem. In [21], while the total duration for CP and windowing is kept constant for all subcarriers, windowing is mostly applied to the edge subcarriers since the leakage of edge subcarriers causes more interference on the adjacent frequencies. In a practical multiuser scenario where users need different CP sizes, this approach can decrease the windowing redundancy compared to



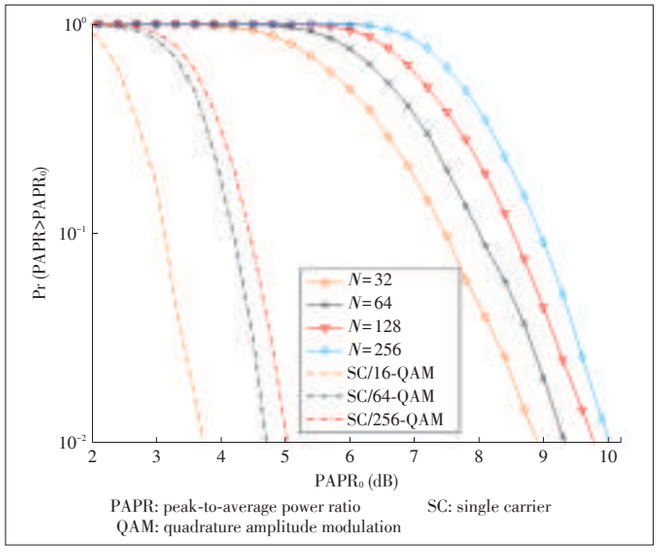
▲**Figure 2. Frequency responses of an OFDM subcarrier (sinc) and various filters.**

the classical approach via a convenient user scheduling. Users with low time-dispersive channels are assigned to the edge subcarriers and users having highly time-dispersive channels are assigned to the inner subcarriers. Thus, the total duration required for CP and windowing could be shorter without causing any problem. In [22]–[25], the OOB leakage is addressed from frequency domain perspective. A set of subcarriers, named as cancellation carriers, are allocated for canceling the sidelobes in [22], [23]. However, such approaches also introduce redundancy in the frequency domain and degrade spectral efficiency similar to classical windowing approach. In [24], sidelobe suppression is done by weighting subcarriers in such a way that sidelobes are combined on adjacent frequencies as destructively as possible. However, weighting leads to a pre-distortion of subcarriers and bit-error rate (BER) performance naturally reduces. In order to limit this distortion, a frequency domain precoder is proposed in [25], which only maintains the spectrum of OFDM signals under the prescribed mask rather than forcing OOB leakage to zero. By doing so, interference on the adjacent frequencies is kept on a reasonable level at the expense of a smaller degradation BER performance.

2.2 PAPR Mitigation

As a consequence of multicarrier transmission, i.e., transmitting multiple signals in parallel, high PAPR is inevitable for OFDM signals due to the probable constructive combination of signals in time domain. In **Fig. 3**, a comparison between SC signals having various modulation orders up to 256-QAM, and OFDM signals having a different number of subcarriers (N) is provided. Obviously, there is a huge difference in PAPR even when the number of subcarriers is as low as 32. It could be ignored for users requiring low power transmission. However, in many scenarios such as the mobile users on cell edges, a reli-

[1] **CP deployment can also be considered among these terms, however, that will be discussed in later sections.**



▲**Figure 3. PAPR comparison between single carrier signals with different modulation orders and OFDM with different number of subcarriers (*N*).**

able transmission requires high power and high PAPR of the signal for this scenario makes the signal vulnerable to non-linear effects of RF front-end components. These components typically have a limited linear range, and any part of the signal exceeding the linear range is non-linearly scaled. Non-linear scaling of a signal can also be referred as multiplying a part of signal components with various coefficients. This makes a time-varying channel effect on the signal, and the signal is distorted as if it is exposed to a Doppler spread effect at the transmitter. As a result, non-linearity of RF components may lead to severe interference not only in the user's band but also for the others operating on neighboring frequencies due to the spectral re-growth. At this point, one may notice that the OOB leakage is not only the function of the waveform itself but also the spectral regrowth of the ideal waveform signals due to the high PAPR in practice. Then, even if the OOB suppression performance of the related studies in the literature is quite satisfactory, a good scheme needs to address PAPR and OOB leakage jointly for fixing these two shortcomings, practically.

PAPR suppression techniques are surveyed well in [26], however, many of them tackle with PAPR individually without considering OOB. On the other hand, some existing studies use PAPR reduction concepts for suppressing OOB. This is achieved by actively selecting some predesigned sequences, i.e., selected mapping (SLM) sequences in [27]. Another well-known PAPR reduction method, partial transmit sequences are applied on OFDM signals partitioned into contiguous blocks in the frequency domain in [28]. Additionally, the optimized phase rotations are multiplied by each sub-block to provide a contiguous transition between the OFDM symbols to suppress OOB leakage along with PAPR. Another joint suppression method is presented in [29], where the constellation points are dynamically extended. For the similar purpose, a method called CP alignment is proposed in [30] similar to the interference alignment method presented in [31]. The key idea in this method is to add a perturbation signal, called alignment signal (AS), to the plain OFDM symbols in order to reduce the PAPR and OOB leakage such that the AS aligns with the CP duration of the OFDM symbols after passing through the channel. However, this method completely relies on the perfect channel estimation and any error might result in an interference on the data part. In order to fix this problem, a recent method called static CP alignment is proposed in [32] where the AS is designed according to a pre-determined filter independent of the channel. These methods also provide physical layer security to some extent by using additional signals that confuse the unauthorized users [33].

Joint PAPR and OOB leakage suppression techniques are definitely offering a comprehensive solution in enhancing characteristics of OFDM signals. However, they require symbol based active optimization, which introduces complexity issues at the transmitter side. Therefore, simpler solutions such as windowing are still needed in this field.

3 Reducing Redundancies of OFDM

Classical CP-OFDM is known as one of the most spectrally efficient transmission schemes and sufficiently satisfying the requirements of LTE-Advanced Pro and currently used IEEE 802.11 systems. However, it still suffers from redundancies, in both time and frequency domains. The redundancy in time comes from the use of CP while the redundancy in frequency domain comes from the use of pilot subcarriers and guard bands. In order to adopt OFDM for future radio access technologies and to achieve the aforementioned goals in data rate, these redundancies should be reduced, significantly.

3.1 CP Reduction

In communication channels with multipath delay spread, a guard interval (GI) is required to prevent leakage in time between the successive symbols. CP is a smart way to utilize the guard interval by copying the samples from the end of a symbol and pasting them to its beginning. Thus, the linear convolution of the channel becomes a circular convolution, which makes the channel matrix diagonalizable only by taking its Fourier transform and enables a simple equalization in the frequency domain. The length of the CP is chosen to be larger than the expected delay spread to avoid any inter-symbol interference (ISI) and ICI. Also, in order to maintain orthogonal coexistence of neighboring transmission blocks, predefined values have been used for the length of the CP and applied to all the blocks. For example, two CP rates are defined in LTE where the normal CP duration in terms of symbol duration (T_{SYM}) is given as $T_{SYM} \times 9/128$ while the extended CP duration is $T_{SYM} \times 32/128$.

Recent works have shown that extending the CP duration might not be the best approach to combat against long delay spreads [34]–[36]. The newly proposed 5G scenarios have introduced their own methods for ISI mitigation. An example is mmWave MIMO systems that employ highly directional transmission using beamforming. For such systems, beam switching reference signals are broadcasted so that the receivers can determine which predefined beam is directed to their way, resulting in higher signal to interference plus noise ratio for all receivers. Inventors have shown in [34] that such signals not only provide support for beam switching but can also be used to estimate delay spread exceeding CP duration. Then, ISI could be canceled from received symbols, which reduces the required CP. In [35], authors claim that MIMO receivers can identify the presence of any residual interference after equalization by evaluating the channel matrix. It was shown that decreasing the modulation and coding index instead of extending the CP length, would increase the throughput. In addition to these methods, advanced signal processing techniques can be employed to mitigate the interference. A good example of that is the bi-directional M-algorithm based equalizer proposed in [36]. It has shown that a system which experiences a delay spread six times longer than the CP duration exhibits the same performance of a conventional system with sufficient CP. It is achieved at only the expense of performing two iterations of the proposed algorithm referred to as trellis-based interference detection and mitigation.

In some scenarios, the maximum excess delay might be much less than the duration of normal CP which makes the minimum CP overhead of 7% a pointless guard for LTE systems. To reduce this overhead, the authors of [37] present the idea of a flexible frame design. A wider range of options in terms of subcarrier spacing and CP length are used for the OFDM symbols inside the proposed frame structure. Then, the users experiencing similar channel dispersions are grouped and the proper symbol parameters are determined for each group within the frame. Thus, overall efficiency is enhanced by avoiding inconvenient parameter selection. An extension of this approach to mmWave single user MIMO systems can also be found in [34], where the use of additional subframe configurations is presented independently for each user. Increasing the number of options for the CP duration is also recommended in the 3GPP standard contributions [38], [39].

CP overhead also constitutes a disadvantage for the low latency required applications as it introduces delays in the transmission which might cause drawbacks for 5G services such as ultra reliable low latency communications (URLLC). In the algorithm proposed in [40], the author removed the CPs entirely from all symbols except the first one for reducing total transmission delay. In the proposed method, the CP used by the first symbol is utilized to obtain a detailed estimation of the channel time and frequency characteristics. Afterwards, the subcarrier spacing is reduced to $\Delta f/(N_{SYM}-1)$, where Δf denotes the subcarrier spacing used in the first symbol and N_{SYM} denotes the total number of symbols including the first one. All symbols sent later are combined to fit in the same bandwidth used by the first symbol using an IFFT size of $(N_{SYM}-1)\times N_{OFDM}$, and sent as a single symbol without CP. Thus, the total transmission time is reduced by $(N_{SYM}-1)\times T_{CP}$, where T_{CP} is the CP duration.

3.2 Pilot Decontamination

Another redundancy that has been with OFDM since it became practical is the use of pilots within the subcarriers. In time domain duplexing (TDD) systems, inside a cell, the mobile stations transmit mutually orthogonal pilot sequences to the base station (BS) so that the BS can estimate the channel in the uplink (UL), and assuming channel reciprocity, precode accordingly for the downlink (DL). In the case of frequency domain duplexing (FDD) systems, because the channel state for the UL and DL is different, a two-stage procedure is required. The BS first transmits pilot symbols and then, the users feedback their channel state information to the BS. For the massive MIMO concept with the TDD case, many beams are established for a vast number of users compared to the past. Each beam requires a different mutually orthogonal sequence, which increases the length of the sequences immensely and decreases the resources available to transmit data symbols. For the FDD case, the same situation happens as the number of transmit antennas at the BS goes to infinity. A proposed method to reduce this overhead is reusing pilot sequences of nearby cells, which introduces inter-cell-interference and gives rise to the "pilot contamination" effect [41]. The high number of lengthy pilots also increases the latency and makes Internet of Things (IoT)-type sporadic and short messages inefficient.

Some researchers allow the use of the pilots but try to reduce the overhead, which can be referred to as soft pilot mitigation. In [42], the authors propose using only the amplitudes of the subcarriers as the pilots, and the phase of the same subcarriers can be used to transmit information in an effort to increase the data rate. In [43], the author proposes many techniques to mitigate pilot contamination for the TDD case. The terminals are suggested to match the DL reference signal powers in the UL, in order to reduce both the overall pilot interference in the neighboring cells and the pilot overhead required for closed loop power control. Another suggestion is reusing pilots softly to avoid inter-cell-interference. Some pilot sequences are proposed to be assigned for use only at the cell edge whereas the same groups of pilot sequences can be transmitted with less power near the BSs. Even further, the author suggests that the angular resolution provided by the massive number of antennas can be used to coordinate pilot allocation between cells and safely reuse the pilot sequences for spatially separated terminals. In [44], the authors have aligned the Power Delay Profiles (PDPs) of the users served by the same BS to orthogonalize the pilots sent within the common OFDM symbol. By

aligning the PDPs for the same number of users that can be sounded within the same symbol, it has been shown that the average signal-to-noise ratio (SNR) can increase by half.

Another group of researchers has proposed blind channel estimation or signal detection techniques to remove pilots completely, which would be appropriately called hard pilot mitigation. In [45], the authors have shown that pilots can completely be removed as the singular value decomposition of the received signal matrix projects the received signal onto an interference-free subspace governed by an easily predictable nonlinear compound. The authors have demonstrated that the proposed subspace projection method outperforms linear channel estimation if a power margin between the users of interest and interfering users are provided, especially when the base station antennas outnumber the coherence time (in number of symbol durations). In [46], the authors have treated the detected UL data as pilot symbols to obtain the least squares estimate of the channel. Also, by estimating the channels of all users sequentially, they obtained extracting vectors that accurately extract the desired data from the mixture signal.

4 Numerology

Wireless users have different requirements in waveform based on the service that they get or channel conditions that they experience. Therefore, there is not any one-size-fits-all solution for waveform design. The ongoing discussions on the usage of different numerologies also confirm this argument and consequently, one crucial expectation from future standards is the allowance to use multiple waveforms in one frame. This would definitely constitute a great relaxation in selecting the most proper waveform based on user needs. However, managing their coexistence is a critical issue. Especially, inter-user interference management for uplink/downlink, synchronous and asynchronous scenarios should be carefully investigated to utilize spectral resources efficiently.

In the context of 3GPP 5G standardization contributions, the term numerology refers to the configuration of waveform parameters, and different numerologies are considered as OFDM-based sub-frames having different parameters such as subcarrier spacing/symbol time, CP size, etc. [47]. By designing such numerologies based on user requirements, industry targets to meet the aforementioned user-specific demands to some extent. A general illustration of such numerologies is provided in **Fig. 4**. Here, numerology-I would be properly assigned to highly mobile users having more time-variant channels and the ones with low latency requirement. On the other hand, numerology-II offers more robustness against frequency selectivity and includes less redundancy due to low CP rate.

Let us give more details on the parameters and the importance of their selection for designing different numerologies:

1) CP length: The basic function of CP is to avoid inter-symbol interference and in-band interference. In order to achieve that, CP length should be specified as longer than the maximum excess delay of the channel impulse response. Therefore, users experiencing a wireless channel causing high dispersion in time (or more selectivity in frequency) should have longer CP lengths compared to the users with low dispersive channels. In addition, CP makes the signal robust against time synchronization errors. This might be very critical especially for asynchronous UL scenarios and low latency demanding services.
2) Subcarrier spacing: It can also be referred to as subcarrier bandwidth and is directly related to the duration of an OFDM symbol. When the CP size is determined on the basis of the channel conditions and application requirements, decreasing subcarrier spacing increases spectral efficiency as the CP rate decreases. However, for highly mobile users, channel responses might vary within a symbol duration which leads to ICI. Therefore, symbol time should be kept smaller by increasing subcarrier spacing in order to make the transmission robust against time-varying, i.e., frequency dispersive channels. Additionally, a proper choice of subcarrier spacing is very critical for immunity against phase noise, which is specifically important for users operating on high frequencies such as mmWave frequencies.

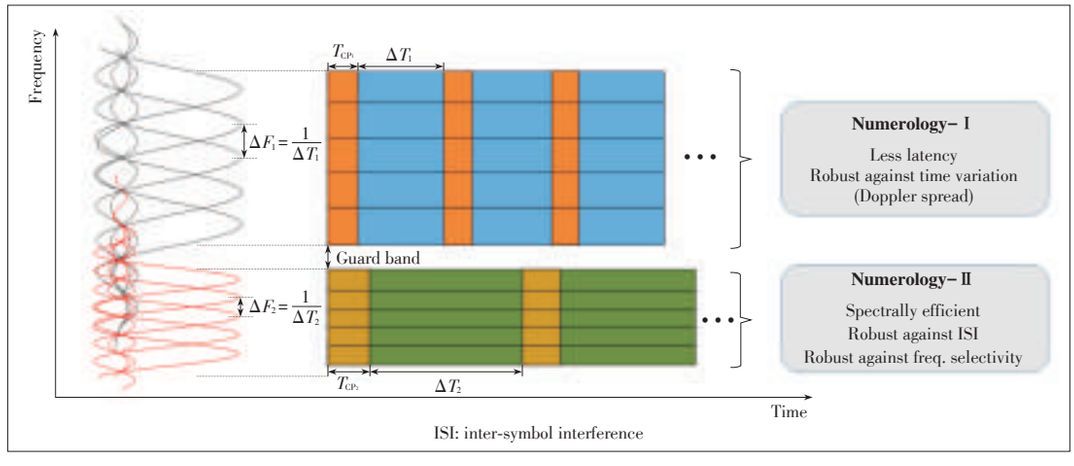


Figure 4. Illustration of different numerologies.

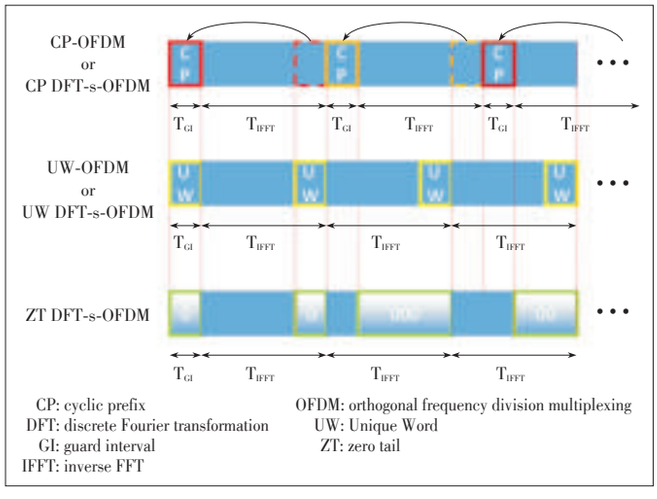
In the light of aforementioned facts, the coexistence of different numerologies offers a great advantage in serving users with different requirements. However, such a design obviously removes the orthogonality between the numerologies, i.e., sinc shaped subcarriers with different spacings as illustrated in Fig. 4, and inter-numerology interference becomes inevitable. This is a major issue in numerology design and guard band determination between numerologies. Therefore, for the sake of communication performance and spectral efficiency, minimization of OOB leakage of each numerology or keeping their orthogonality with various methods should be investigated carefully.

5 Other OFDM-Based Signaling Schemes

While many researchers take a stand on enhancing the characteristics of OFDM without changing its conventional structure, alternative OFDM based approaches, e.g., Unique Word-OFDM (UW-OFDM), discrete Fourier transformation-spread-OFDM (DFT-s-OFDM), etc., are also very popular. In this section, we will discuss the advantages and disadvantages of these technologies, illustrated in **Fig. 5**, along with a comparison with the classical CP-OFDM.

5.1 Utilizing GI with UW

In Subsection 3.1, we have denoted that CP is a smart way of utilizing the GI. There are other methods that utilize the GI as good as CP, which have recently been gaining attention. The most prominent one among these methods is UW-OFDM [48]. In UW-OFDM, the GI is filled with a deterministic sequence called the unique word. To obtain the UW, data subcarriers are multiplied by a precoding matrix, which depends on the desired UW, before the IFFT operation. This multiplication generates data dependent redundant subcarriers. Then, these redundant subcarriers are given as input to the IFFT along with the data symbols. At the output of the IFFT, UW is obtained in the time symbol, without needing any other operations such as copying and pasting as performed for CP.



▲**Figure 5. The modified structures' symbols in time below the OFDM symbol in time, to scale according to LTE extended CP specifications.**

The advantages of UW-OFDM comes from the fact that the UW is a natural part of the IFFT interval. Due to this property, symbols with different unique words can be multiplexed in time and frequency without destroying the orthogonality among various users and applications as long as the IDFT length is kept the same. Therefore, UW length can be adjusted by selecting the number of redundant subcarriers and UW-OFDM symbols can be flexibly designed based on the delay spreads experienced by different receivers [49]. Furthermore, the corresponding correlation introduced between the redundant and data subcarriers can be exploited to improve the BER performance [50]. However, due to the increased complexity stemmed from the precoding and decoding processes, UW-OFDM suffers from the complicated receiver and transmitter structures [51].

5.2 Quasi-Single Carrier Structures

DFT-s-OFDM can be obtained by adding an M-DFT block before the conventional N-IFFT operation where $M < N$. It is a midway between multicarrier and single carrier (SC), and is usually categorized as a quasi-single carrier structure due to this generation process. There are two main reasons this well-known modification of OFDM has been used in the UL of LTE [52]. Firstly, although higher from pure SC, it exhibits lower PAPR compared to the CP-OFDM and requires much lower power amplifier back-off resulting in a higher power efficiency. Secondly, since it is an OFDM-based structure, the scheduling flexibility provided by orthogonal frequency division multiple access (OFDMA) can still be used [53].

The circularity of DFT-s-OFDM symbols is also satisfied with the help of CP just like the conventional OFDM implemented in LTE. Recently, methods that fill the GI with different sequences have also been proposed for DFT-s-OFDM. Despite the similarity to the UW-OFDM approach discussed in the previous subsection, filling the GI with specific sequences does not require any precoding operation in DFT-s-OFDM because of its inherent structure [54]. When the sequence, desired to fill the GI, is appended to the data symbols at the input of the M-DFT, the interpolated form of this sequence is obtained at the output of the N-IFFT at no expense of complexity. Furthermore, these schemes can be used by existing DFT-s-OFDM receivers without any modifications as the guard sequences do not impact the data symbols [55]. Considering these facts, a popular alternative to the CP-based DFT-s-OFDM is zero tail (ZT)-based DFT-s-OFDM [56]. The main motivation behind using a ZT is the ability of adaptation to different channel conditions and data rates just by modifying the number of zeroes [57]. Compared to conventional CP-DFT-s-OFDM, this scheme offers a better BLER performance and reduced OOB leakage as the interference power leaking to the consecutive symbols is reduced and the zeroes are a natural

part of the IFFT output [57]. Having a ZT, however, decreases the average power of the transmitted signal, resulting in a PAPR penalty [55]. This penalty recently forced this approach to evolve into what is called Generalized DFT-s-OFDM, and ZT DFT-s-OFDM remained as a special case where the head and tail are set to zeroes [58].

The UW concept can also be combined with ZT DFT-s-OFDM and gives rise to UW DFT-s-OFDM. It replaces the ZT with nonzero low energy redundant symbols that further reduce the OOB leakage, PAPR and energy in the tail compared to both UW-OFDM and ZT-DFT-s-OFDM [55]. An enhanced version of UW DFT-s-OFDM concept is given in [59], where an additional perturbation signal is introduced to suppress the ISI energy between the consecutive symbols, which remains even less than the ISI between ZT-DFT-s-OFDM symbols. However, this scheme suffers from the increased receiver complexity and transmitter complexity due to the linear precoding [55].

6 Conclusions

In this paper, we discussed various aspects of OFDM as the strongest candidate waveform technology for 5G RAN. After providing a brief discussion on other candidate waveform schemes (FBMC, UFMC and GFDM), we firstly addressed the two major shortcomings of OFDM, high PAPR and OOB leakage, and reviewed the potential solutions for handling them individually or jointly. Then, we discussed the redundancies in OFDM, e.g., CP and pilots, and provided the proposed methods in the literature for reducing them. They are specifically critical for massive-MIMO applications and majority of the proposed techniques are presented in this context. Following the redundancy reduction techniques, we discussed the concept of numerology which plays an important role for 5G technologies in terms of delivering reliable service to the users with various requirements. Since OFDM is the most prominent waveform considered in the standard contributions, we just focused on how OFDM related parameters can meet different user requirements stemming from personal and environmental conditions, and from the types of provided service. However, authors believe the concept of numerology will evolve to a more general and flexible notion that encompasses any waveform technology, not limited to OFDM. We finally went through the other OFDM-based waveforms along with their pros and cons in comparison with classical OFDM.

References

[1] J. G. 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] FP7. (2012). *European Project 318555 5G NOW (5th Generation Non-Orthogonal Waveforms for Asynchronous Signaling)* [Online]. Avail- able: http://www.5gnow.eu

[3] FP7. (2012). *European Project 317669 METIS (Mobile and Wireless Communications Enablers for the Twenty-Twenty Information Society)* [Online]. Available: https://www.metis2020.com

[4] FP7. (2016). *European Project (FP7-ICT-619563) miWaveS (Beyond 2020 Heterogeneous Wireless Networks with Millimeter-Wave Small Cell Access and Backhauling)* [Online]. Available: http://www.miwaves.eu

[5] Fantastic 5G. (2016). *Horizon 2020 project (ICT-671660) FANTASTIC-5G (Flexible Air Interface for Scalable Service Delivery within Wireless Communication Networks of the 5th Generation)* [Online]. Available: http://fantastic5g.eu

[6] S. Methley, W. Webb, S. Walker, and J. Parker, "5G candidate band study: study on the suitability of potential candidate frequency bands above 6 GHz for future 5G mobile broadband systems," Quotient Associates Ltd, Tech. Rep., 2015.

[7] B. Farhang Boroujeny, "Filter bank multicarrier modulation: a waveform candidate for 5G and beyond," *Advances in Electrical Engineering*, vol. 2014, Dec. 2014. doi:10.1155/2014/482805

[8] A. Şahin, I. Güvenç, 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, Aug. 2014. doi: 10.1109/SURV.2013.121213.00263.

[9] F. Hu, *Opportunities in 5G Networks: A Research and Development Perspective*. Boca Raton, USA: CRC Press, 2016.

[10] 5G Forum. (2016, Mar.). *5G white paper: 5G vision, requirements, and enabling technologies* [Online]. Available: http://kani.or.kr/5g/whitepaper/5G%20Vision,%20Requirements,%20and%20Enabling%20Technologies.pdf

[11] 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, 2014. doi: 10.1109/MCOM.2014.6736749.

[12] J. Li, E. Bala, and R. Yang, "Resource block filtered-OFDM for future spectrally agile and power efficient systems," *Physical Communication*, vol. 11, pp. 36–55, 2014. doi: 10.1016/j.phycom.2013.10.003.

[13] I. F. Akyildiz, S. Nie, S. Lin, and M. Chandrasekaran, "5G roadmap: 10 key enabling technologies," *Computer Networks*, vol. 106, pp. 17 – 48, 2016. doi: 10.1016/j.comnet.2016.06.010.

[14] G. Fettweis, M. Krondorf, and S. Bittner, "GFDM—generalized frequency division multiplexing," in *Proc. IEEE 69th Vehicular Technology Conference*, Barcelona, Spain, Apr. 2009, pp. 1–4. doi: 10.1109/VETECS.2009.5073571.

[15] R. Datta, G. Fettweis, Z. Kollár, and P. Horváth, "FBMC and GFDM interference cancellation schemes for flexible digital radio PHY design," in *Proc. 14th Euromicro Conference on Digital System Design*, Washington DC, USA, 2011, pp. 335–339. doi:10.1109/DSD.2011.48.

[16] Qualcomm Inc., "Waveform candidates," 3GPP Standard Contribution (R1-162199), Busan, Korea, Apr. 2016.

[17] Huawei and HiSilicon, "5G waveform: requirements and design principles," 3GPP Standard Contribution (R1-162151), Busan, Korea, Apr. 2016.

[18] Huawei and HiSilicon, "F-OFDM scheme and filter design," 3GPP Standard Contribution (R1-165425), Nanjing, China, May 2016..."

[19] W. Jiang and M. Schellmann, "Suppressing the out-of-band power radiation in multi-carrier systems: A comparative study," in *Proc. IEEE Global Telecommunications Conference*, Anaheim, USA, Dec. 2012, pp. 1477–1482. doi: 10.1109/GLOCOM.2012.6503322.

[20] T. Weiss, J. Hillenbrand, A. Krohn, and F. K. Jondral, "Mutual interference in OFDM-based spectrum pooling systems," in *Proc. 59th IEEE Vehicular Technology Conference*, Milan, Italy, May 2004, vol. 4, pp. 1873–1877. doi: 10.1109/VETECS.2004.1390598.

[21] A. Sahin and H. Arslan, "Edge windowing for OFDM based systems," *IEEE Communications Letters*, vol. 15, no. 11, pp. 1208–1211, Nov. 2011. doi: 10.1109/LCOMM.2011.090611.111530.

[22] S. Brandes, I. Cosovic, and M. Schnell, "Reduction of out-of-band radiation in OFDM systems by insertion of cancellation carriers," *IEEE Communications Letters*, vol. 10, no. 6, pp. 420–422, Jun. 2006. doi: 10.1109/LCOMM.2006.1638602.

[23] D. Qu, Z. Wang, and T. Jiang, "Extended active interference cancellation for sidelobe suppression in cognitive radio OFDM Systems with cyclic prefix," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1689–1695, May 2010. doi: 10.1109/TVT.2010.2040848.

[24] I. Cosovic, S. Brandes, and M. Schnell, "Subcarrier weighting: a method for sidelobe suppression in OFDM systems," *IEEE Communications Letters*, vol. 10, no. 6, pp. 444–446, Jun. 2006. doi: 10.1109/LCOMM.2006.1638610.

[25] A. Tom, A. Sahin, and H. Arslan, "Mask compliant precoder for OFDM spectrum shaping," *IEEE Communications Letters*, vol. 17, no. 3, pp. 447–450, Mar. 2013. doi: 10.1109/LCOMM.2013.020513.122495.

[26] Y. Rahmatallah and S. Mohan, "Peak-to-average power ratio reduction in OFDM systems: a survey and taxonomy," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1567–1592, Fourth Quarter 2013. doi: 10.1109/

SURV.2013.021313.00164.

[27] A. Ghassemi, L. Lampe, A. Attar, and T. A. Gulliver, "Joint sidelobe and peak power reduction in OFDM-based cognitive radio," in *Proc. IEEE 72nd Vehicular Technology Conference*, San Francisco, USA, pp. 1–5, Sept. 2010. doi: 10.1109/VETECF.2010.5594133.

[28] E. Güvenkaya, A. Tom, and H. Arslan, "Joint sidelobe suppression and PAPR reduction in OFDM using partial transmit sequences," in *Proc. IEEE Military Communications Conference*, San Diego, USA, Nov. 2013, pp. 95–100. doi: 10.1109/MILCOM.2013.26.

[29] C. Ni, T. Jiang, and W. Peng, "Joint PAPR reduction and sidelobe suppression using signal cancelation in NC-OFDM-based cognitive radio systems," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 3, pp. 964–972, Mar. 2015. doi: 10.1109/TVT.2014.2327012.

[30] A. Tom, A. Sahin, and H. Arslan, "Suppressing alignment: An approach for out-of-band interference reduction in OFDM systems," in *Proc. IEEE International Conference on Communications*, London, UK, Jun. 2015, pp. 4630–4634. doi: 10.1109/ICC.2015.7249053.

[31] M. Maso, M. Debbah, and L. Vangelista, "A distributed approach to interference alignment in OFDM-based two-tiered networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 1935–1949, Jun. 2013. doi: 10.1109/TVT.2013.2245516.

[32] Z. E. Ankarali, A. Sahin, and H. Arslan, "Static cyclic prefix alignment for OFDM-based waveforms," in *Proc. IEEE Global Communications Conference and Workshops*, Washington, USA, Dec. 2016. doi: 10.1109/GLOCOMW.2016.7849051.

[33] Z. E.Ankarali, and H. Arslan. "Joint physical layer security and PAPR mitigation in OFDM systems," U.S. Patent No. 9,479,375. Oct. 2016.

[34] S.Rajagopal, S. Abu Surra, A. Gupta, et al., "Methods and apparatus for cyclic prefix reduction in mmwave mobile communication systems," U.S. Patent US20 130 315 321 A1, Nov. 2013.

[35] E. Zoechmann, S. Pratschner, S. Schwarz, and M. Rupp, "MIMO transmission over high delay spread channels with reduced cyclic prefix length," in *Proc. 19th International ITG Workshop on Smart Antennas (WSA)*, Ilmenau, Germany, Mar. 2015, pp. 1–8.

[36] T. Pham, T. Le Ngoc, G. Woodward, P. A. Martin, and K. T. Phan, "Equalization for MIMO-OFDM systems with insufficient cyclic prefix," in *Proc. IEEE 83th Vehicular Technology Conference*, Nanjing, China, May 2016, pp. 1–5. doi: 10.1109/VTCSpring.2016.7504240.

[37] A. Sahin and H. Arslan, "Multi-user aware frame structure for OFDMA based system," in *Proc. IEEE 76th Vehicular Technology Conference*, Quebec City, Canada, Sept. 2012, pp. 1–5. doi:10.1109/VTCFall.2012.6399155.

[38] ZTE, "Consideration of cyclic prefix for NR," 3GPP Standard Contribution (R1-166406), Gothenburg, Sweden, Aug. 2016."

[39] ZTE, "Support of Multiple CP Families for NR," 3GPP Standard Contribution (R1-1608962), Lisbon, Portugal, Oct. 2016."

[40] J. Lorca, "Cyclic prefix overhead reduction for low-latency wireless communications in OFDM," in *Proc. IEEE 81st Vehicular Technology Conference*, Glasgow, UK, May 2015, pp. 1–5. doi: 10.1109/VTCSpring.2015.7145767.

[41] O. Elijah, C. Y. Leow, T. A. Rahman, S. Nunoo, and S. Z. Iliya, "A comprehensive survey of pilot contamination in massive MIMO–5G system," *IEEE Communications Surveys & Tutorials.*, vol. 18, no. 2, pp. 905–923, 2016. doi: 10.1109/COMST.2015.2504379.

[42] P. Walk, H. Becker, and P. Jung, "OFDM channel estimation via phase retrieval," in *Proc. 49th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, USA, Nov. 2015, pp. 1161–1168. doi: 10.1109/ACSSC. 2015. 7421323.

[43] V. Saxena, "Pilot contamination and mitigation techniques in massive MIMO systems," M.S. Thesis, Lund University, Stockholm, Sweden, 2014.

[44] X. Luo, X. Zhang, H. Qian, and K. Kang, "Pilot decontamination via PDP alignment," in *Proc. IEEE Global Communications Conference*, Washington DC, USA, Dec. 2016, pp. 1-6. doi: 10.1109/GLOCOM.2016.7842147.

[45] R. R. Müller, L. Cottatellucci, and M. Vehkaperä, "Blind pilot decontamination," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 773–786, Oct. 2014. doi: 10.1109/JSTSP.2014.2310053.

[46] D. Hu, L. He, and X. Wang, "Semi-blind pilot decontamination for massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 525–536, Jan. 2016. doi: 10.1109/TWC.2015.2475745.

[47] A. A. Zaidi, R. Baldemair, H. Tullberg, et al., "Waveform and numerology to support 5G services and requirements," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 90–98, Nov. 2016. doi: 10.1109/MCOM.2016.1600336CM.

[48] C. Hofbauer, M. Huemer, and J. B. Huber, "Coded OFDM by unique word prefix," in *Proc. IEEE International Conference on Communication Systems*, Singapore, Singapore, Nov. 2010, pp. 426–430. doi: 10.1109/ICCS.2010.5686520.

[49] M. Huemer, C. Hofbauer, and J. B. Huber, "Non-systematic complex number RS coded OFDM by unique word prefix," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 285–299, Jan. 2012. doi: 10.1109/TSP.2011.2168522.

[50] A. Onic and M. Huemer, "Noise interpolation for unique word OFDM," *IEEE Signal Processing Letters*, vol. 21, no. 7, pp. 814–818, Jul. 2014. doi: 10.1109/LSP.2014.2317512.

[51] M. Huemer, A. Onic, and C. Hofbauer, "Classical and Bayesian linear data estimators for unique word OFDM," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6073–6085, Dec. 2011. doi: 10.1109/TSP.2011.2164912.

[52] H. Holma and A. Toskala, Eds., *LTE for UMTS: Evolution to LTE- Advanced*, 2nd ed. Chichester, UK: Wiley, 2011.

[53] S. Sesia, I. Toufik, and M. P. J. Baker, Eds., *LTE - the UMTS Long Term Evolution: from Theory to Practice*. Chichester, UK: Wiley, 2009.

[54] C. F. Gauss, "Nachlass: Theoria interpolationis methodo nova tractata," pp. 265–303, *Carl Friedrich Gauss, Werke, Band 3*, Gottingen: Koniglichen Gesellschaft der Wissenschaften, 1866.

[55] A. Sahin, R. Yang, M. Ghosh, and R. L. Olesen, "An improved unique word DFT-spread OFDM scheme for 5G systems," in *Proc. IEEE Global Telecommunications Conference and Workshops*, San Diego, USA, Dec. 2015, pp. 1–6. doi: 10.1109/GLOCOMW.2015.7414173.

[56] G. Berardinelli, F. M. L. Tavares, T. B. Sorensen, P. Mogensen, and Pajukoski, "Zero-tail DFT-spread-OFDM signals," in *Proc. IEEE Global Telecommunications Conference and Workshops*, Atlanta, USA, Dec. 2013, pp. 229–234. doi: 10.1109/GLOCOMW.2013.6824991.

[57] G. Berardinelli, F. Tavares, T. Sorensen, P. Mogensen, and K. Pajukoski, "On the potential of Zero-Tail DFT-spread-OFDM in 5G networks," in *Proc. IEEE 80th Vehicular Technology Conference*, Sept. 2014, pp. 1–6. doi: 10.1109/VTCFall.2014.6966089.

[58] G. Berardinelli, K. I. Pedersen, T. B. Sorensen, and P. Mogensen, "Generalized DFT-spread-OFDM as 5G waveform," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 99–105, Nov. 2016. doi: 10.1109/MCOM.2016.1600313CM.

[59] A. Sahin, R. Yang, E. Bala, M. C. Beluri, and R. L. Olesen, "Flexible DFT-S-OFDM: solutions and challenges," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 106–112, Nov. 2016. doi: 10.1109/MCOM.2016.1600330CM.

Manuscript received: 2016-11-30

Biographies

Zekeriyya Esat Ankaralı (zekeriyya@mail.usf.edu) received his B.Sc. degree in control engineering from Istanbul Technical University (ITU), Turkey in 2011 with honors degree and M.Sc. in electrical engineering from University of South Florida (USF), USA in December 2012. Since January 2013, he has been pursuing his Ph. D. as a member of the Wireless Communication and Signal Processing (WCSP) Group at USF. His current research interests are waveform design, multicarrier systems, physical layer security and in vivo communications.

Berker Peköz (pekoz@mail.usf.edu) received his B.Sc. degree in electrical and electronics engineering from Middle East Technical University (METU), Turkey in 2015 with high honors degree. Since August 2015, he has been pursuing his Ph.D. as a member of the Wireless Communication and Signal Processing (WCSP) Group at University of South Florida (USF), USA. His current research interests are mmWave communications, multidimensional modulations and waveform design.

Huseyin Arslan (arslan@usf.edu) received his B.S. degree from Middle East Technical University (METU), Turkey in 1992; M.S. and Ph.D. degreesfrom Southern Methodist University (SMU), USA in 1994 and 1998. From January 1998 to August 2002, he was with the research group of Ericsson Inc., USA, where he was involved with several project related to 2G and 3G wireless communication systems. Since August 2002, he has been with the Electrical Engineering Dept. of University of South Florida (USF), USA. Also, he has been the dean of the College of Engineering and Natural Sciences of Istanbul Medipol University, Turkey since 2014. In addition, he has worked as part time consultant for various companies and institutions including Anritsu Company (USA) and The Scientific and Technological Research Council of Turkey. His current research interests include physical layer security, mmWave communications, small cells, multi-carrier wireless technologies, co-existence issues on heterogeneous networks, aeronautical (high altitude platform) communications and in vivo channel modeling, and system design.

DOI: 10.3969/j. issn. 1673-5188. 2017. S1. 003
http://kns.cnki.net/kcms/detail/34.1294.TN.20170606.1856.004.html, published online June 6, 2017

An Overview of Non-Orthogonal Multiple Access

Anass Benjebbour
(NTT DOCOMO, INC., Kanagawa 239-8536, Japan)

Abstract

In recent years, non-orthogonal multiple access (NOMA) has attracted a lot of attention as a novel and promising power-domain user multiplexing scheme for Long-Term Evolution (LTE) enhancement and 5G. NOMA is able to contribute to the improvement of the tradeoff between system capacity and user fairness (i.e., cell-edge user experience). This improvement becomes in particular emphasized in a cellular system where the channel conditions vary significantly among users due to the near-far effect. In this article, we provide an overview of the concept, design and performance of NOMA. In addition, we review the potential benefits and issues of NOMA over orthogonal multiple access (OMA) such as orthogonal frequency division multiple access (OFDMA) adopted by LTE, and the status of 3GPP standardization related to NOMA.

Keywords

multiple access; non-orthogonal multiple access (NOMA); power-domain; multi-user detection; MUST

1 Introduction

Significant gains in system capacity and quality of user experience (QoE) are required to respond to the anticipated exponential increase in the volume of mobile traffic in the next decade and the merge of enhanced mobile broadband (eMBB) services [1]. In cellular mobile communications, the design of the radio access technology (RAT) is one important aspect for improving system capacity in a cost-effective manner. Radio access technologies are typically characterized by the radio frame design, waveform design, multiple-input and multiple-output (MIMO) transmission scheme, and multiple access scheme. In particular, the design of the multiple access scheme is of great importance from a system perspective, since it provides the means for multiple users to access and share the system resources efficiently and simultaneously, e.g., frequency division multiple access (FDMA), time division multiple access (TDMA), code division multiple access (CDMA), and orthogonal frequency division multiple access (OFDMA). In the 3.9G and 4G mobile communication systems such as Long-Term Evolution (LTE) and LTE-Advanced [2], standardized by the 3rd Generation Partnership Project (3GPP), orthogonal multiple access (OMA) based on OFDMA for downlink and single carrier (SC)-FDMA for uplink are adopted. Orthogonal multiple access is a good choice for achieving good system-level throughput performance in packet-domain services with a simplified receiver design. However, non-orthogonal designs become of interest toward further enhancement of the system efficiency and QoE especially at the cell edge.

Recently, there have been several investigations on advanced schemes for non-orthogonal signal transmission within a user and non-orthogonal user multiplexing among multiple users. For example, Faster-than-Nyquist (FTN) signaling [3] is one approach for non-orthogonal signal transmission within a user by exploiting the excess bandwidth of the signal. Interleaved division multiple access (IDMA), where the channelization of respective user is achieved by the user-specific channel interleaver and multiuser detection at the receiver, is investigated to accommodate a large number of low-rate users [4], [5]. However, these schemes do not exploit the channel difference among users and generally require high complexity receivers for signal separation.

As a novel multiple access approach, a non-orthogonal multiple access (NOMA) scheme was proposed by NTT DOCOMO [6]–[18] including the author of this article. In the proposed NOMA, multiple users of different channel conditions are multiplexed in the power-domain on the transmitter side and multi-user signal separation on the receiver side is conducted. From an information-theoretic perspective, it is well-known that by using superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver, non-orthogonal user multiplexing not only outperforms orthogonal multiplexing, but also can achieve the capacity region of the downlink broadcast channel [12], [19], [20]. NOMA can be also applied to the uplink (multiple access channel) [12], [15], [19]. For uplink, al-

though both NOMA and OMA can achieve the capacity region, NOMA also provides improvements in the tradeoff of system capacity and user fairness (i.e., the cell-edge user experienced data rate) [12], [19].

Assuming a proportional fairness (PF) scheduler, the performance of NOMA has been heavily investigated for downlink and uplink from the system-level and link-level perspectives [6]–[18]. In addition, the transmitter and receiver designs for NOMA were considered for both closed-loop and open-loop MIMO and for both successive interference cancellation and non-SIC receivers [16], [17]. When applied to either downlink or uplink, NOMA is shown able to contribute to the improvement of the tradeoff between system capacity and user fairness. This improvement becomes in particular emphasized in a cellular system where the channel conditions vary significantly among users due to the near-far effect.

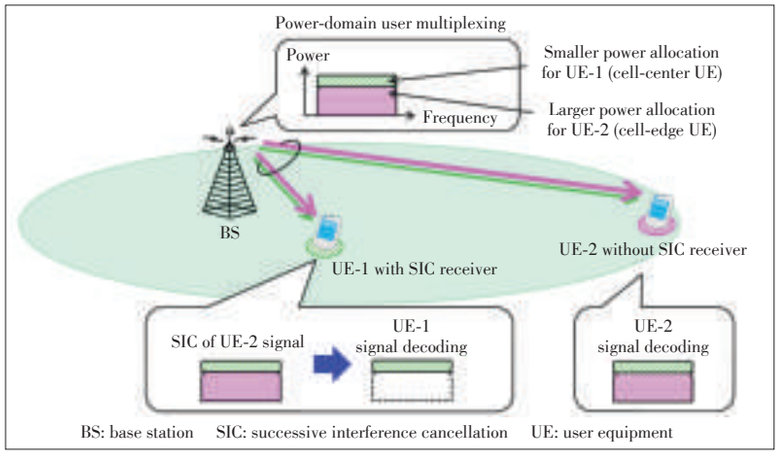
In this article, we introduce an overview of the concept, the design with a combination of MIMO and the performance of NOMA. We also review the potential benefits and issues of NOMA over orthogonal multiple access, and the status of standardization related to downlink NOMA, known as multi-user superposition transmission (MUST) in 3GPP.

The rest of this paper is organized as follows. Section 2 describes the concept. Section 3 discusses the expected benefits and issues of NOMA. Section 4 explains the combination of NOMA with MIMO. Section 5 reviews the performance of NOMA from link-level and system-level evaluations and trial results. In addition, the status of NOMA standardization in 3GPP LTE Release 14 is briefly summarized. Finally, Section 6 concludes the paper.

2 NOMA Concept

2.1 Downlink

Fig. 1 illustrates downlink NOMA with SIC for the case of one base station (BS) and two user equipments (UEs).



▲**Figure 1. Illustration of downlink NOMA with SIC.**

For the sake of simplicity, we assume in the following descriptions the case of single transmit and receive antennas. The overall system transmission bandwidth is assumed to be 1 Hz. The base station transmits a signal for UE-i (i = 1, 2), x_i, where $E[|x_i|^2] = 1$, with transmit power P_i and the sum of P_i is equal to P. In NOMA, x_1 and x_2 are superposed in the power-domain as follows:

$$x = \sqrt{P_1}x_1 + \sqrt{P_2}x_2 . \quad (1)$$

Thus, the received signal at UE-i is represented as

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

where h_i is the complex channel coefficient between UE-i and the BS. The variable w_i denotes additive white Gaussian noise (AWGN) including inter-cell interference. The power spectral density of w_i is $N_{0,i}$. In downlink NOMA, the SIC process is implemented at the UE receiver for the case where the decoding of the signal of desired UE and that of the superposed signals of other UEs are needed. The optimal order for SIC decoding is in the order of decreasing channel gain normalized by noise and inter-cell interference power, $|h_i|^2/N_{0,i}$ (called as simply channel gain in the following). Given this decoding order and assuming that any user can correctly decode the signals of other users whose decoding order comes before the corresponding user, each UE-i can remove the inter-user interference from the j-th user whose $|h_j|^2/N_{0,j}$ is lower than $|h_i|^2/N_{0,i}$. In a 2-UE case, assuming that $|h_1|^2/N_{0,1} > |h_2|^2/N_{0,2}$, UE-2 does not perform interference cancellation since it comes first in the decoding order. UE-1 first decodes x_2 and subtracts its component from the received signal y_1, then next, x_1 is decoded without interference from x_2. Assuming successful decoding and no error propagation, the throughput of UE-i, R_i, can be represented as

$$R_1 = \log_2\left(1 + \frac{P_1|h_1|^2}{N_{0,1}}\right), R_2 = \log_2\left(1 + \frac{P_2|h_2|^2}{P_1|h_2|^2 + N_{0,2}}\right). \quad (3)$$

From (3), it can be seen that power allocation for each UE greatly affects the user throughput performance and thus the modulation and coding scheme (MCS) used for data transmission of each UE. By adjusting the power allocation ratio, P_1/P_2, the BS can flexibly control the throughput of each UE and also optimize tradeoff between the system capacity and user fairness. By flexibly adjusting power allocation, the BS can control the throughput of each UE such that the signal designated to each UE is decodable at its corresponding receiver. Also, since the channel gain of the cell-center UE is higher than cell-edge UE, as long as the cell-edge UE signal is decodable at cell-edge UE receiver, its decoding at the cell-center UE receiver can be successful with high probability.

For OMA as orthogonal user multiplexing, the bandwidth of α Hz ($0 < \alpha < 1$) is assigned to UE-1

and the remaining bandwidth, 1—α Hz, is assigned to UE-2. The throughput of UE-i, R_i, is represented as

$$R_1 = \alpha \log_2\left(1+\frac{P_1|h_1|^2}{\alpha N_{0,1}}\right), R_2 = (1-\alpha)\log_2\left(1+\frac{P_2|h_2|^2}{(1-\alpha)N_{0,2}}\right). \quad (4)$$

In NOMA, the performance gain compared to OMA increases when the difference in channel gains, e.g., path loss between UEs, is large. For example, as shown in **Fig. 2**, we assume a 2-UE case with a cell-interior UE and a cell-edge UE, where $|h_1|^2/N_{0,1}$ and $|h_2|^2/N_{0,2}$ are set to 20 dB and 0 dB, respectively. For OMA with equal bandwidth and equal transmission power are allocated to each UE (a = 0.5, $P_1 = P_2 = 1/2P$), the user rates are calculated according to (4) as R_1 = 3.33 bps and R_2 = 0.50 bps, respectively. On the other hand, for NOMA, when the power allocation is conducted as $P_1 = 1/5P$ and $P_2 = 4/5P$, the user rates are calculated according to (3) as R_1 = 4.39 bps and R_2 = 0.74 bps, respectively. The corresponding gains of NOMA over OMA are 32% and 48% for UE-1 and UE-2, respectively. According to this example where a 20 dB signal-to-noise ratio (SNR) difference between the 2 UEs is assumed, it can be seen that NOMA provides a higher sum rate than OMA.

2.2 Uplink

Fig. 3 illustrates uplink NOMA where two UEs transmit signals to the BS on the same frequency resource and at the same time, and SIC is conducted at BS for UE multi-user signal separation. Similar to downlink, we assume the case of single transmit and receive antennas, and the overall system transmission bandwidth is 1 Hz. The signal transmitted by UE-i (i = 1, 2) is denoted as x_i, where $E[|x_i|^2] = 1$, with transmit power P_i. In uplink NOMA, the received signal at BS is a superposed signal of x_1 and x_2 as follows:

$$y = h_1\sqrt{P_1}x_1 + h_2\sqrt{P_2}x_2 + w, \quad (5)$$

where h_i denotes the complex channel coefficient between UE-i and the BS. The variable w denotes inter-cell interference and noise observed at the BS with a power spectral density of N_0. We assume UE-1 is the cell-center user and UE-2 is the cell-edge user, i.e. $|h_1|^2/N_0 > |h_2|^2/N_0$, and the BS conducts SIC according to the descending order of channel gains. The throughput of UE-i, denoted as R_i, assuming no error propagation can be calculated as

$$R_1 = \log_2\left(1+\frac{P_1|h_1|^2}{P_2|h_2|^2+N_0}\right), R_2 = \log_2\left(1+\frac{P_2|h_2|^2}{N_0}\right). \quad (6)$$

If the BS conducts SIC according to the ascending order of channel gains, the throughput of UE-i can be calculated as

$$R_1 = \log_2\left(1+\frac{P_1|h_1|^2}{N_0}\right), R_2 = \log_2\left(1+\frac{P_2|h_2|^2}{P_1|h_1|^2+N_0}\right). \quad (7)$$

Interestingly, the total UE throughput is the same regardless of the SIC order of descending order or ascending order of channel gain, i.e.

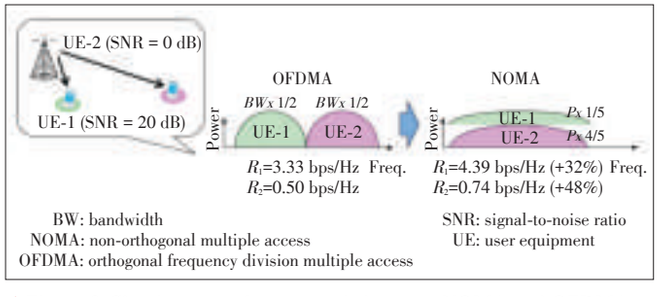
$$R_1 + R_2 = \log_2\left(1+\frac{P_1|h_1|^2+P_2|h_2|^2}{N_0}\right). \quad (8)$$

However, the conclusion that the total UE throughputs of different SIC orders are equal only holds under the assumption of no error propagation. In practical systems where we have error propagation, the best SIC order is in the decreasing order of channel gains.

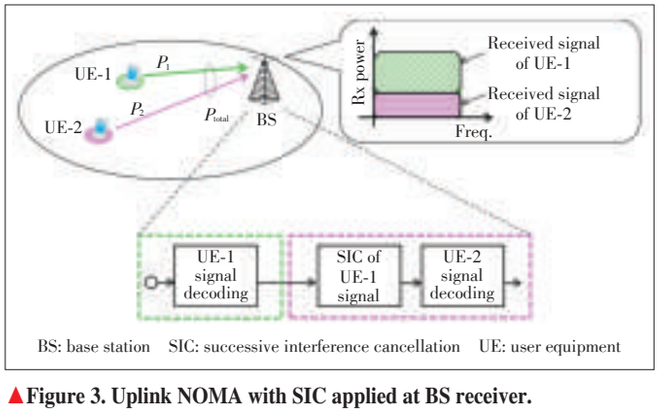
For OMA, we assume the bandwidth of α Hz ($0 < \alpha < 1$) is assigned to UE-1 and the remaining bandwidth, 1—α Hz, is assigned to UE-2. The throughput of UE-i can be calculated as

$$R_1 = \alpha \log_2\left(1+\frac{P_1|h_1|^2}{\alpha N_0}\right), R_2 = (1-\alpha)\log_2\left(1+\frac{P_2|h_2|^2}{(1-\alpha)N_0}\right). \quad (9)$$

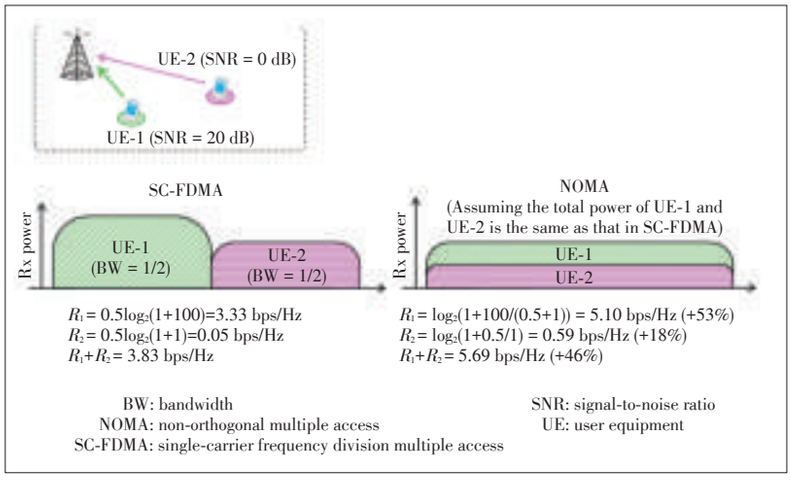
One comparison example of OMA and NOMA is shown in **Fig. 4** by assuming a 2-UE case with a cell-center UE and a cell-edge UE, where $|h_1|^2/N_0$ and $|h_2|^2/N_0$ are set to 20 dB and 0 dB, respectively. For OMA, we assume equal bandwidth is allocated to each UE (i.e., a = 0.5), the user rates are calculated according to (9) as R_1 = 3.33 bps and R_2 = 0.50 bps, respectively. On the other hand, in NOMA the total transmission power of each UE is assumed the same as that in OMA, the user rates are calculated according to (6) as R_1 = 5.10 bps and R_2 = 0.59



▲**Figure 2. Simple comparison example between NOMA and OFDMA for downlink.**



▲**Figure 3. Uplink NOMA with SIC applied at BS receiver.**



▲**Figure 4. Simple comparison example of NOMA and SC-FDMA for uplink.**

bps, respectively. The total UE throughput gain of NOMA over OMA is 46%. Therefore, for uplink NOMA, we can obtain similar performance gain as that for downlink NOMA.

3 Expected Benefits and Issues of NOMA

3.1 Benefits

NOMA is a promising multiple access scheme for the future owing to the following expected benefits.

1) Exploitation of channel gain difference among users

Unlike OMA (OFDMA) where the channel gain difference among users is translated into multi-user diversity gains via frequency-domain scheduling, in NOMA the channel gain difference is translated into multiplexing gains by superposing in the power-domain the transmit signals of multiple users of different channel gains. As shown in Figs. 1 and 2, by exploiting the channel gain difference in downlink NOMA, both UEs of high and low channel gains are in a win-win setup. Indeed, UEs with high channel gain (bandwidth-limited UEs) lose a little by being allocated less power, but can gain much more by being allocated more bandwidth, while UEs with low channel gain (power-limited UEs) also lose only a little by being allocated little less power and "effective" bandwidth (because of being interfered by the signal designated to the other UEs with high channel gain) but gain much more by being allocated more bandwidth. This win-win situation is also the main reason why NOMA gains over OMA increase when the difference in channel gains between NOMA paired UEs becomes larger [8].

2) Intentional non-orthogonality via power-domain user multiplexing and advanced receiver processing

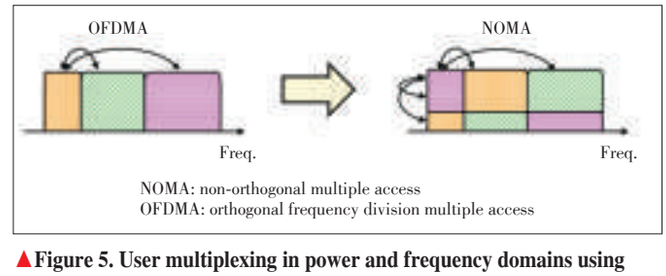
NOMA is a multiplexing scheme that utilizes an additional new domain, i.e., the power domain, which is not sufficiently utilized in previous systems. For downlink NOMA, non-orthogonality is intentionally introduced via power-domain user multiplexing as shown in **Fig. 5**; however, quasi-orthogonality can still be achieved. In fact, user demultiplexing is ensured via the allocation of large power difference between paired UEs and the application of SIC in power-domain. The UE with high channel gain (e.g., UE-1 in Figs. 1 and 2) is allocated less power and the UE with low channel gain (e.g., UE-2 in Figs. 1 and 2) is allocated more power. Such large power difference facilitates the successful decoding and the cancellation of the signal designated to UE-2 (being allocated high power) at UE-1 receiver and thus less complex receivers such as SIC can be used. In addition, at UE-2 receiver, the signal designated to UE-2 is decoded directly by treating the interference from the signal designated to UE-1 (being allocated low power) as noise.

On the other hand, NOMA captures well the evolution of device processing capabilities, generally following Moore's law, by relying on more advanced receiver processing such as SIC. In the same spirit, but for the purpose of inter-cell interference mitigation, network-assisted interference cancellation and suppression (NAICS), including SIC, is being discussed in LTE Release 12 [21]. Thus, NOMA is in fact a natural direction to extend the work in 3GPP on NAICS in LTE Release 13 and beyond, as it should be much easier to apply advanced receiver to deal with intra-cell interference than inter-cell interference. Moreover, the issue of the increased overhead is common to both intra-cell and inter-cell SIC since the signaling the information related to the demodulation and decoding of other UEs is needed. The signaling overhead issue is discussed later.

3) Robust performance gain in practical wide area deployments and high mobility scenarios

NOMA relies on power-domain instead of spatial domain for user multiplexing. Therefore, the knowledge of the transmitter of the instantaneous frequency-selective fading channels such as the frequency-selective channel quality indicator (CQI) or channel state information (CSI) is mainly used at the receiver for user pairing and multi-user power allocation. Thus, NOMA does require less fine CSI feedback compared to multi-user MIMO (MU-MIMO) and a robust performance gain in practical wide area deployments can be expected irrespective of UE mobility or CSI feedback latency.

In [9], downlink NOMA is shown to maintain good gains



▲**Figure 5. User multiplexing in power and frequency domains using NOMA.**

compared to OMA in particular with wideband scheduling. Thus, NOMA can be a promising multiple access to provide a good robustness to mobility by mainly relying on receiver side CSI and signal processing.

3.2 Issues

In the following, we discuss several issues regarding downlink NOMA, such as signaling overhead and receiver design.

3.2.1 Signaling Overhead

1) Multi-user scheduling

For OFDMA, both subband and wideband multi-user scheduling can be generally considered for frequency-domain scheduling. For the case of LTE which adopts OFDMA, irrespective of subband or wideband scheduling, the same channel coding rate (including rate matching) and data modulation scheme are assumed over all the subbands allocated to each single user. Thus, MCS selection is always wideband. However, when NOMA is applied over LTE and the user pairing and power allocation are conducted over each subband, a mismatch occurs between MCS selection granularity (i.e., wideband) and power allocation granularity (i.e., subband). Such a mismatch prevents the full exploitation of NOMA gains [14]. Thus, MCS selection over each subband, if introduced in 5G, could be beneficial for NOMA. On the other hand, when the NOMA user pairing and power allocation are conducted over each subband, the signaling overhead increases linearly with the number of subbands. Therefore, considerations on signaling overhead and performance tradeoffs need to be taken into account in the design of NOMA.

2) Multi-user power allocation

Because of the power-domain user multiplexing of NOMA, the transmit power allocation (TPA) to one user affects the achievable throughput of that user and also the throughput of other users. The best performance of downlink NOMA can obviously be achieved by exhaustive full search of user pairs and dynamic transmit power allocations. In case of full search power allocation (FSPA), all possible combinations of power allocations are considered for each candidate user set. However, FSPA remains computationally complex. Moreover, with such dynamic TPA, the signaling overhead associated with decoding order and power allocation ratio increases significantly. In order to reduce the signaling overhead associated with multi-user transmit power allocation of NOMA and to clarify the degree of impact of user pairing on the performance of NOMA, both exhaustive and simplified user pairing and power allocation schemes were explored [9]. In NOMA, users with large channel gain difference (e.g., large path-loss difference) are paired with high probability; thus, considering practical implementations, user pairing and TPA, could be simplified by using pre-defined user grouping and fixed per-group power allocation (FPA), where users are divided into multiple user groups according to the magnitude of their channel gains using pre-defined thresholds or according to their selected MCS level [15]. Pre-defined user grouping and fixed TPA can be promising in practical usage when the potential saving in signaling overhead is taken into account. For example, the order of SIC and information on power assignment do not need to be transmitted in every sub-frame but rather on a longer time scale.

For uplink NOMA, since the user separation process is implemented at the base station, we do not see a significant increase in the signaling overhead. In addition, the conventional control signaling assumed in LTE or LTE-Advanced may be reused in a straightforward manner.

3.2.2 Receiver Design and Resource Alignment

In practice, the impact of the receiver on NOMA performance remains as one concern. For the cell-edge UE, advanced receiver technologies may not necessarily be applied since the received signal power for this UE is greater than that for the cell-center UE, i.e., interfering UE. On the other hand, in order to decode the received signal for the cell-center UE, the application of interference cancellation is inevitable since the signal for the cell-center UE is significantly contaminated by that for the cell-edge UE in the same time and frequency resources. There are two types of interference cancellation receivers: symbol-level interference cancellation (SLIC) and codeword level interference cancellation (CWIC). For both receivers, the received data symbols for the cell-edge UE are first de-modulated by multiplying the received signal with the maximal ratio combining (MRC) weight or minimum mean squared error (MMSE) receiver, then the Log-likelihood ratio (LLR) corresponding to those de-modulated symbols are calculated.

For the CWIC, a sequence of LLRs which is called codeword is input to the Turbo decoder and a sequence of posteriori-LLRs is generated. After interleaving the sequence of posteriori-LLRs, the interleaved LLRs are used to calculate a soft symbol replica for the cell-edge UEs. On the other hand, for the SLIC, those LLRs are directly used to generate a symbol replica for the cell-edge UE.

The decoding performance of CWIC is basically better than that of SLIC. However, it is important to note that resource alignment and transmission power alignment highly impact the system performance and largely affect the receiver complexity and signaling overhead. For example, resource alignment among the paired UEs would be needed to facilitate the CWIC; however such a scheduling restriction may degrade the system-level performance due to reduction in scheduler flexibility and thus in the gains of frequency-domain scheduling. Also, some limitations on the UE pairing for the retransmissions need to be taking into account. On the other hand, when SLIC (e.g., reduced complexity ML (R-ML)) is applied, such restrictions related to resource allocation and retransmission can be relaxed and the frequency-domain scheduling gain can be obtained [25]. These tradeoffs need to be taken into account in the re-

ceiver choice.

4 Combination of NOMA and MIMO

MIMO is one of the key technologies to improve spectrum efficiency in LTE/LTE-Advanced. In general, MIMO techniques can be categorized into single-user MIMO (SU-MIMO), where only one UE is served in data transmission, and MU-MIMO, where more than one UE are served in data transmission. Because MIMO technology exploits spatial domain and NOMA exploits power domain, these two technologies can be combined to further boost the system performance. In single-input single-output (SISO) and single-input multiple-output (SIMO) downlink, the broadcast channel is degraded where superposition coding with SIC and dirty paper coding (DPC) are equivalent and optimal from the viewpoint of the achievable capacity region. However, for the downlink MIMO case, the broadcast channel is non-degraded and the superposition coding with SIC receiver becomes non-optimal, although DPC remains optimal [12], [19], [20]. These aspects need to be taken into account when NOMA is combined with MIMO.

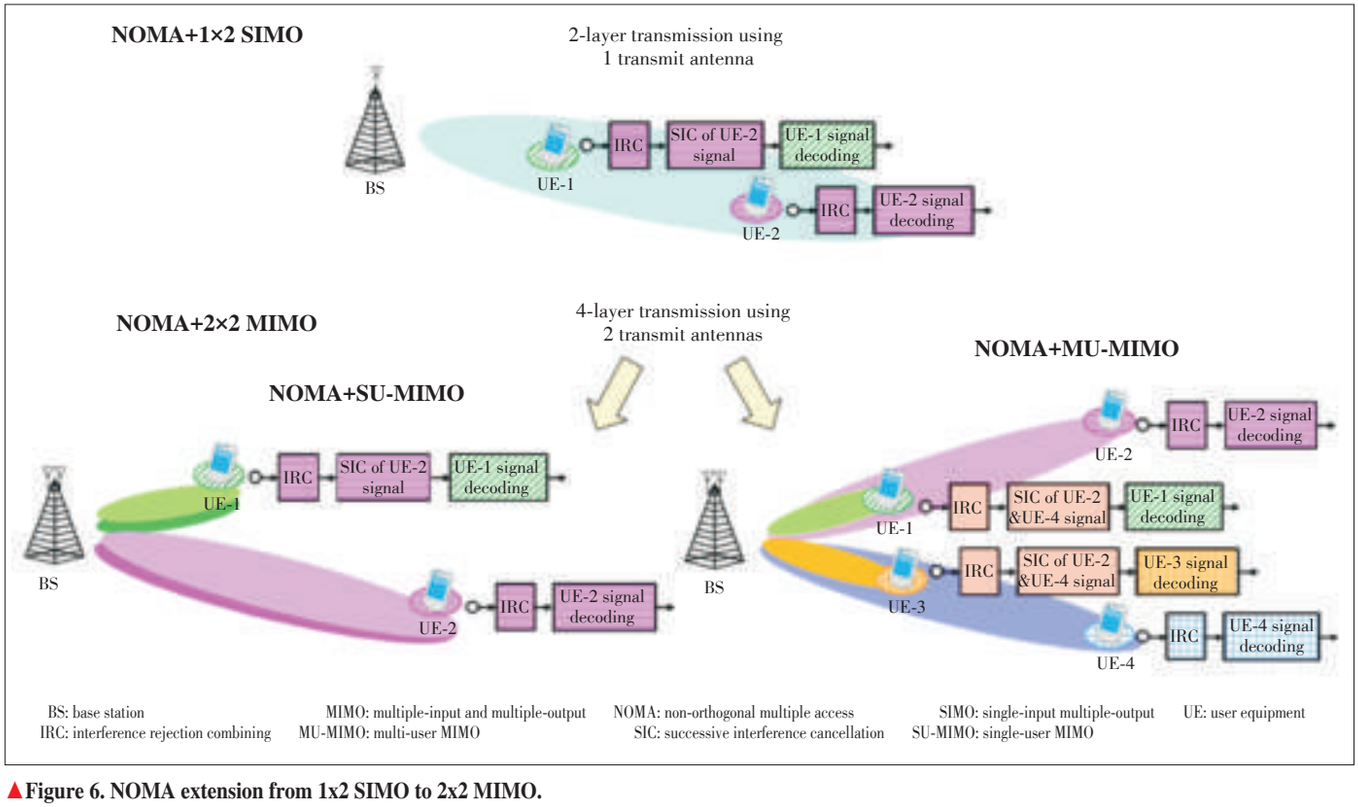
4.1 Downlink

There are two major approaches to combine downlink NOMA and MIMO technologies (**Fig. 6**).

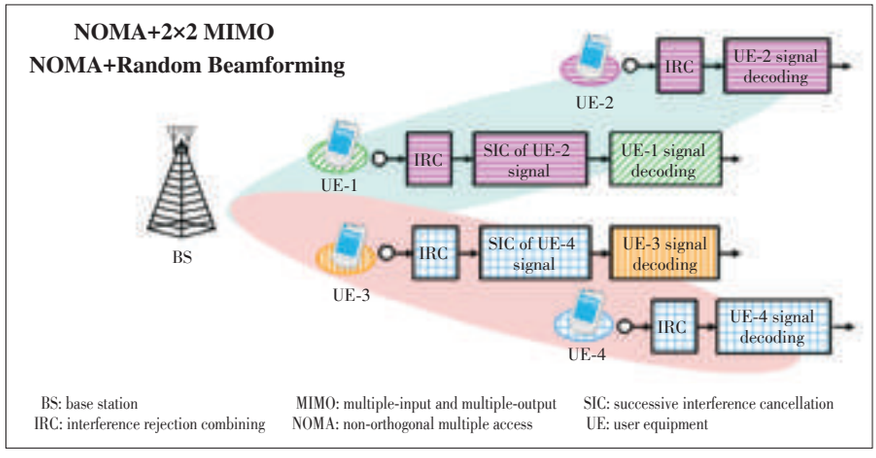
One approach is to use NOMA technique to create multiple power levels and apply SU-MIMO and/or MU-MIMO technique inside each power level. For example for NOMA with SU-MIMO (2x2), with up to 2 user multiplexing in the power-domain, non-orthogonal beam multiplexing enables up to 4 beam multiplexing using only 2 transmit antennas. In addition, the combination of NOMA with SU-MIMO can involve both open-loop MIMO (e.g., space frequency block coding (SFBC), large delay cyclic delay diversity (CDD)) and closed-loop MIMO (based on CSI such as the precoder indicator, channel quality indicator (CQI), rank indicator feedback by users)). Open-loop MIMO schemes when combined with NOMA are expected to provide robust performance in high mobility scenarios.

The other approach is to convert the non-degraded 2x2 MIMO channel into two degraded 1x2 SIMO channels, where NOMA is applied over each equivalent 1x2 SIMO channel separately, as shown in **Fig. 7** [10]. For this scheme, multiple transmit beams are created and superposition coding of signals designated to multiple users is applied within each transmit beam (i.e., intra-beam superposition coding). At the user terminal, the inter-beam interference is first suppressed by spatial filtering only by using multiple receive antennas, then multi-signal separation (e.g., SIC) is applied within each beam. This scheme can be considered as a combination of NOMA with MU-MIMO where fixed rank 1 transmission is applied to each user; thus, a large number of users would be required to obtain sufficient gains [11].

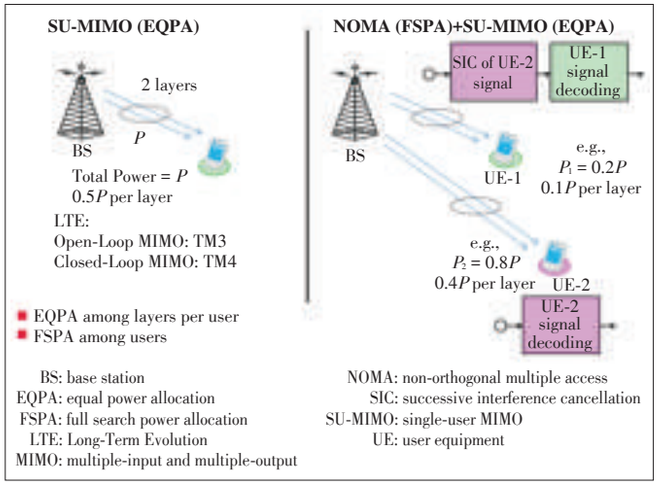
The combination of NOMA with SU-MIMO is illustrated in **Fig. 8** [20]. At the left side is the case of 2x2 SU-MIMO and at



▲**Figure 6. NOMA extension from 1x2 SIMO to 2x2 MIMO.**



▲**Figure 7. Downlink NOMA combined with 2x2 MIMO using random beamforming and applying IRC-SIC receivers.**



▲**Figure 8. Downlink NOMA with SIC combined with SU-MIMO (2x2 MIMO, 2-UE).**

the right side is the combination of NOMA with 2x2 SU-MIMO (N_t = N_r = 2), where the number of multiplexed UEs is 2. UE-1 and UE-2 are NOMA paired cell-center and cell-edge users, respectively.

By combining NOMA with SU-MIMO, up to 4-layer (4-beam) transmission is enabled using only 2 transmit antennas.

4.2 Uplink

Examples about uplink NOMA combined with MIMO assuming 2x2 antenna configuration are shown in **Fig. 9**. For the case of NOMA combined with SU-MIMO (left side), the UEs are separated in the power domain, and the spatial domain is used to multiplex multiple data streams of a single UE. For the case of NOMA combined with MU-MIMO (right side), UEs are separated in both power and spatial domains, i.e., within each user group of {UE-1, UE-2} and {UE-3, UE-4}, users are separated in the power domain. Among the {UE-1, UE-2} and {UE-3, UE-4} user groups, MU-MIMO transmission is applied to further separate the two user groups in the spatial domain. It can be seen that for the same MIMO antenna configuration, the same number of data streams are supported in uplink and downlink.

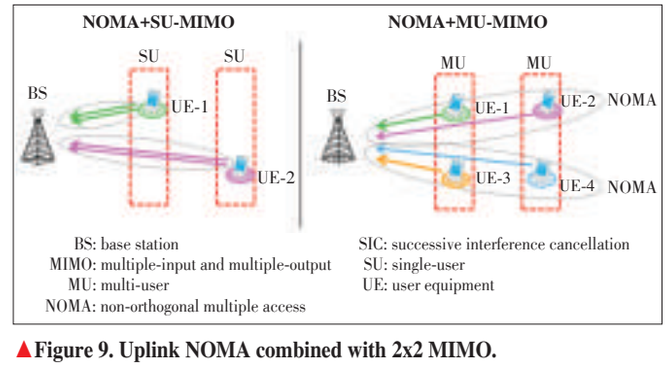
5 Performance of NOMA

5.1 NOMA Link-Level Performance

For the case of downlink NOMA with two UEs, we verified the effectiveness of CWIC for NOMA and the block error rate (BLER) performance of cell-center UE by evaluating link-level simulations. The number of transmit and receive antennas are both set to 2 and open-loop transmission mode 3 (TM3) is assumed as SU-MIMO transmission [16]. It is shown that almost the same BLER performance is obtained for NOMA with ideal SIC and CWIC. In addition, the power ratio (UE-1: UE-2 = P_1:P_2, P_1 + P_2 = 1.0, P_1 < P_2) for which the error propagation becomes dominant was investigated. The SIC error propagation increases for power sets with larger power ratio. In addition, the performance degrades with smaller values for the power ratio due to the increase of channel estimation error for the cell-center UE. Therefore, it would be important to limit the power sets to be used by the scheduler in order to maximize NOMA gains by limiting error propagation and the impact of channel estimation error, and ensuring that all chosen MCS combinations are decodable.

5.2 NOMA System-Level Performance

NOMA system-level performance has been investigated heavily with and without MIMO for both downlink and uplink. The multi-cell system-level simulation parameters are basically compliant with existing LTE specifications for an urban macro (UMa) scenario. The cell radius of the macro cells is set to 289 m (inter-site distance (ISD) = 500 m). 10 UEs are dropped randomly following a uniform distribution and full buffer traffic is assumed. Assuming proportional fairness scheduling, the performance gains of NOMA are measured in terms of cell



▲**Figure 9. Uplink NOMA combined with 2x2 MIMO.**

throughput (Mbps) and cell-edge user throughput (Mbps). The cell throughput is defined as the average aggregated throughput for users scheduled per a single cell, while the cell-edge user throughput is defined as the 5% value of the cumulative distribution function (CDF) of the user throughput.

The proportional fairness scheduler maximizes the geometric mean of the user throughputs, thus the tradeoff between user fairness and system throughput as shown in (10):

$$\sqrt[n]{\prod_{i\in U} R_i} = \frac{\sqrt[n]{\prod_{i\in U} R_i}}{\frac{1}{n}\sum_{i\in U} R_i} \times \frac{1}{n}\sum_{i\in U} R_i , \quad (10)$$

where U is the set of users scheduled. The first term on the right hand is the geometric mean of the user throughputs normalized with their arithmetic mean, representing a metric for user fairness, while the second term is the arithmetic mean of user throughputs, representing a metric for total system throughput.

5.2.1 Downlink

In [14], the user throughput of downlink NOMA is compared to that of OFDMA for both 1x2 SIMO and 2x2 MIMO. For MIMO, a comparison is made between the NOMA with SU-MIMO case and the OFDMA with SU-MIMO case for open-loop TM3 and closed-loop transmission mode 4 (TM4) MIMO. It is shown that NOMA with SU-MIMO provides gains over OMA with SU-MIMO by covering the entire user throughput region for both TM3 and TM4. The performance gains increase with the number of power sets. However, a hefty portion of the gains could be still achieved even with a few power sets.

5.2.2 Uplink

In [15], single-carrier frequency division multiple access (SC-FDMA) and NOMA are compared for uplink while taking uplink power control and resource contiguity constraint into account. A large performance gain in cell throughput is achieved for NOMA with very practical assumptions. This gain can be further increased by applying larger number of multiplexed users and/or enhanced schemes, e.g., advanced transmit power control (TPC). The large gain of NOMA mainly comes from the non-orthogonal multiplexing of users with large channel gain difference, which improves the resource utilization efficiency compared to SC-FDMA where only one UE exclusively occupies the radio resources.

When the user throughputs of SC-FDMA and NOMA are compared, it is observed that NOMA can achieve higher UE throughput than SC-FDMA for the most region of the CDF curve. However, for the cell-edge user throughput, i.e. 5% UE throughput, NOMA performance is worse than that of SC-FDMA. This is mainly due to two reasons. One reason is the increase of inter-cell interference in NOMA compared with SC-FDMA because more than one UE can be scheduled for simultaneous uplink transmission. The other reason is that the used TPC algorithm [8] is not fully optimized, where the total transmission power is controlled by a predefined parameter and the UEs in non-orthogonal transmission get less transmission power than what they get in SC-FDMA. Furthermore, the transmission power of the UEs is determined from large scale fading without considering instantaneous channel conditions.

To further balance between cell throughput and cell-edge throughput, three approaches are possible as listed below:

1) Introduction of weighted PF scheduling such that more resources are allocated to cell-edge user [12].

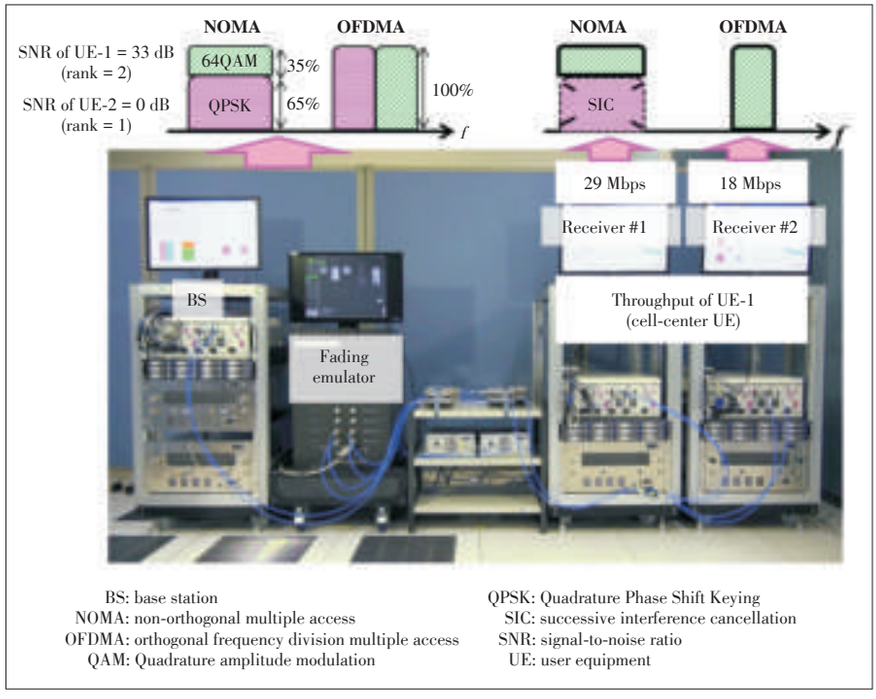
2) Combination of NOMA with other cell-edge performance enhancing technologies such as fractional frequency reuse (FFR) [15].

3) Introduction of sophisticated TPC algorithms designed for NOMA.

Taking the second approach above as an example, reference [15] shows that NOMA with FFR improves both the cell-edge throughput gain and the overall cell throughput gain. This possible improvement is due to the reduction in the inter-cell interference for both the cell-center UEs and cell-edge UEs.

5.2.3 NOMA Experimental Trial

A test-bed was developed to conduct experimental trials on NOMA, and to confirm NOMA performance with a real SIC receiver taking into account hardware (RF) impairments such as error vector magnitude (EVM) and the number of quantization bits of analog/digital (A/D) converter, etc. The test-bed assumed two UEs and used a carrier frequency of 3.9 GHz and bandwidth per user of 5.4 MHz for NOMA and of 2.7 MHz for OFDMA (a total bandwidth of 5.4 MHz for 2 users). LTE Release 8 frame structure is adopted and channel estimation is based on cell-specific reference signal (CRS). At the transmitter side, for each UE data, Turbo encoding, data modulation and multiplication by precoding vector are applied, then the precoded signal of the two UEs is superposed according to a predefined power ratio and goes through digital/analog (D/A) converter before up conversion to the carrier frequency of 3.9 GHz and transmission from two antennas. For MIMO transmission, LTE TM3 is utilized for open-loop 2-by-2 single user MIMO transmission. At the receiver side, two receive antennas are used to receive the RF signal, which is first down-converted and then goes through a 16-bit A/D converter. At the cell-center UE (UE-1), CWIC is applied. Using the fading emulator, for simplicity we set each link of the 2-by-2 MIMO channel to a 1-path channel with maximum Doppler frequency of 0.15 Hz. **Fig. 10** shows that the user throughputs of the cell-center UE, UE-1 (green color), with NOMA and SIC applied (29 Mbps) and with OFDMA only applied (18 Mbps). The user throughput of cell-edge UE, UE-2 (pink color) was adjusted for NOMA to be equal to the case of OFDMA. The measured gains of NOMA over OFDMA are the result of enabling three-layer transmission over a 2×2 MIMO channel while using twice the



▲Figure 10. NOMA test-bed.

bandwidth compared to OFDMA.

5.3 NOMA Standardization

NOMA was proposed to 3GPP LTE Release 13 [22] and a new study item (SI) under the name of "downlink multi-user superposition transmission (MUST)" was approved [23]. In 3GPP RAN1, the target scenarios, evaluation methodology, and the candidate non-orthogonal multiple access were discussed during the SI phase [24]–[26]. NOMA system-level performance with non-full buffer traffic and link-level performance for different receivers were evaluated [27], [28]. Based on Gray-mapped composite constellation with the same precoder but different transmit powers being applied to the superposed UEs, another NOMA multiplexing scheme is also considered in order to reduce signaling overhead and the receiver complexity compared to NOMA with SIC [26]. In such a scheme, coded bits for both the superposed UEs are jointly mapped onto the signal constellation based on Gray mapping, and then a reduced-maximum likelihood (R-ML) receiver is used for symbol-level interference cancellation [17]. The outcome of the SI in Release 13 was summarized under a technical report [29]. Later in Release 14, a work item (WI) was established to specify the necessary mechanisms to enable LTE to support downlink intra-cell multiuser superposition transmission for data channels with assistance information from serving BS to a UE regarding its experienced intra-cell interference [30]. In the WI, a MUST UE receiver is assumed to be capable to cancel or suppress intra-cell interference between co-scheduled MUST users for the following three cases:

- Case 1: Superposed data channels (i.e., Physical Downlink Shared Channels (PDSCHs)) are transmitted using the same transmission scheme and the same spatial precoding vector.
- Case 2: Superposed PDSCHs are transmitted using the same transmit diversity scheme.
- Case 3: Superposed PDSCHs are transmitted using the same transmission scheme, but their spatial precoding vectors are different.

During the WI phase, what was considered is up to 2 transmitter (Tx) CRS-based transmission schemes for cases 1 and 2, and up to 4 Tx CRS-based or up to 8 Tx DMRS-based transmission schemes for all three cases. The RAN1 agreements with RAN1 specification impacts made within the Release 14 WI are summarized in [31]. For example, for MUST Case 1 and Case 2, the higher layer and dynamic signaling mechanisms of MUST ON/OFF and of the power information of MUST users are specified.

6 Conclusions

This article presents an overview of the NOMA concept, design and its potential performance. Different from OFDMA, NOMA superposes multiple users in the power-domain, exploiting the channel gain difference between multiple UEs. NOMA contributes to the maximization of the tradeoff between system performance and user fairness. NOMA involves several aspects that need careful design, including the granularity in time and frequency of multi-user scheduling and multi-user power allocation, signaling overhead, receiver design, and combination with MIMO. NOMA can also be applied to uplink. For uplink, new issues arise including power control design to balance intra-cell and inter-cell interference and the design of the scheduling algorithm in case of single carrier transmission where consecutive resource allocation of non-orthogonally multiplexed UEs is taken into account.

From performance perspective, NOMA has shown promising gains for both downlink and uplink. These gains were investigated by link-level simulations, system-level simulations, and in experimental trials. Downlink NOMA was studied and specified in 3GPP RAN1 as MUST during LTE Release 13 and 14.

The design of sophisticated uplink power control schemes and of uplink reference signal for channel estimation to enable multiple user transmissions within the same frequency block is of interest to the future work.

Moreover, NOMA gains are expected to increase with more

users, which correspond to the case of massive machine type communications (mMTC), i.e., massive sensors and devices with small packets being simultaneously transmitted over the cellular network. Further investigations and optimizations of NOMA for mMTC are also of interest.

References

[1] Y. Kishiyama, A. Benjebbour, H. Ishii, and T. Nakamura, "Evolution concept and candidate technologies for future steps of LTE - A," in *Proc. IEEE ICCS2012*, Singapore, Nov. 2012, pp. 473–477. doi: 10.1109/ICCS.2012.6406193.
[2] 3GPP, "Physical layer aspects for Evolved UTRA," TR 25.814,V7.1.0, Oct. 2006.
[3] F. Rusek and J. B. Anderson, "Constrained capacities for faster-than-Nyquist signaling," *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 764–775, Feb. 2009. doi: 10.1109/TIT.2008.2009832.
[4] L. Ping, 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.
[5] L. Ping, Q. Guo, and J. Tong, "The OFDM-IDMA approach to wireless communication systems," *IEEE Wireless Communications*, vol. 14, no. 3, pp. 18–24, Jun. 2007. doi: 10.1109/MWC.2007.386608.
[6] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with successive interference cancellation for future radio access," in *Proc. APWCS 2012*, Kyoto, Japan, Aug. 2012.
[7] A. Benjebbour, Y. Saito, Y. Kishiyama, et al., "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proc. IEEE ISPACS*, Okinawa, Japan, Nov. 2013, pp. 770–774. doi: 10.1109/ISPACS.2013.6704653.
[8] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE PIMRC* 2013, London, UK, Sept. 2013, pp. 611–615. doi: 10.1109/PIMRC.2013.6666209
[9] A. Benjebbour, A. Li, Y. Saito, et al., "System-level performance of downlink NOMA for future LTE enhancements," in *Proc. IEEE Globecom*, Atlanta, USA, Dec. 2013, pp. 66–70. doi: 10.1109/GLOCOMW.2013.6824963.
[10] K. Higuchi and Y. Kishiyama, "Non-orthogonal access with random beamforming and intra-beam SIC for cellular MIMO downlink," in *Proc. IEEE VTC2013-Fall*, Las Vegas, USA, Sept. 2013, pp. 1–5. doi: 10.1109/VTC Fall.2013.6692307.
[11] A. Li, A. Benjebbour, and A. Harada, "Performance evaluation of non-orthogonal multiple access combined with opportunistic beamforming," in *Proc. IEEE VTC Spring 2014*, Seoul, Korea, May 2014, pp. 1– 5. doi: 10.1109/VTC-Spring.2014.7023050.
[12] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. E98-B, no. 3, pp. 403–414, Jan. 2015.
[13] Y. Lan, A. Benjebbour, X. Chen, A. Li, and H. Jiang, "Considerations on downlink non-orthogonal multiple access (NOMA) combined with closed-loop SU-MIMO," in *Proc. ICSPCS 2014*, Gold Coast, Australia, pp. 1–5. doi: 10.1109/ICSPCS.2014.7021086.
[14] A. Benjebbour, A. Li, Y. Kishiyama, H. Jiang, and T. Nakamura, "System-level performance of downlink NOMA combined with SU-MIMO for future LTE enhancements," in *Proc. IEEE Globecom*, Austin, USA, Dec. 2014, pp. 706–710. doi: 10.1109/GLOCOMW.2014.7063515.
[15] A. Li, A. Benjebbour, X. Chen, H. Jiang, and H. Kayama, "Uplink non-orthogonal multiple access (NOMA) with single-carrier frequency division multiple access (SC-FDMA) for 5G Systems," *IEICE Transactions on Communications*, vol. E98-B, no. 8, pp. 1426–1435, Aug. 2015.
[16] K. Saito, A. Benjebbour, Y. Kishiyama, Y. Okumura, and T. Nakamura, "Performance and design of SIC receiver for downlink NOMA with open-loop SU-MIMO," in *Proc. IEEE ICC*, London, UK, Jun. 2015, pp. 1161– 1165, doi: 10.1109/ICCW.2015.7247334.
[17] C. Yan, A. Harada, A. Benjebbour, et al., "Receiver design for downlink non-orthogonal multiple access (NOMA)," *Proc. IEEE VTC Spring 2015*, Glasgow, Scotland, May 2015, pp. 1–6. doi: 10.1109/VTCSpring.2015.7146043.
[18] A. Benjebbour, A. Li, K. Saito, Y. Kishiyama, and T. Nakamura, "Downlink non-orthogonal multiple access (NOMA) combined with single user MIMO (SU-MIMO), " *IEICE Transactions on Communications*, vol. E98-B, no. 8, pp. 1415-1425, Aug. 2015.
[19] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, England: Cambridge University Press, 2005.
[20] G. Caire and S. Shamai, "On the achievable throughput of a multi-antenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1692–1706, Jul. 2003. doi: 10.1109/TIT.2003.813523.
[21] 3GPP, "Study on network-assisted interference cancellation and suppression for LTE," RP-130404, Feb. 2013.
[22] 3GPP, "Justification for NOMA in new study on enhanced multi-user transmission and network assisted interference cancellation for LTE," RP-141936, Dec. 2014.
[23] 3GPP, "New SI proposal: Study on downlink multiuser superposition transmission for LTE," RP-150496, Mar. 2015.
[24] 3GPP, "Deployment scenarios for downlink multiuser superposition transmissions," R1-152062, Apr. 2015.
[25] 3GPP, "Evaluation methodologies for downlink multiuser superposition transmissions," R1-153332, May 2015.
[26] 3GPP, "Candidate non-orthogonal multiplexing access scheme," R1-153335, May 2015.
[27] 3GPP, "System-level evaluation results for downlink multiuser superposition schemes," R1-154536, Aug. 2015.
[28] 3GPP, "Link-level evaluation results for downlink multiuser superposition schemes," R1-154537, Aug. 2015.
[29] 3GPP, "Study on downlink multiuser superposition transmission (MUST) for LTE (Release 13)," TR 36.859 V13.0.0, Dec. 2015.
[30] 3GPP, "New work item proposal: Downlink multiuser superposition transmission for LTE," RP-160680, Mar. 2016.
[31] 3GPP, "Summary of RAN1 agreements for Rel-14 DL MUST," R1-1613802, Nov. 2016.

Manuscript received: 2016-10-11

Biography

Anass Benjebbour (benjebbour@nttdocomo.com) obtained his Ph.D. and M.Sc. degrees in telecommunications in 2004 and 2001, respectively, and his B.Sc. diploma degree in electrical engineering in 1999, all from Kyoto University, Japan. In 2004, he joined NTT DOCOMO, INC. Since 2010, he has been a leading member of its 5G team. His research interests include novel system design concepts and radio access techniques for next generation mobile communication systems (5G), such as massive MIMO, NOMA, and waveform design. Dr. Benjebbour served as a 3GPP and ITU-R standardization delegate, a secretary of the IEICE RCS conference from 2012 to 2014, an associate editor for the *IEICE Communications Magazine* from 2010 to 2014, and an associate editor for the *IEICE Transactions on Communications* from 2014 to 2018. He is an author or a coauthor of 100+ technical publications, 4 book chapters and is an inventor of 50+ patent applications. He is a senior member of IEEE and IEICE.

DOI: 10.3969/j. issn. 1673-5188. 2017. S1. 004
http://kns.cnki.net/kcms/detail/34.1294.TN.20170623.1022.004.html, published online June 23, 2017

Uplink Multiple Access Schemes for 5G: A Survey

YANG Shan, CHEN Peng, LIANG Lin, ZHU Jianchi, and SHE Xiaoming
(Technology Innovation Center, China Telecom, Beijing 102209, China)

Abstract

In non-orthogonal multiple access (NMA) system, signal transmitter and receiver are jointly optimized, so that multiple layers of data from more than one user can be simultaneously delivered in the same resource. To meet the 5G requirements on the number of connections and spectral efficiency, uplink NMA is becoming an important candidate technology and has been extensively studied in 3GPP. A number of uplink NMA schemes from different industrial companies have been proposed in recent 3GPP meetings. In terms of their basic technique principles, this paper classifies these NMA schemes into three categories, namely: scrambling based NMA schemes, interleaving based NMA schemes, and spreading based NMA schemes. Moreover, the key characteristics of these schemes are summarized, and the detailed introduction of each scheme is provided according to the comprehensive survey of the latest progress in 3GPP 5G standardization work.

Keywords

5G; non-orthogonal multiple access; scrambling; interleaving; spreading

1 Introduction

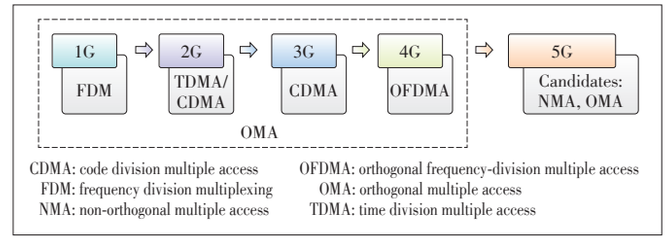
In wireless communications, multiple access technology allows several user devices to share one radio transmission resource. Over the past twenty years, the innovation on multiple access technology has been an essential part for each new generation of cellular mobile systems.

Long-Term Evolution (LTE) and LTE-Advanced networks are now being more and more widely deployed by global mobile operators. Meanwhile, the 5G research towards the year 2020 and beyond has been started in the academia and industry worldwide. ITU has defined three usage scenarios including enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and low latency communications (URLLC), as well as the key technology capabilities for IMT-2020 (5G) [1]. Moreover, 3GPP launched the study on 5G core network (CN) and radio access network (RAN) in October 2015 and March 2016, receptively [2].

Compared with 4G system, two of the key 5G capabilities are to provide higher connection density and spectral efficiency [1], [3]. As seen in **Fig. 1**, the 1G to 4G cellular systems are mainly based on orthogonal multiple access (OMA) technologies. In recent years, non-orthogonal multiple access has attracted more and more interests and has become an important candidate technology for 5G system [4], [5].

Non-orthogonal multiple access (NMA) allows the simultaneous transmission of more than one layer of data for more than one piece of user equipment (UE) without time, frequency or spatial domain separation. Different layers of data can be separated by utilizing interference cancellation or iterative detection at the receiver. On one hand, the point-to-point link performance of LTE is quite close to the single UE channel capacity, thus the improvement in link performance would be limited. On the other hand, NMA can be used to further enhance the spectral efficiency over OMA, in order to achieve the multiple UE channel capacity, as shown in the previous research in [6] and [7]. Furthermore, NMA can significantly increase the number of UE connections, which is quite beneficial for MTC services. In addition, NMA does not rely on the knowledge of instantaneous channel state information (CSI) of frequency-selective fading, and thus a robust performance gain in practical wide area deployments can be expected irrespective of UE mobility or CSI feedback latency.

Uplink NMA schemes have been studied in 3GPP RAN WG1 (working group 1) since March 2016. It has been agreed that NMA should be investigated for diversified 5G usage scenarios and use cases [8], and 5G should target to support up-



▲**Figure 1. Multiple access technology for cellular systems.**

link NMA, at least for mMTC scenario [9].

A number of NMA schemes have been proposed to 3GPP by different industrial companies, including operators, base station (BS) and UE vendors as well as chipset vendors. This paper gives a comprehensive survey on these candidate NMA schemes, and provides an insight on the latest progress in 3GPP 5G standardization work.

The rest of this paper is organized as follows. Section 2 outlines all the candidate NMA schemes submitted to 3GPP. The three categories of NMA schemes, namely scrambling based NMA, interleaving based NMA, and spreading based NMA schemes are presented in Section 3, Section 4 and Section 5, respectively. Conclusions are drawn in Section 6.

2 Uplink NMA Schemes

In an uplink NMA system, signal transmitter and receiver are jointly optimized, so that multiple layers of data from more than one UE can be simultaneously delivered in the same resource (**Fig. 2**). At the transmitter side, the information of different UEs can be delivered using the same time, frequency and spatial resource. At the receiver side, the information of different UEs can be recovered by advanced receivers such as interference cancellation or iterative detection receivers.

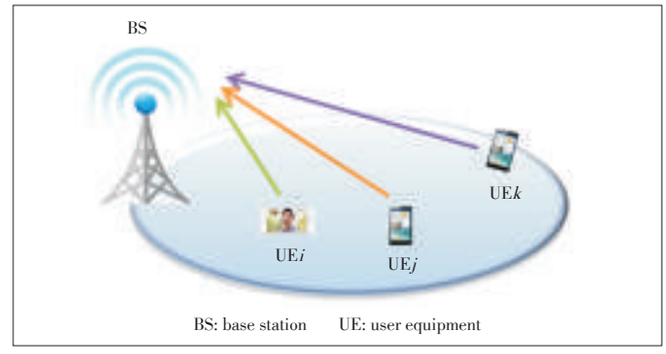
As mentioned above, a number of uplink NMA schemes have been proposed in the recent RAN1 meetings [10]–[21]. The difference of these schemes is mainly on UE's signature design, i.e., whether the scrambling sequence, interleaver or spreading code is used to differentiate UEs. Therefore, these schemes can be classified into the following three categories:

- Category 1: scrambling based NMA schemes
- Category 2: interleaving based NMA schemes
- Category 3: spreading based NMA schemes.

The key characteristics of the three categories are summarized in **Table 1**. The general descriptions on the 12 candidate schemes are given in **Table 2**, while the detailed introduction is provided in Section 3 to Section 5.

3 Scrambling Based NMA Schemes

As discussed in Section 2, a key characteristic of scrambling based NMA schemes is that different scrambling sequences are used to distinguish different UEs, and that an successive interference cancellation (SIC) algorithm is applied at the BS receiver to separate different UEs' information. In the following, we will introduce three typical scrambling based NMA schemes proposed in 3GPP 5G study phase.



▲**Figure 2. Illustration of uplink NMA.**

3.1 Non-Orthogonal Multiple Access (NOMA)

In NOMA, the improvements in UE spectral efficiency and the number of connection can be expected by sharing the same radio resources among multiple UEs and allocating more radio resource per UE (**Fig. 3** [10]). These multiplexed UEs can be separated by assigning different scrambling sequences to different UEs and creating received power difference among paired UEs. An advanced baseband demodulation receiver, e.g., SIC receiver, is employed at the receiver side as shown in **Fig. 4**. Note that the received power difference can be created through either large-scale channel difference or small-scale channel variance, i.e., NOMA is also applicable for multiple UEs with similar wideband signal to interference plus noise ratio (SINR) thanks to the variation in small-scale channel.

3.2 Resource Spread Multiple Access (RSMA)

In RSMA, a group of different UEs' signals are super-positioned on top of each other, and each UE's signal is spread to the entire frequency/time resource assigned for the group [11]. Different UEs' signals within the group are not necessarily orthogonal to each and could potentially cause inter-UE interfer-

▼**Table 1. Categories of NMA schemes**

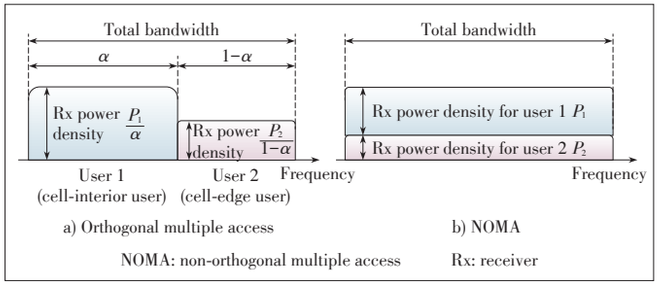
	Category 1: Scrambling based	Category 2: Interleaving based	Category 3: Spreading based	
Key characteristics	• Use different scrambling sequences to distinguish different UEs • Can be used together with low code rate FEC	• Use different interleavers to distinguish different UEs • Can be used together with low code rate FEC	Use different spreading codes to distinguish different UEs	
			LDS code based	Non-LDS code based
Candidate schemes	• NOMA[10] • RSMA [11] • LSSA [12]	• IDMA [13] • IGMA [14]	• SCMA [15] • PDMA [16] • LDS-SVE [17]	• MUSA [18] • NOCA [19] • NCMA [20] • LCRS [21]

FEC: forward error correction
IDMA: interleave division multiple access
IGMA: interleave-grid multiple access
LCRS: low code rate spreading
LDS: low density signature
LDS-SVE: low density signature-signature vector extension
LSSA: low code rate and signature based shared access
MUSA: multiple user shared access
NCMA: non-orthogonal coded multiple access
NOCA: non-orthogonal coded access
NOMA: non-orthogonal multiple access
PDMA: pattern defined multiple access
RSMA: resource spread multiple access
SCMA: sparse code multiple access
UE: user equipment

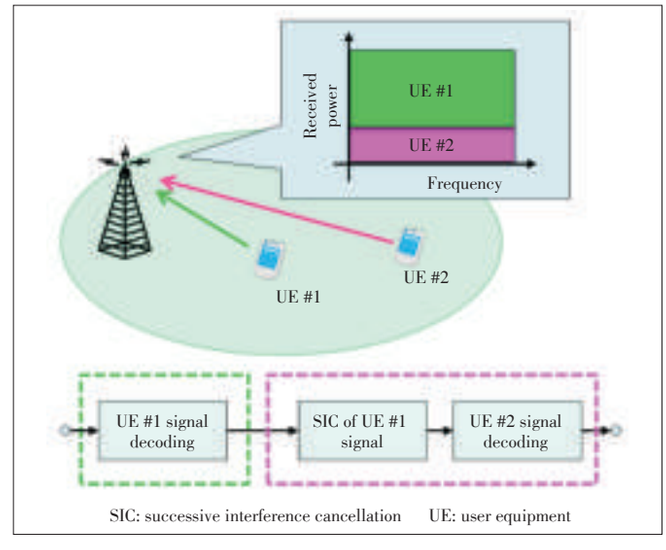
▼Table 2. Candidate uplink NMA schemes

NMA scheme		Description	Standardization impact	Receiver algorithm
Category 1: Scrambling based	NOMA	• Multiple UEs with different scrambling sequences are transmitted on the same resource • NOMA can also bring in performance gain for multiple UEs with similar wideband SINR, thanks to the fast fading	• Define new scrambling sequence if needed • Power control enhancement if needed	SIC
	RSMA	Use combination of low rate channel codes and scrambling codes (and optionally different interleavers) with good correlation properties	• Define scrambling sequence and interleaver if needed • Define single carrier based new waveform for asynchronous transmission	SIC
	LSSA	Each UE's data is bit or symbol level multiplexed with UE specific signature pattern which is unknown to others	Define signature pattern	SIC
Category 2: Interleaving based	IDMA	Use bit level interleavers to separate UEs	Define bit-level interleaver	ESE
	IGMA	Use bit level interleavers and/or grid mapping pattern to separate UEs	• Define bit-level interleaver • Define sparse symbol-to-RE grid mapping pattern	ESE or chip-by-chip MAP
Category 3: Spreading based	SCMA	The coded bits of a data stream are directly mapped to a codeword from a codebook built based on a multi-dimensional constellation, and low density spreading is utilized	Define LDS code and multi-dimensional constellation	MPA, or MPA with SIC
	PDMA	A code is used to define sparse mapping from data to a group of resources, and different codes may have different diversity orders	Define LDS code matrix	BP based iterative detection and decoding
	LDS-SVE	For LDS spreading, consider UE signature vector extension, e.g., transforming and concatenating two element signature vectors into a larger signature vector	• Define LDS code • Define signature vector extension method	MPA
	MUSA	Use random complex spreading codes with short length, and the real part and imaginary part of each element in the complex spreading code are drawn from a multi-level real value set uniformly, for example, {-1, 1} or {-1, 0, 1}	Define spreading code	SIC
	NOCA	Use LTE defined low correlation sequences as spreading codes, e.g., LTE defined sequences for uplink reference signal for 1 RB case	Reuse the LTE defined sequence as spreading code	SIC
	NCMA	Spreading codes are obtained by Grassmannian line packing problem	Define spreading code	PIC
	LCRS	Apply direct spreading of modulation symbols and to transmit the spread symbols in time-frequency resources allocated for non-orthogonal transmission	Define spreading code	SIC

BP: belief propagation
ESE: elementary signal estimator
IDMA: interleave division multiple access
IGMA: interleave-grid multiple access
LCRS: low code rate spreading
LDS: low density signature
LDS-SVE: low density signature-signature vector extension
LSSA: low code rate and signature based shared access
LTE: Long-Term Evolution
MAP: maximum a posteriori
MPA: message passing algorithm
MUSA: multiple user shared access
NCMA: non-orthogonal coded multiple access
NOCA: non-orthogonal coded access
NOMA: non-orthogonal multiple access
PDMA: pattern defined multiple access
PIC: parallel interference cancellation
RB: resource block
RE: resource element
RSMA: resource spread multiple access
SCMA: sparse code multiple access
SIC: successive interference cancellation
SINR: signal to interference plus noise ratio
UE: user equipment



▲Figure 3. Comparison of orthogonal multiple access and NOMA in uplink.



▲Figure 4. Uplink NOMA.

ence. Spreading of bits to the entire resources enables decoding at a signal level below background noise and interference.

RSMA uses the combination of low rate channel codes and scrambling codes (and optionally different interleavers) with good correlation properties to separate different UEs' signals. Depending on application scenarios, the RSMA to be discussed includes:

• Single carrier RSMA: optimized for battery power consumption and coverage extension for small data transactions by using single carrier waveforms and very low peak to average power ratio (PAPR) modulations. Moreover, an asynchronous access is potentially allowed in this case.

• Multi-carrier RSMA: optimized for low latency access for radio resource control (RRC) connected state UEs (i.e., timing with BS already acquired).

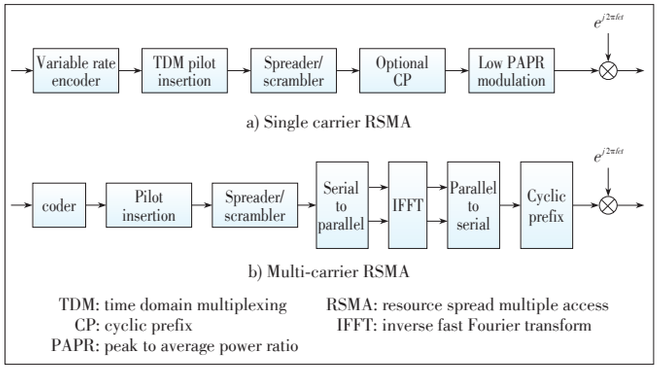
In **Fig. 5** that shows these two modes, "TDM pilot insertion" indicates inserting pilot which is time domain multiplexing (TDM) with the data, "Option CP" indicates that cyclic prefix (CP) is optional for single carrier RSMA, "IFFT" indicates the inverse fast Fourier transform(IFFT) operation, which is similar to that used in LTE.

RSMA will prioritize the usage of low rate channel codes to fully leverage the coding gain, except in the very low spectral efficiency region where capacity scales linear with power and repetition accounts for most of the gain. Also, the SIC receiver is used at the receiver side. For scrambling, the uplink scrambling sequence designed for uplink WCDMA system can be reused.

3.3 Low Code Rate and Signature Based Shared Access (LSSA)

In LSSA (**Fig. 6**), each UE's information is encoded with very low rate channel coding [12]. The channel encoding part can be optionally replaced with slightly higher channel coding rate with non-orthogonal spreading sequence. Then the output of channel encoder is bit or symbol level multiplexed with UE specific signature.

A UE's signature is a set that consists of complex or binary sequence and permutation pattern of a short length vector. The short length signature vector, however, does not mean that the number of multiplexed uplink UEs is limited to the length of the signature. UE overloading feature can well be supported since the receiver separates the targeted UE signal from other UE's contribution to multi-UE interference without relying on orthogonal multiplexing codes.



▲**Figure 5. RSMA block diagrams.**

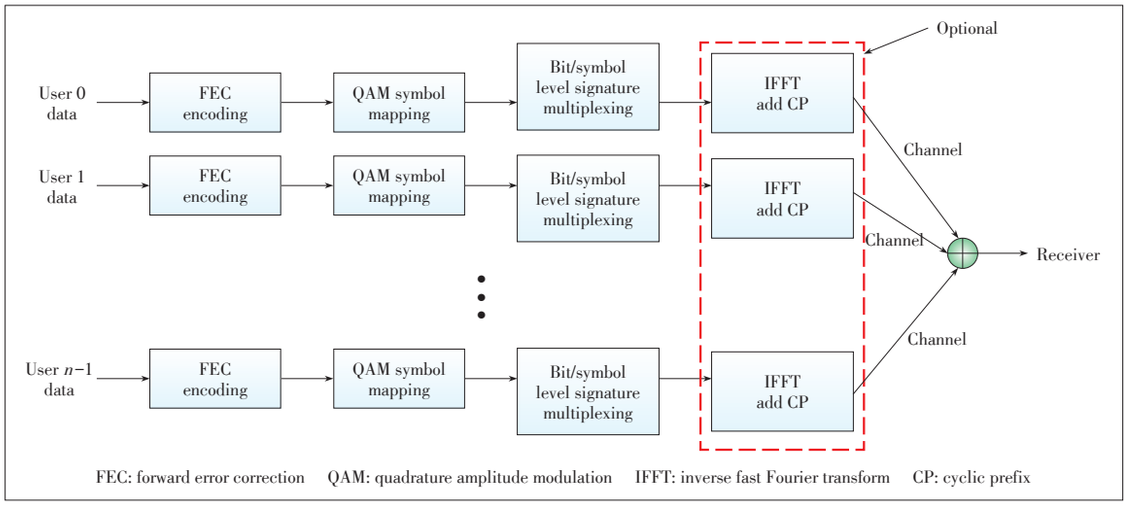
4 Interleaving Based NMA Schemes

The key characteristic of interleaving based NMA schemes is that different interleavers are used to distinguish different UEs, and a low code rate forward error correction (FEC) can be applied together. At the receiver side, an elementary signal estimator (ESE) with or without iteration is used [22], and maximum a posteriori (MAP)/message passing algorithm (MPA) can also be used if there exists zero entries. In the following, we will introduce two typical interleaving based NMA schemes proposed in 3GPP 5G study.

4.1 Interleave Division Multiple Access (IDMA)

IDMA was proposed in [22], which originally targeted to the performance enhancement for asynchronous code division multiple access (CDMA) system. Further studies revealed that IDMA exhibits strong robustness against asynchronicity and tolerance upon UEs' overloading. Meanwhile, the IDMA receiver, denoted as ESE and recognized in literature, turns out to be simple and effective.

In **Fig. 7**, a scenario for performance comparison between frequency division multiple access (FDMA) and IDMA is presented [13]. In the FDMA scheme, each UE is allocated to a relatively narrow frequency bandwidth F/K, considering an FEC rate R_c and temporal frame-length T, and mapping M_F bits



◀**Figure 6. Example of LSSA transmitter structure.**

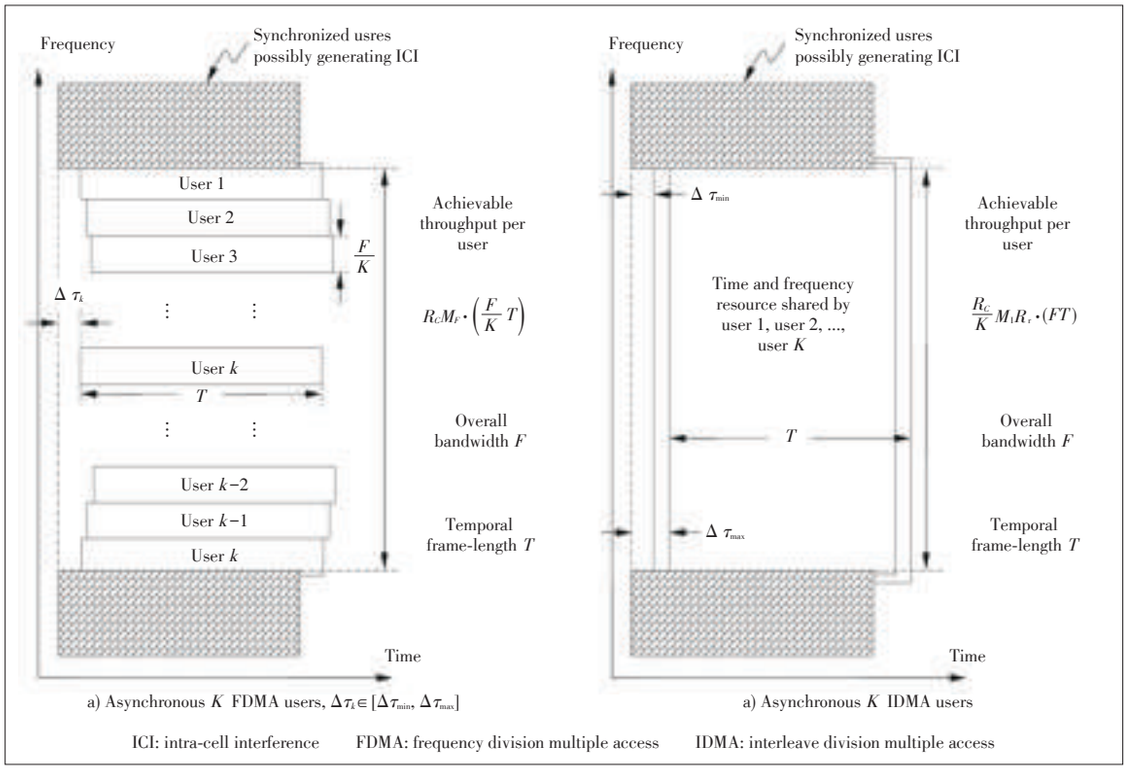


Figure 7. Scenario for FDMA and IDMA comparison.

to a modulated symbol. The achievable throughput per UE is $R_c\ M_r\ (F/K\times T)$. In the IDMA scheme, all of the UEs are allowed to share the complete bandwidth F, by reducing the FEC rate to R_c/K, optionally deploying M_l bits to a modulated symbol and a repetition code with rate R_r. The achievable throughput per UE is thus $R_c/K\ M_l R_r\ (FT)$.

4.2 Interleave-Grid Multiple Access (IGMA)

Basically, this IGMA scheme could distinguish different UEs on the basis of:

- Different bit-level interleavers
- Different grid mapping patterns
- Different combinations of bit-level interleaver and grid mapping pattern.

The typical transmitter system structure using IGMA is shown in **Fig. 8** [14].

The channel coding process can be either using simple repetition (or spreading) of a moderate coding rate FEC or directly using low coding rate FEC. The sufficient source of bit-level interleavers and/or grid mapping patterns is able to provide enough scalability to support 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 could be achieved.

In the grid mapping process, sparse mapping based on zero padding and symbol-level interleaving is introduced, which could provide another dimension for UE multiplexing (**Fig. 9**). Moreover, the density ρ of the grid mapping pattern is defined as the occupied resource elements(REs) N_{used} dividing the total assigned REs N_{all}, i.e., $\rho = N_{used}/N_{all}$. Different densities could be flexibly configured. It is noted that the symbol sequence order will be randomized after the grid mapping process due to symbol-level interleaving. This may further bring benefits in terms of combating frequency selective fading and inter-cell interference in comparison with resource mapping using direct code-matrices/codebooks.

At the receiver side, the low complexity multi-UE detector, i.e., ESE, takes advantage of the special property of interleaving [22], and can be utilized with a simple de-mapping operation

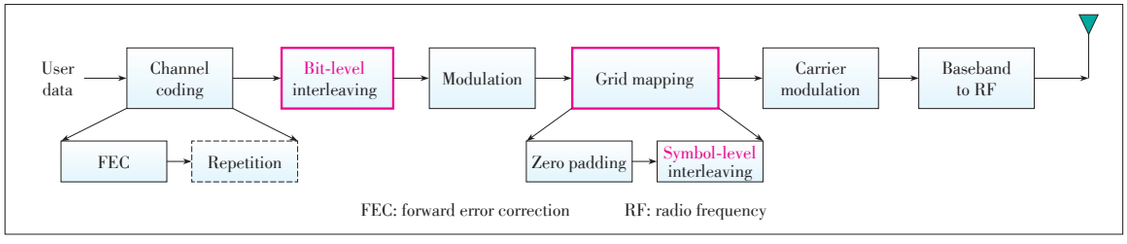
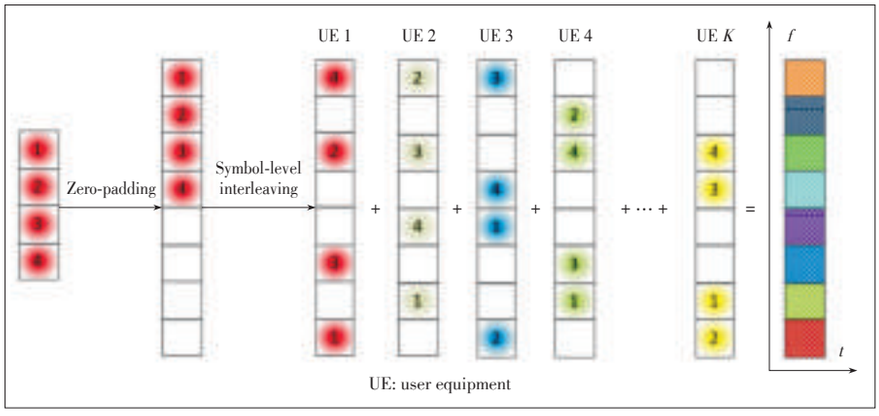


Figure 8. Schematic of IGMA transmitter.



▲Figure 9. Example of grid mapping process.

on the top. Note that lower density of the grid mapping pattern could further reduce detection complexity of ESE for IGMA. In addition, MAP and MPA detectors are applicable for IGMA and they can significantly improve the detection performance in comparison to ESE at the cost of complexity. The complexity of MAP/MPA for IGMA can probably be alleviated when sparse grid mapping is used, due to the similar property of low density signature (LDS).

5 Spreading Based NMA Schemes

The key characteristic of spreading based NMA schemes is that different spreading codes are used to distinguish different UEs. Depending on whether there exist zero entries in the spreading code, spreading based NMA schemes can be further classified into two sub-categories, i.e., the LDS based and non-LDS based. For LDS spreading based NMA schemes, an MPA or belief propagation (BP) receiver is usually employed; while for non-LDS spreading based NMA schemes, SIC or parallel interference cancellation (PIC) is applied at the receiver. Organized into these two groups, seven typical spreading based NMA schemes proposed in 3GPP 5G study are presented in this section. More specifically, LDS spreading based NMA schemes include sparse code multiple access (SCMA), pattern defined multiple access (PDMA), low density signature - signature vector extension (LDS-SVE).

Non-LDS spreading based NMA schemes consists of multiple user shared access (MUSA), non-orthogonal coded access (NOCA), non-orthogonal coded multiple access (NCMA), low code rate spreading (LCRS).

5.1 SCMA

Fig. 10 illustrates a basic system model with SCMA [23], [24] where the coded bits of a data stream are directly mapped to a codeword from a built codebook according to a multi-dimensional constellation. As can be seen in Fig. 10, the basic structure of SCMA implementation would be similar to LTE transmission model, with a key difference of joint design of modulation and spreading [15]. SCMA utilizes low density spreading, also named as sparse spreading, which has been used in the LDS technique in CDMA system. An introduction of multi-dimensional modulation for 5G New Radio (NR) can be found in [25].

The SCMA mapping and multiple access procedures are explained as follows.

1) Codebook mapping

Similar to LTE, SCMA supports layer mapping, i.e. one or multiple SCMA layers can be assigned to a data stream. Different from LTE, the SCMA also conducts mapping from information bits to codewords at each SCMA layer, i.e. the SCMA modulator maps input bits to a complex multi-dimensional codeword selected from a layer-specific SCMA codebook. SCMA codewords are sparse, i.e. only few of their entries are non-zero and the rest are zero. All SCMA codewords corresponding to a SCMA layer have a unique location of non-zero entries, referred to as sparsity pattern for simplicity.

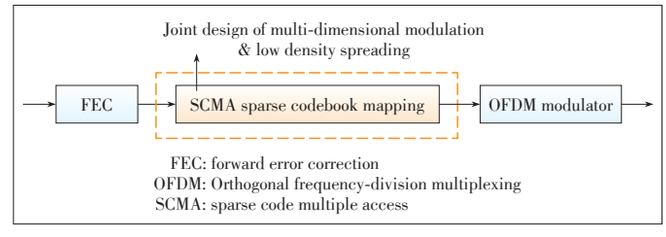
Fig. 11 shows an example of a codebook set containing 6 codebooks for transmitting 6 data layers. As can be seen, each of the codebook has 8 multi-dimensional complex codewords that correspond to 8 points of constellation, respectively. The length of each codeword is 4, which is the same as the spreading length. Upon transmission, the codeword of each layer is selected on the basis of the input bit sequence.

2) Multiple access procedure

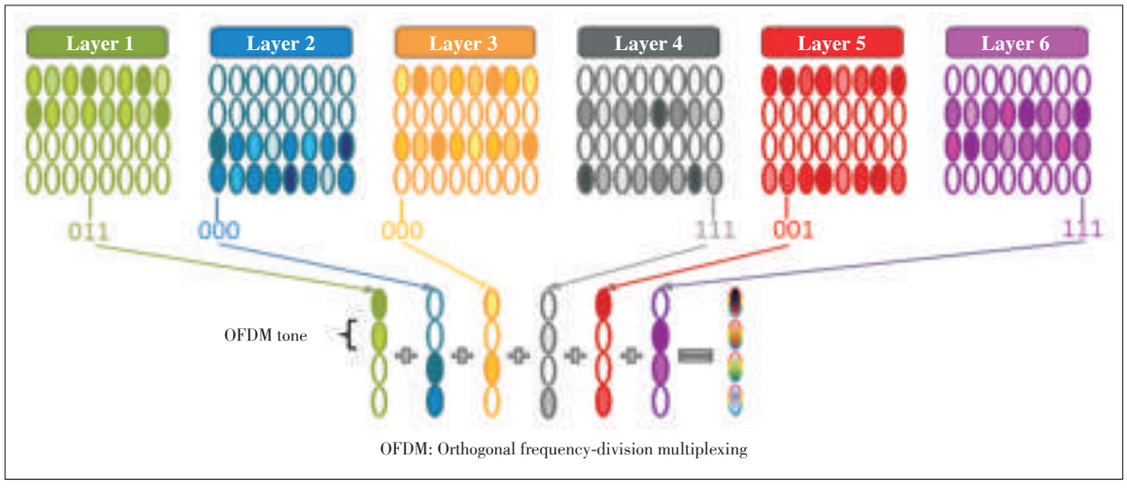
Fig. 12 shows an example of multiple access of 6 UEs with the SCMA layer-specific codebooks illustrated in Fig. 11. Each UE is assigned with one SCMA codebook. In the example, UE i takes codebook for layer i, i = 1, 2, ⋯, 6. After the FEC encoder, each UE's coded bits are then mapped to the SCMA codeword according to its assigned codebook. The SCMA codewords are further combined over OFDM tones and symbols are transmitted in terms of SCMA blocks, similar to resource block concept in LTE.

The main characteristics of multiple access with SCMA can be summarized as follow:

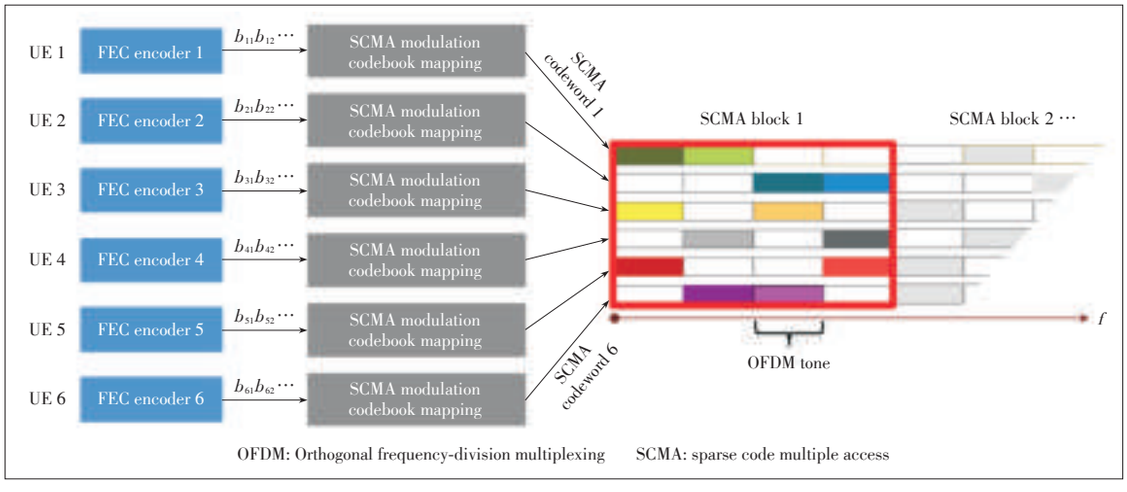
- Code domain non-orthogonal signal superposition: It allows superposition of multiple symbols from different UEs on each RE. For example, in Fig. 12, on RE 1, symbols from



▲Figure 10. SCMA codebook mapping.



◀Figure 11. SCMA codebook bit-to-codeword mapping.



◀Figure 12. Multiple access with SCMA.

UE 1, 3, and 5 are overlapped with each other. The superposition pattern on each RE can statically be configured or semi-statically.

- Sparse spreading: SCMA uses sparse spreading to reduce the number of symbol collisions. For example, in Fig. 12, there are 3 symbols from different UEs are colliding over each RE, instead of 6 in the case of non-sparse spreading.
- Multi-dimensional modulation: SCMA uses multi-dimensional modulation instead of linear spreading as in CDMA [25].

5.2 PDMA

A code 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 number of resource 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. BP based iterative detection and decoding (BP-IDD) is applied at the receiver.

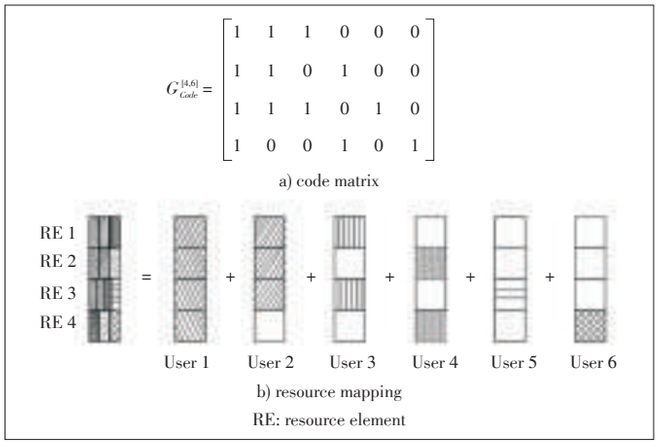
Assuming six users multiplexing on four REs, **Fig. 13** shows an example of code matrix and related resource mapping. User 1's data is mapped to all four resources in the group, and user 2's data is mapped to the first three resources, etc. The order of transmission diversity of the six users is 4, 3, 2, 2, 1, and 1 respectively.

It can be seen that a code with heavier weight (i.e., number of '1' elements in the pattern) provides higher diversity order, more reliable data transmission can be anticipated, and detection complexity is also increased. Moreover, codes shall have as many different diversity orders as possible to fasten convergence of BP receiver.

5.3 LDS-SVE

LDS-SVE is also one type of LDS spreading based NMA schemes. The main difference of LDS-SVE from the basic LDS spreading based NMA schemes is to consider UE signature vector extension, e.g., transforming and concatenating two element signature vectors into a larger signature vector for LDS spreading. Moreover, the MPA receiver is applied at the BS receiver side.

Uplink Multiple Access Schemes for 5G: A Survey
YANG Shan, CHEN Peng, LIANG Lin, ZHU Jianchi, and SHE Xiaoming



▲Figure 13. Six users sharing on four REs.

To combat inter-UE interference or collision, there is actually implicit information dependency, i.e., some kind of information replication within one UE signature vector. Typically, a UE transmits a block of signature vectors; If dependency can be further introduced among these vectors, more robustness against inter-UE interference or higher order of diversity can be obtained. UE signature vector extension is one way to achieve this. The extension can be accomplished by transforming and concatenating e.g. two element signature vectors into a larger signature vector. Assuming LDS type of multiple access, (1) shows an example of UE signature vector extension. Define $\boldsymbol{s}_R$ as a real number vector obtained by stacking the real and imaginary parts of signature vector 1 $\boldsymbol{s}_1$ and vector 2 $\boldsymbol{s}_2$ into one column. Similarly, define $\boldsymbol{x}_R$ as a real number vector obtained by stacking the real and imaginary parts of the extended UE signature vector $\boldsymbol{x}$. UE signature vector extension can be achieved by multiplying $\boldsymbol{s}_R$ with a transformation matrix $\boldsymbol{U}$.

$$\left.\begin{aligned}\boldsymbol{s}_R &= [\text{real}(\boldsymbol{s}_1)\ \text{real}(\boldsymbol{s}_2)\ \text{imag}(\boldsymbol{s}_1)\ \text{imag}(\boldsymbol{s}_2)]^T \\ \boldsymbol{x}_R &= [\text{real}(\boldsymbol{x})\ \text{imag}(\boldsymbol{x})]^T\end{aligned}\right\} \boldsymbol{x}_R = \boldsymbol{U}\boldsymbol{s}_R . \quad (1)$$

5.4 MUSA

MUSA is a non-orthogonal multiple access scheme operating in the code domain. Conceptually, each UE's modulated data symbols are spread by a specially designed sequence which can facilitate robust SIC implementation compared to the sequences employed by traditional direct-sequence CDMA (DS-CDMA). Then each UE'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 UE's data from superimposed signal can be performed at the BS side using SIC technology.

By determining the interference between different UEs and system performance, the design of spreading sequence is crucial to MUSA. The spreading sequences should have low cross-correlation and can be non-binary.

For MUSA, a family of complex spreading sequence can be studied to achieve relatively low cross-correlation at very short length. The complex sequence exhibits lower cross-correlation than traditional pseudo random noise (PN) due to the utilization of additional freedom of the imaginary part. The real and imaginary parts of the complex element in the spreading sequence are drawn from a multi-level real value set with uniform distribution. For example, for a 3-value set {-1, 0, 1}, every bit of the complex sequence is drawn from the constellation depicted in **Fig. 14** with equal probability.

5.5 NOCA

Similarly with other spreading based NMA schemes, the basic idea of NOCA is that the data symbols are spread using non-orthogonal sequences before transmission. The spreading can be applied in frequency domain and/or time domain based on configuration. The basic transmitter structure of NOCA is shown in **Fig. 15**, where SF denotes the spreading factors and C^j is the spreading sequence of the jth UE. The original modulated data sequence is first converted into P parallel sequences, then each sequence is mapped onto SF subcarriers. The total number of subcarriers is therefore $P \times SF$ for transmitting the data stream.

The sequences used for NOCA shall have good properties like constant modulus, good auto-correlation and cross-correlation, and low memory and complexity requirements. Multiple spreading factors for flexible adaptation might also be supported in this scheme. As a starting, the sequences could be LTE defined sequences for uplink reference signal for 1 RB case, and the spreading factor equals to 12 [26]. For the case in which the spreading factor equals to 12, there are 30 available

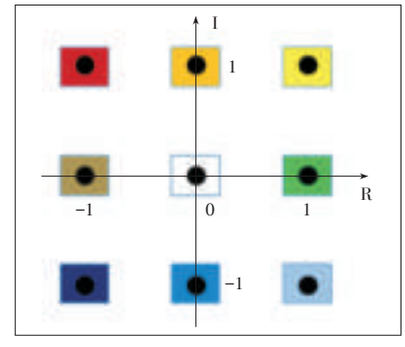
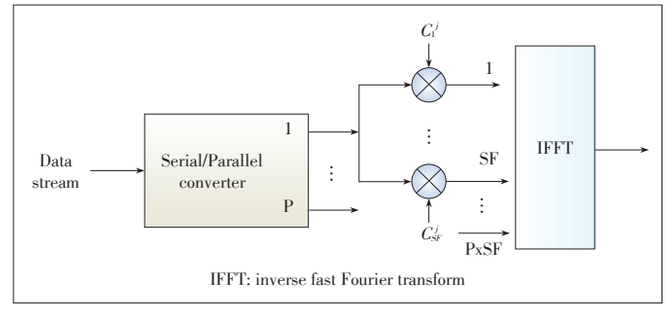


Figure 14. ▶ Elements of complex spreading sequence.



▲Figure 15. NOCA transmitter structure.

roots as specified for LTE system; therefore, together with 12 available cyclic shifts for each root, the number of available sequences for spreading is 360. Besides, these sequences are QPSK sequences, and thus have constant modulus and low cubic metric.

5.6 NCMA

For NMA, multi-UE interference is inherently induced. NCMA is proposed to minimize the multi-UE interference theoretically, based on the spreading codes with the minimum correlation. NCMA is a NMA scheme based on the resource spreading by using non-orthogonal codewords, which is composed of the codewords obtained by Grassmannian line packing problem [27].

The transceiver structure of NCMA is illustrated in **Fig. 16**, where the UE specific non-orthogonal code cover (NCC) represents a non-orthogonal codeword allocated to each UE.

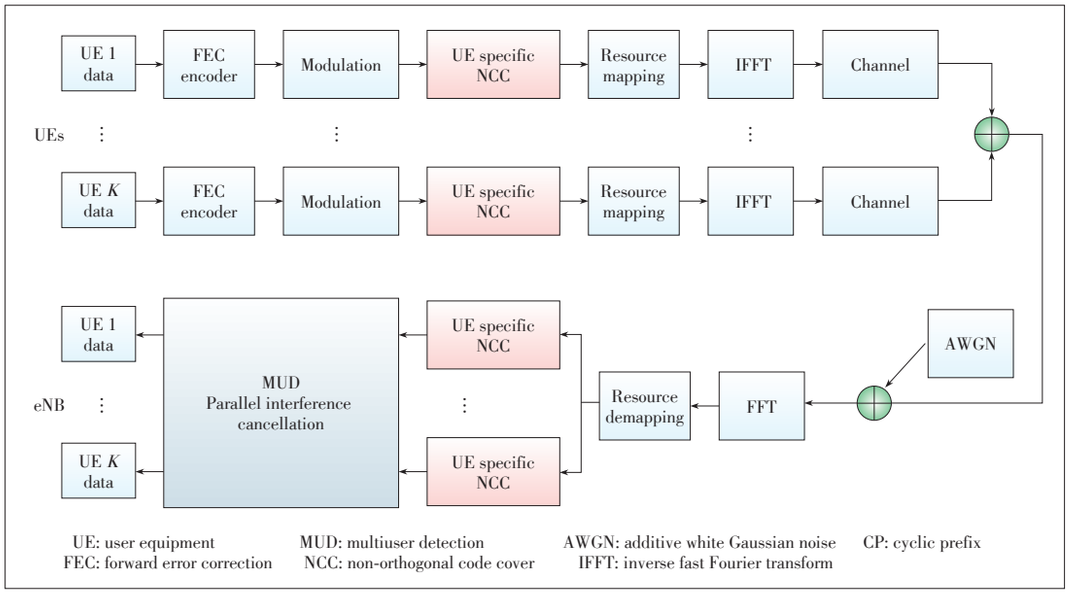
NCMA can provide the additional throughput or improved connectivity with a small loss of block error rate (BLER) in specific environments, by exploiting additional layers through superposed symbol. It can also satisfy quality of service (QoS) constraints. Since the receiver of NCMA system is available for PIC, the multi-UE detection can be implemented with low complexity.

5.7 LCRS

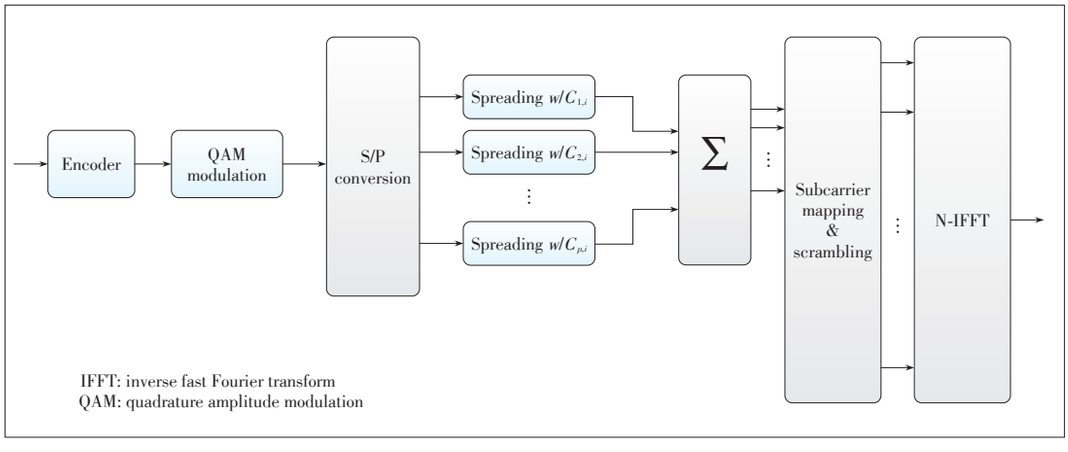
The basic idea of LCRS is to spread information bits over the entire non-orthogonal transmission zone with repetition and rate matching, i.e. combining channel coding with spreading via low rate codes to maximize the coding gain. In this case, a UE-specific channel interleaver [22] can be further employed for improved multi-UE signal separation at the receiver. **Fig. 17** shows a block diagram of transmitter processing at the UE.

6 Conclusions

Uplink NMA is identified as an important candidate 5G technology to provide higher spectral efficiency and support



◀Figure 16. Example of transceiver structure of uplink NCMA.



◀Figure 17. Direct full-frequency spreading with multiple codes at the UE transmitter, assuming that OFDM waveform is used.

connection of more user devices. Based on the latest standardization progress, this paper gives a comprehensive survey on the twelve NMA schemes being considered in 3GPP. As a matter of fact, extensive investigations on these NMA schemes are still ongoing in 3GPP and it is believed that only one or a few NMA schemes will be selected to the final stage of the 5G New Radio (NR) standard according to the consensus reached among the individual members of 3GPP organization.

References

[1] *IMT Vision - Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, ITU-R M.2083-0, Sept. 2015.
[2] NTT DOCOMO, "New SID proposal: study on new radio access technology," RP-160671, 3GPP TSG RAN Meeting #71, Göteborg, Sweden, Mar. 2016.
[3] IMT-2020 (5G) Promotion Group, "White paper on 5G vision and requirements," May 2014.
[4] METIS, "Components of a new air interface-building blocks and performance," ICT-317669-METIS/D2.3, Apr. 2014.
[5] IMT-2020 (5G) Promotion Group, "White paper on 5G wireless technology architecture," May 2015.
[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, USA: Wiley, 2006.
[7] Q. Bi, L. Liang, S. Yang, and P. Chen, "Non-orthogonal multiple access technology for 5G systems," *Telecommunication Science*, vol. 31, no. 5, article ID 2015137, May 2015.
[8] ETSI, "Final report of 3GPP TSG RAN1 #84bis v1.0.0," R1-165448, 3GPP TSG RAN WG1 Meeting #84bis, Busan, Korea, Apr. 2016.
[9] ETSI, "Final report of 3GPP TSG RAN1 #86 v1.0.0," R1-1608562, 3GPP TSG RAN WG1 Meeting #86, Gothenburg, Sweden, Aug. 2016.
[10] NTT DOCOMO, "Initial views and evaluation results on non-orthogonal multiple access for NR," R1-165175, 3GPP TSG RAN WG1 Meeting #85, Nanjing, China, May 2016.
[11] Qualcomm, "RSMA," R1-164688, 3GPP TSG RAN WG1 Meeting #85, Nanjing, China, May 2016.
[12] ETRI, "Low code rate and signature based multiple access scheme for New Radio," R1-164869, 3GPP TSG RAN WG1 Meeting #85, Nanjing, China, May 2016.
[13] Nokia, Alcatel-Lucent Shanghai Bell, "Performance of Interleave Division Multiple Access (IDMA) in combination with OFDM family waveforms," R1-165021, 3GPP TSG RAN WG1 Meeting #85, Nanjing, China, May 2016.
[14] Samsung, "Non-orthogonal multiple access candidate for NR," R1-163992, 3GPP TSG RAN WG1 Meeting #85, Nanjing, China, May 2016.
[15] Huawei and HiSilicon, "LLS results for uplink multiple access," R1-164037, 3GPP TSG RAN WG1 Meeting #85, Nanjing, China, May 2016.
[16] CATT, "Candidate solution for new multiple access," R1-163383, 3GPP TSG RAN WG1 Meeting #84bis, Busan, Korea, Apr. 2016.
[17] Fujitsu, "Initial LLS results for UL non-orthogonal multiple access," R1-164329, 3GPP TSG RAN WG1 Meeting #85, Nanjing, China, May 2016.
[18] ZTE, "Contention-based non-orthogonal multiple access for UL mMTC," R1-164269, 3GPP TSG RAN WG1 Meeting #85, Nanjing, China, May 2016.
[19] Nokia and Alcatel-Lucent Shanghai Bell, "Non-orthogonal multiple access for new radio," R1-165019, 3GPP TSG RAN WG1 Meeting #85, Nanjing, China, May 2016.
[20] LG Electronics, "Considerations on DL/UL multiple access for NR," R1-162517, 3GPP TSG RAN WG1 Meeting #84bis, Busan, Korea, Apr. 2016.
[21] Intel Corporation, "Multiple access schemes for new radio interface," R1-162385, 3GPP TSG RAN WG1 Meeting #84bis, Busan, Korea, Apr. 2016.
[22] L. Ping, 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.
[23] H. Nikopour and H. Baligh, "Sparse code multiple access," in *IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, London, UK, 2013. doi: 10.1109/PIMRC.2013.6666156.
[24] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *IEEE 80th Vehicular Technology Conference (VTC Fall)*, Vancouver, Canada, 2014. doi: 10.1109/VTCFall.2014.6966170.
[25] Huawei and HiSilicon, "Applying multi-dimensional modulation to non-orthogonal multiple access," R1-162163, RAN1#84bis, Busan, South Korea, Apr. 2016.
[26] *Evolved Universal Terrestrial Radio Access (E-UTRA): Physical Channels And Modulation*, 3GPP TS 36.211 V12.2.0, Jun. 2014.
[27] 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: 2016-09-30

Biographies

YANG Shan (yangshan.bri@chinatelecom.cn) received her M.S. degree in commnuication and information system in 2012 from Beijing University of Posts and Telecommunications (BUPT), China. She works with China Telecom Technology Innvotation Center. She is now the 3GPP RAN4 and RAN1 delegate of China Telecom, and has submitted more than 200 contributions to 3GPP. Her reasearch interestes include 5G multiple access technology, baseband advacned receiver, and radio frequency requirements.

CHEN Peng (chenpeng.bri@chinatelecom.cn) received his Ph.D. degree in commnuication and information system in 2006 from BUPT, China. He is the director of China Telecom Technology Innvotation Center, and his reasearch interestes include 5G air interface, network artchitecutre, and 3GPP 5G standaridzaion work.

LIANG Lin (lianglin.bri@chinatelecom.cn) received his M.S. degree in commnuication and information system in 2013 from BUPT, China. He works with China Telecom Technology Innvotation Center. His reasearch interestes include baseband advacned receiver, channel coding, massive MIMO and 3GPP 5G standaridzaion work.

ZHU Jianchi (zhujc.bri@chinatelecom.cn) received his M.S. degree in commnuication and information system in 2007 from BUPT, China. He works with China Telecom Technology Innvotation Center. His reasearch interestes include 5G multiple access techonology, ultra dense network, massive MIMO and 3GPP 5G standaridzaion work.

SHE Xiaoming (shexm.bri@chinatelecom.cn) received his Ph.D. degree in commnuication and information system in 2004 from Tsinghua University, China. He is the vice director of China Telecom Technology Innvotation Center, and his reasearch interestes include 5G air interface, network artchitecutre, and 3GPP 5G standaridzaion work.

DOI: 10.3969/j. issn. 1673-5188. 2017. S1. 005
http://kns.cnki.net/kcms/detail/34.1294.TN.20170628.1459.002.html, published online June 28, 2017

Massive MIMO 5G Cellular Networks: mm-Wave vs. μ-Wave Frequencies

Stefano Buzzi and Carmen D'Andrea
(University of Cassino and Lazio Meridionale, I-03043 Cassino (FR), Italy)

Abstract

Enhanced mobile broadband (eMBB) is one of the key use-cases for the development of the new standard 5G New Radio for the next generation of mobile wireless networks. Large-scale antenna arrays, a.k.a. massive multiple-input multiple-output (MIMO), the usage of carrier frequencies in the range 10–100 GHz, the so-called millimeter wave (mm-Wave) band, and the network densification with the introduction of small-sized cells are the three technologies that will permit implementing eMBB services and realizing the Gbit/s mobile wireless experience. This paper is focused on the massive MIMO technology. Initially conceived for conventional cellular frequencies in the sub-6 GHz range (μ-Wave), the massive MIMO concept has been then progressively extended to the case in which mm-Wave frequencies are used. However, due to different propagation mechanisms in urban scenarios, the resulting MIMO channel models at μ-Wave and mm-Wave are radically different. Six key basic differences are pinpointed in this paper, along with the implications that they have on the architecture and algorithms of the communication transceivers and on the attainable performance in terms of reliability and multiplexing capabilities.

Keywords

millimeter wave; microwave; channel modeling; massive MIMO; doubly massive MIMO

1 Introduction

Fifth-generation (5G) wireless networks are expected to provide 1000x improvement on the supported data rate, as compared to current LTE networks. Such an improvement will be mainly achieved through the concurrent use of three factors [1]: (a) the reduction in the size of the radio-cells, so that a larger data-rate density can be achieved; (b) the use of large-scale antenna arrays at the base stations (BSs), i.e., massive multiple-input multiple-output (MIMO) [2], so that several users can be multiplexed in the same time-frequency resource slot through multiuser MIMO (MU-MIMO) techniques; and (c) the use of carrier frequencies in the range 10 GHz–100 GHz, a.k.a. millimeter-waves (mm-Waves) [3], so that larger bandwidths become available. The factor (a), i.e. the densification of the network, is actually a trend that we have been observing for some decades, in the sense that the size of the radio cells has been progressively reduced over time from one generation of cellular networks to the next one. Differently, factor (b) can be seen as a sort of 4.5G technology, in the sense that the latest 3GPP LTE releases already include the possibility to equip BS with antenna arrays of up to 64 elements. This trend will certainly continue in the future 5G New Radio standard, since the potentialities of massive MIMO are currently being tested worldwide in a number of real-world experiments (for instance, [4] and [5]). The use of mm-Waves, on the contrary, is a more recent technology, at least as far as wireless cellular applications are concerned, and, although there is no doubt that future cellular networks will rely on these technologies, mm-Waves can be certainly classified as a true 5G technology.

Focusing on the massive MIMO technology, most of the research and experimental work has mainly considered its use at conventional cellular frequencies (e.g. sub-6 GHz). We denote here such a range of frequencies as μ-Wave, to contrast them with the above-6 GHz frequencies that we denote as mm-Wave[1]. Only recently, the combination of the massive MIMO concept with the use of mm-Wave frequency bands has started being considered [6], [7]. As a matter of fact, the channel propagation mechanisms at μ-Wave frequencies are completely different from those at mm-Waves. As an instance, the rich-scattering environment at μ-Wave in urban environments is observed [8], thus implying that the MIMO channel is customarily modeled as the product of a scalar constant when the shadowing effects and path loss times a matrix with independent

[1] Strictly speaking, the mm-Wave bands correspond to carrier frequencies larger than 30 GHz.

and identically distributed (i.i.d.) entries are taken into account. At mm-Waves, instead, propagation is mainly based on Line-of-Sight (LOS) propagation and on one-hop reflections, and blockage phenomena are more frequent. To capture these mechanisms, a finite-rank clustered channel model is usually employed [9]–[11]. This paper compares massive MIMO systems at μ-Waves with massive MIMO systems at mm-Waves. We observe that these two different channel models have key implications on the achievable performance, on the multiplexing capabilities of the channels themselves, on the beamforming strategies that can be employed, on the transceiver algorithms and on the adopted channel estimation procedures. Six key differences between massive MIMO systems at μ-Waves and massive MIMO systems at mm-Waves are thus identified and critically discussed.

The rest of this paper is organized as follows. Section 2 describes the considered transceiver model and the massive MIMO channel models at μ-Waves and at mm-Wave frequencies. Section 3, the core of the paper, is divided in six subsections, each one describing a key difference between the massive MIMO channels at μ-Wave and at mm-Wave frequencies; numerical results are also shown here in order to provide experimental evidence of the theoretical discussion. Finally, concluding remarks are given in Section 4.

2 System and Channel Models

In this section, we briefly illustrate the considered transceiver architecture and review the main characteristics of the MIMO wireless channel at μ-Wave and mm-Wave carrier frequencies.

We consider a MIMO wireless link with N_T antennas at the transmitter and N_R antennas at the receiver. We denote by d the distance between the transmitter and receiver, and by M the number of transmitted parallel data streams (i.e., the multiplexing order). The considered transceiver model is shown in **Fig. 1**.

2.1 μ-Wave Channel Model

Assuming frequency-flat fading (i.e. either multipath may be neglected or it is nulled through the use of OFDM modulation), at channel frequencies below 6 GHz, the propagation channel is customarily modelled through an $(N_R \times N_T)$-dimensional matrix, whose $(i,j)^{\text{th}}$ entry, $[\boldsymbol{H}_\mu]_{i,j}$, has the following structure [12], [13]:

$$\left[\boldsymbol{H}_\mu\right]_{i,j} = \sqrt{\beta}\, g_{i,j}, \tag{1}$$

where $g_{i,j}$ represents the small-scale (fast) fading between the i^{th} receive antenna and the j^{th} transmit antenna, and β represents the (slow) large-scale fading (shadowing) and the path-loss between the transmitter and the receiver. In a rich scattering environment, the coefficients $g_{i,j}, i=1,\cdots,N_R, j=1,\cdots,N_T$ are i.i.d. CN (0,1) random variables. The factor β is assumed constant across the transmit and receive antennas (i.e., it does not depend on the indices i,j), and is usually expressed as:

$$\beta = PL 10^{0.1\sigma_{\text{sh}} z}, \tag{2}$$

where PL represents the path loss and $10^{0.1\sigma_{\text{sh}} z}$ represents the shadow fading with the standard deviation σ_{sh} and $z \sim N(0,1)$. With regard to the path loss PL, several models have been derived over the years, based on theoretical models and/or on empirical heuristics. According to the popular three-slope model [13], [14], the path loss in logarithmic units is given by:

$$PL = \begin{cases} -L - 35\log_{10} d, & \text{if } d > d_1 \\ -L - 15\log_{10} d_1 - 20\log_{10} d, & \text{if } d_0 < d \leqslant d_1 \\ -L - 15\log_{10} d_1 - 20\log_{10} d_0, & \text{if } d \leqslant d_0 \end{cases}, \tag{3}$$

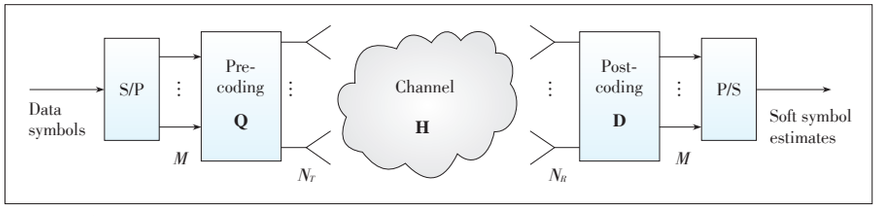
where

$$L = 46.3 + 33.9\log_{10} f - 13.82\log_{10} h_T - \left(1.1\log_{10} f - 0.7\right)h_R + 1.56\log_{10} f - 0.8, \tag{4}$$

with f the carrier frequency in MHz, h_T the transmitter antenna height in meters, and h_R the receiver antenna height in meters. Given the fact that the small-scale fading contribution to the entries of the matrix $\boldsymbol{H}_\mu$ are i.i.d random variates, the channel matrix has full-rank with probability 1, and its rank is equal to the minimum value between N_T and N_R.

2.2 mm-Wave Channel Model

At mm-Waves, propagation mechanisms are different from those at μ-Waves. Indeed, path loss is much larger, while diffraction effects are practically negligible, thus implying that the typical range in cellular environments is usually not larger than 100 m, and the non-LOS component is mainly based on reflections. Moreover, signal blockages, due to the presence of macroscopic obstacles between the transmitter and the receiver, are much more frequent than those at μ-Wave frequencies. In order to catch these peculiarities, general consensus has been reached on the so-called clustered channel model [7], [15]– [18]. This model is based on the assumption that the propagation environment is made of N_{cl} scat-



▲**Figure 1. The considered transceiver model.**

tering clusters, each of which contributes with N_{ray} propagation paths, plus a possibly present LOS component. Apart from the LOS component, the transmitter and the receiver are linked through single reflections on the N_{cl} scattering clusters. Assuming again frequency-flat fading and focusing on a bi-dimensional model for the sake of simplicity, the baseband equivalent of the propagation channel is now represented by an $(N_R \times N_T)$-dimensional matrix expressed as:

$$\boldsymbol{H} = \gamma \sum_{i=1}^{N_{cl}} \sum_{l=1}^{N_{ray}} \alpha_{i,l} \sqrt{L(r_{i,l})}\, \boldsymbol{a}_r(\phi_{i,l}^r)\boldsymbol{a}_t^H(\phi_{i,l}^t) + \boldsymbol{H}_{LOS}. \quad (5)$$

In the above equation, we denote by $\phi_{i,l}^r$ and $\phi_{i,l}^t$ the angles of arrival and departure of the l^{th} ray in the i^{th} scattering cluster, respectively. The quantities $\alpha_{i,l}$ and $L(r_{i,l})$ are the complex path gain and the attenuation associated to the $(i,l)^{\text{th}}$ propagation path. Following [10], the attenuation $L(r_{i,l})$ of the $(i,l)^{\text{th}}$ path is written in logarithmic units as:

$$L(r_{i,l}) = -20\log_{10}\left(\frac{4\pi}{\lambda}\right) - 10n\left[1 - b + \frac{bc}{\lambda f_0}\right]\log_{10}(r_{i,l}) - X_\sigma, \quad (6)$$

with λ the wavelength, c the speed of light, n the path loss exponent, X_σ the zero-mean, σ^2 −variance Gaussian-distributed shadow fading term in logarithmic units, b a system parameter, and f_0 a fixed reference frequency, the centroid of all the frequencies represented by the path loss model. The values for all these parameters for the four-different use-case scenarios discussed in [10] (Urban Microcellular (UMi) Open-Square, UMi Street-Canyon, Indoor Hotspot (InH) Office, and InH Shopping Mall) are reported in **Table 1**.

The complex gain $\alpha_{i,l} \sim CN(0, \sigma_{\alpha_i}^2)$, with $\sigma_{\alpha_i}^2 = 1$ [15]. The factors $\boldsymbol{a}_r(\phi_{i,l}^r)$ and $\boldsymbol{a}_t(\phi_{i,l}^t)$ represent the normalized receive and transmit array response vectors evaluated at the corresponding angles of arrival and departure; for an uniform linear array (ULA) with half-wavelength inter-element spacing we have $\boldsymbol{a}_t(\phi_{i,l}^t) = \frac{1}{\sqrt{N_T}}\left[1, e^{-j\pi\sin\phi_{i,l}^t}, \cdots, e^{-j\pi(N_T-1)\sin\phi_{i,l}^t}\right]^T$. A similar expression can be also given for $\boldsymbol{a}_t(\phi_{i,l}^t)$. Finally, $\gamma = \sqrt{\frac{N_T N_R}{N_{cl} N_{ray}}}$ is a normalization factor that ensures the received signal power scales linearly with the product $N_T N_R$. Regarding the LOS component, the arrival and departure angles corresponding to the LOS link are denoted by ϕ_{LOS}^r and ϕ_{LOS}^t, and we assume that

$$\boldsymbol{H}_{LOS} = I_{LOS}(d)\sqrt{N_T N_R L(d)}\, e^{j\vartheta}\boldsymbol{a}_r(\phi_{LOS}^r)\boldsymbol{a}_t^H(\phi_{LOS}^t). \quad (7)$$

In the above equation, $\theta \sim U(0, 2\pi)$ and $I_{LOS}(d)$ is a random variate indicating the existence of a LOS link between transmitter and receiver. A detailed description of all the parameters needed for the generation of sample realizations for the channel model in (5) is reported in [9]. Comparing the channel model in (5) for mm-Wave frequencies with the one in (1) for μ-Wave frequencies, it is immediately evident that the channel in (5) is a parametric channel model whose rank is tied to the number of clusters and reflectors contributing to the transmitter-receiver link. The next section will provide an accurate description of the implications that these two radically different channel models have on the architecture and on the attainable performance of massive MIMO multiuser wireless systems operating at μ-Wave and at mm-Wave frequencies.

Table 1. Parameters for the path loss model at mm-Waves for four different use-case scenarios

Scenario	Model Parameters
UMi Street Canyon LOS	$n = 1.98$, $\sigma = 3.1$ dB, $b = 0$
UMi Street Canyon NLOS	$n = 3.19$, $\sigma = 8.2$ dB, $b = 0$
UMi Open Square LOS	$n = 1.85$, $\sigma = 4.2$ dB, $b = 0$
UMi Open Square NLOS	$n = 2.89$, $\sigma = 7.1$ dB, $b = 0$
InH Indoor Office LOS	$n = 1.73$, $\sigma = 3.02$ dB, $b = 0$
InH Indoor Office NLOS	$n = 3.19$, $\sigma = 8.29$ dB, $b = 0.06$, $f_0 = 24.2$ GHz
InH Shopping Mall LOS	$n = 1.73$, $\sigma = 2.01$ dB, $b = 0$
InH Shopping Mall NLOS	$n = 2.59$, $\sigma = 7.40$ dB, $b = 0.01$, $f_0 = 39.5$ GHz

3 mm-Wave vs. μ-Wave Massive MIMO

In the following, we highlight and discuss six key differences between μ-Wave and mm-Wave massive MIMO systems.

3.1 Doubly Massive MIMO at mm-Waves

The idea of a large scale antenna array was originally launched by Marzetta in his pioneering paper [12] with reference to BSs. The paper showed that in the limit of a large number of base station antennas small-scale fading effects vanish by virtue of channel hardening, and that channel vectors from the BS to the users tend to become orthogonal; consequently, plain channel-matched beamforming at the BS permits serving several users on the same time-frequency resource slot with (ideally) no interference, and the only left impairment is imperfect channel estimates due to the fact that orthogonal pilots are limited and they must be re-used throughout the network (this is the so-called pilot contamination effect, discussed in the following). Reference [12] considered a system where mobile users were equipped with just one antenna. Successive studies have extended the massive MIMO idea at μ-Wave frequencies to the case in which the mobile devices have multiple antennas, but this number is obviously limited to few units. Indeed, at μ-Wave frequencies the wavelength is in the order of several centimeters, and it is thus difficult to pack many antennas on small-sized user devices. At μ-Waves, thus, massive MIMO just refers to BSs. Things are instead different at mm-Waves,

wherein multiple antennas are necessary first and foremost to compensate for the increased path loss with respect to conventional sub-6 GHz frequencies. At mm-Waves, the wavelength is on the order of millimeters, and, at least in principle, a large number of antennas can be mounted not only on the BS, but also on the user device. As an example, at a carrier frequency of 30 GHz the wavelength is 1 cm, and for a planar antenna array with $\lambda/2$ spacing, more than 180 antennas can be placed in an area as large as a standard credit card (8.5 cm x 5.5 cm); this number climbs up to 1300 at a carrier frequency of 80 GHz. This consideration leads to the concept of doubly massive MIMO system [7], which is defined as a wireless communication system where the number of antennas grows large at both the transmitter and the receiver. Of course, there are a number of serious practical constraints—e.g., large power consumption, low efficiency of power amplifiers, hardware complexity, ADC and beamformer implementation—that currently prevent the feasibility of a user terminal equipped with hundreds of antennas. Mobile devices with a massive number of antennas thus will not be available in a few years, but, given the intense pace of technological progress, sooner or later they will become reality. As far as long-term forward-looking theoretical research is concerned, we believe that doubly-massive MIMO systems at mm-Waves will be a popular research topic for years to come.

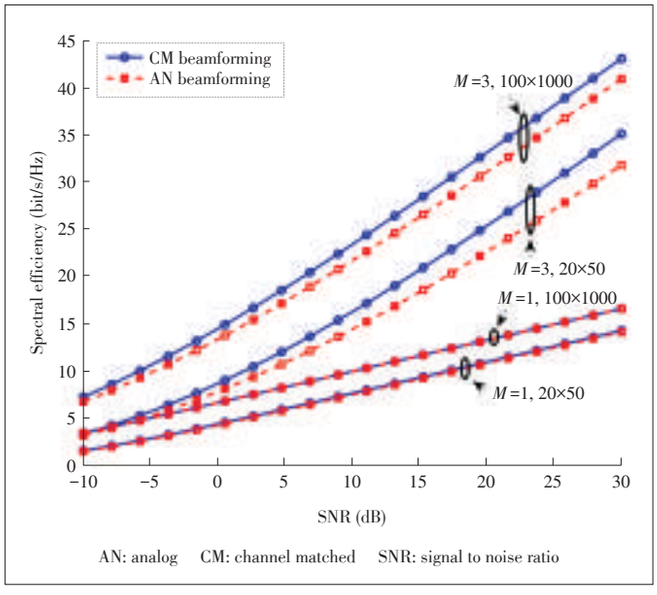
3.2 Analog (Beam-Steering) Beamforming Optimal

One problem with massive MIMO systems is the cost and the complexity of needed hardware to efficiently exploit a so large number of antennas. If fully digital beamforming is to be made, as many RF chains will be needed as the number of antennas; consequently, energy consumption will also grow linearly with the number of antennas. In order to circumvent this problem, lower complexity architectures have been proposed, encompassing, for instance, 1-bit quantization of the antenna outputs [19] and hybrid analog/digital beamforming structures [11], [18], [20], wherein an RF beamforming matrix (whose entries operate as simple phase shifters) is cascaded to a reduced-size digital beamformer. The authors of the paper [21] has shown that if the number of RF chains is twice the multiplexing order, the hybrid beamformer is capable of implementing any fully digital beamformer. Now, while at μ-Waves the use of hybrid beamformer brings an unavoidable performance degradation, at mm-Waves something different happens in the limiting regime of large number of antennas by virtue of the different propagation mechanisms. Indeed, the channel matrix in (5) can be compactly re-written as:

$$\boldsymbol{H}=\gamma\sum_{i=1}^{N}\alpha_i\boldsymbol{a}_r\left(\phi_i^r\right)\boldsymbol{a}_t^H\left(\phi_i^t\right), \tag{8}$$

where we lump the coefficients α_i into the path-loss term, and group the two summations over the clusters and the rays in just one summation, with N being the number of propagation paths from the transmitter to the receiver. Given the continuous random location of the scatterers, the set of arrival angles will be different with probability 1, i.e. there is a zero probability that two distinct scatterers will contribute to the channel with the same departure and arrival angles. Since, for a large number of antennas, we have $\boldsymbol{a}_x^H\left(\phi_p^x\right)\boldsymbol{a}_x\left(\phi_q^x\right)\to 0$, provided that $\phi_p^x\neq\phi_q^x, x=\{r,t\}$, we can conclude that for large N_T, the vectors $a_t\left(\phi_i^t\right)$ for all $i=1,\cdots,N$ converge to an orthogonal set, and, similarly, for large N_R, the vectors $\boldsymbol{a}_r\left(\phi_i^r\right)$ for all $i=1,\cdots,N$ converge to an orthogonal set as well. Accordingly, in the doubly massive MIMO regime, the array response vectors $\boldsymbol{a}_r(\cdot)$ and $\boldsymbol{a}_t(\cdot)$ become the left and right singular vectors of the channel matrix, i.e. the channel representation (8) coincides with the singular-value-decomposition of the channel matrix. Under this situation, purely analog (beam-steering) beamforming becomes optimal. Otherwise stated, we have two main consequences. First, in a single-user link, the channel eigendirections associated to the largest eigenvalues are just the beam-steering vectors corresponding to the arrival and departure angles and associated with the predominant scatterers. This suggests that pre-coding and post-coding beamforming simply require pointing a beam towards the predominant scatterer at the transmitter and at the receiver respectively. Second, in a multiuser environment, assuming that the links between the several users and the BS involve separate scatterers and different sets of arrival and departure angles[2], beam-steering analog beamforming automatically results in no-cochannel interference (in the limiting regime of infinite number of antennas) since the beams pointed towards different users tend to become orthogonal. **Fig. 2** provides some experimental evidence of the above statements. We have considered a single-user MIMO link at mm-Waves; the carrier frequency is 73 GHz, the transmitting antenna height is 15 m, while the receiving antenna height is 1.65 m. All the parameters needed for the generation of the mm-Wave channel matrix in (5) are the ones reported in [9] for the "open square model". Fig. 2 shows the system spectral efficiency measured in bit/s/Hz, versus the received signal to noise ratio (SNR), and it compares the performances of the channel matched (CM) fully digital beamforming and the analog (AN) beam-steering beamforming. With CM beamforming the pre-coding and post-coding beamformers are the left and singular eigenvectors of the channel matrix in (5) associated to the M largest eigenvalues respectively; with AN beamforming, instead, the pre-coding and post-coding beamformers are simply the array responses corresponding to the departure and arrival angles associated to the M dominant scatterers respectively. From the figure it is seen that AN beamforming achieves practically the same performance as CM beamforming for multiplexing order $M=1$, even in the case of not-so-large

[2] This is a quite reasonable assumption for sufficiently spaced mobile user locations.



▲**Figure 2. Spectral efficiency of a mm-Wave MIMO wireless link vs. received SNR for CM-FD beamforming and AN (beam-steering) beamforming, for two different values of the number of transmit and receive antennas and of the multiplexing order of the system.**

number of antennas, while there is a small gap for $M=3$; this gap is supposed to get reduced as the number of antennas increases.

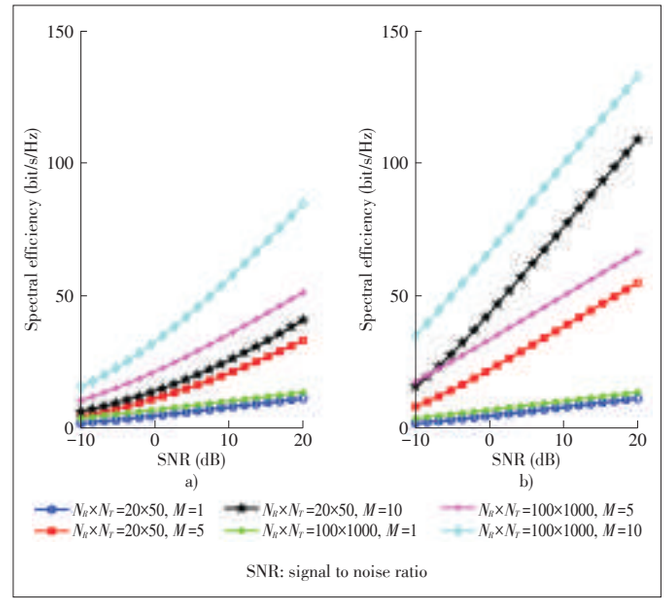
3.3 Rank of the Channel Not Increasing with N_T and N_R

At μ-Wave frequencies, the i.i.d. assumption for the small-scale fading component of the channel matrix $\boldsymbol{H}$ guarantees that with probability 1 the matrix has rank equal to min (N_T, N_R). Consequently, as long as the rich-scattering environment assumption holds and the number of degrees of freedom of the radiated and scattered fields is sufficiently high [22], the matrix rank increases linearly with the number of antennas. At mm-Wave frequencies, instead, the validity of the channel model in (5) directly implies that, including the LOS component, the channel has at most the rank $N_{cl}N_{ray}+1$, since it is expressed as the sum of $N_{cl}N_{ray}+1$ rank-1 matrices. This rank is clearly independent of the number of transmit and receive antennas, so, mathematically, as long as min $(N_T, N_R) > N_{cl}N_{ray}+1$, increasing the number of antennas has no effect on the channel rank. However, it is also suggested that, for increasing number of antennas, the directive beams become narrower and narrower and more scatterers can be resolved, which implies that the channel rank increases (even though probably not linearly) with the number of antennas. However, this is still a conjecture that would need experimental validation.

With respect to the number of antennas, the described different behavior of the channel rank has a profound impact on the multiplexing capabilities of the channel. Indeed, for μ-Wave systems, the increase in the channel rank leads to an increase of the multiplexing capabilities of the channel; on the other hand, the multiplexing capabilities depend on the number of scatterers in the propagation environment in mm-Wave systems, while the number of antennas just contributes to the increase of the received power that can increase proportionally to the product $N_T N_R$. **Fig. 3** provides experimental evidence of such a different behavior. The figure shows the system spectral efficiency for mm-Wave and μ-Wave wireless MIMO links, for two different values of the number of receive and transmit antennas, and for three different values of the multiplexing order M. The parameters of the mm-Wave channel are the same as those in Fig. 2. Regarding the μ-Wave channel, a carrier frequency equal to 1.9 GHz is considered and the standard deviation of the shadow fading σ_{sh} is taken equal to 8 dB, while the parameters of the three-slope path loss model in (3) are $d_1=50$ m and $d_2=100$ m. It is clearly seen from Fig. 3 that the μ-Wave channel has larger multiplexing capabilities than the mm-Wave channel; the gap between the two scenarios is mostly emphasized for the large values of M and for $N_R \times N_T = 100 \times 1000$.

3.4 Channel Estimation Simpler

In μ-Wave massive MIMO systems, channel estimation is a rather difficult and resource-consuming task, since it requires the separate estimation of each entry of the matrix $\boldsymbol{H}$. It thus follows that in a multiuser system with K users equipped with N_R antennas each, the number of parameters to be estimated is KN_RN_T, where N_T denotes the number of antennas at the BS. The attendant computational complexity needed to perform channel estimation is a growing function of the number of used

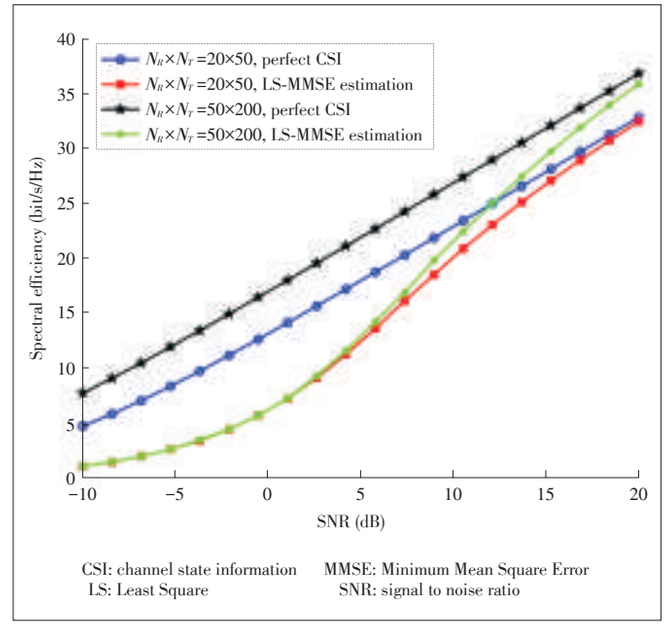


▲**Figure 3. a) Spectral efficiency vs. received SNR for an mm-Wave channel varying the number of transmit and receive antennas and multiplexing order, and b) spectral efficiency vs. received SNR for an μ-Wave channel varying the number of transmit and receive antennas and multiplexing order.**

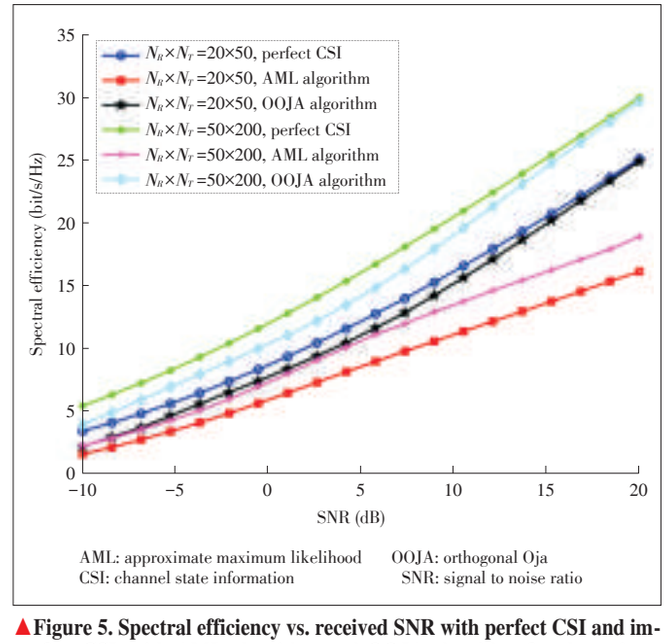
antennas. Additionally, the increase of the number of antennas N_R at the mobile devices has a direct impact on the network capacity. Indeed, let τ_c denote the duration (in discrete samples) of the channel coherence time and τ_p the length (again in discrete samples) of the pilot sequences used on the uplink for channel estimation; since the length of pilot sequences must be a fraction (typically no more than 1/2) of the channel coherence length, and since the use of orthogonal pilots across users requires that $KN_R \leqslant \tau_p < \tau_c$, it is readily seen that we have a physical bound on the maximum number of users and the number of transceiver antennas at the mobile device. Such a bound is the main underlying motivation for the fact that a considerable share of the available literature on massive MIMO systems at μ-Waves focuses on the case of single-antenna mobile devices, and with $N_R = 1$, the number of users K can be taken larger. Additionally, to increase the number of supported users, pseudo-orthogonal pilots with low cross-correlation are used, even though this leads to the well-known pilot contamination problem that, as discussed in the sequel, is the ultimate performance limit in μ-Waves massive MIMO systems [12].

At mm-Wave frequencies, instead, the clustered channel model of (5) is basically a parametric model, and the number of parameters is essentially independent of the number of antennas. Based on this consideration, the computational complexity of the channel estimation schemes at mm-Waves may be smaller than that at μ-Waves. Channel estimation for mm-Wave frequencies is a research track that is currently under development, whereas for μ-Waves this is a rather mature area. Among the several existing approaches to perform channel estimation at mm-Waves, the most considered ones rely either on compressed sensing or on subspace methods. As an example, reference [23] shows that at mm-Waves, for increasing number of antennas, the most significant components of the received signal lie in a low-dimensional subspace due to the limited angular spread of the reflecting clusters. This low-dimensionality feature can be exploited in order to obtain channel estimation algorithms based on the sampling of only a small subset rather than of the whole number of antenna elements. Consequently, channel estimation can be performed using a reduced number (with respect to the number of receive antennas) of required RF chains and A/D converters at receiver front-end. Reference [24], instead, develops subspace-based channel estimation methods exploiting channel reciprocity in TDD systems, using the well-known Arnoldi iteration and explicitly taking into account the adoption of hybrid analog/digital beamforming structures at the transmitter and at the receiver. Subspace methods are particularly attractive in those situations where it is of interest to estimate the principal left and right singular eigenvectors of the channel matrix $\boldsymbol{H}$, which, in the doubly massive MIMO regime, are well-approximated by the array response vectors corresponding to the dominant scatterers. As done in [25], applying fast subspace estimation algorithms such as the Oja's one [26], the dominant channel eigenvectors can be directly obtained by the sample estimate of the data covariance matrix, with no need to directly estimate the whole channel matrix $\boldsymbol{H}$.

Figs. 4 and **5** show numerical results concerning channel estimation at μ-Wave and at mm-Wave channel frequencies. In particular, both figures report the spectral efficiency vs. the received SNR for two different antenna configurations and by contrasting the case of perfect channel state information (CSI)



▲**Figure 4. Spectral efficiency vs. received SNR with perfect CSI and imperfect CSI, with LS-MMSE algorithm for the estimation of μ-Wave channel. The multiplexing order is 3.**



▲**Figure 5. Spectral efficiency vs. received SNR with perfect CSI and imperfect CSI, with AML algorithm and OOJA algorithm for the estimation of mm-Wave channel. The multiplexing order is 3.**

with the case in which the channel is estimated based on training pilots. In both figures a single-user MIMO link is considered, and channel estimation is carried out assuming that each transmit antenna sends an orthogonal pilot. The number of signaling intervals devoted to channel estimation coincides with the number of transmit antennas. Note that this is the minimum possible duration in order to be able to send orthogonal pilots. Channel estimation at μ-Wave frequencies (Fig. 4) is made using the linear minimum mean square errors criterion ([27]), while at mm-Wave frequencies (Fig. 5) the approximate maximum likelihood (AML) algorithm of [23] and the orthogonal Oja (OOJA) algorithm [25] are used. Comparing the figures, it is clearly seen that the gap between the case of estimated channel and the case of perfect CSI is smaller at mm-Wave frequencies, especially when the OOJA algorithm is considered. Conversely, this gap is larger at μ-Waves, and it grows with the dimension of the user antenna arrays. This behavior can be intuitively explained by virtue of the parametric form of the mm-Wave channel model in (5), which permits the development of efficient channel estimation algorithms.

3.5 Pilot Contamination Less Critical

Pilot contamination is the ultimate disturbance in massive MIMO systems operating at μ-Waves. As already discussed in the previous paragraphs, the impossibility to have a number of orthogonal pilots larger than the number of signaling intervals devoted to channel estimation leads to the use of pseudo-orthogonal, low cross-correlation sequences. Accordingly, in a massive MIMO system, when the MSs transmit their own pilot sequences in the uplink training phase to enable channel estimation at the BSs, every BS learns the channel from the intended MS, and also small pieces of the channels from the other MSs using pilots that are correlated to the one used by the intended MS. This phenomenon, in turn, causes a saturation in the achieved Signal-to-Interference plus Noise-Ratio (SINR) both in the downlink and in the uplink. The deceitful nature of pilot contamination was unveiled by Marzetta in his landmark paper [12] and since then, many authors have deeply investigated its effects and proposed strategies to counterbalance its effects [28], [29], [30]. All of these papers deal with the case of a μ-Wave massive MIMO system.

Pilot contamination at mm-Wave frequencies is instead a much less-studied topic (some initial results are reported in [31]). This is in part due to the fact that massive MIMO at mm-Waves is a more recent research topic than massive MIMO at μ-Waves. On the other hand, it may be envisioned that pilot contamination may be less critical at mm-Waves than it has revealed at μ-Waves, mainly for the short-range nature of mm-Wave links. In particular, while the range of μ-Wave links can be in the order of thousands of meters, the range for mm-Wave links will be more than one order of magnitude smaller, due to the increased path loss and a larger relevance of signal blockages. mm-Wave frequencies will be used for short-range communications in small cells, which, by nature, usually serve a smaller number of users than conventional micro-cells and macro-cells. Therefore, on one hand, the signals transmitted by the MSs during uplink training fade rapidly with the distance, and thus they should not be a serious impairment to surrounding BSs learning the channel from their intended MSs; on the other hand, the reduced number of users in each cell will lead to a less severe shortage of orthogonal pilots. The results in [31] seem to confirm such increased resilience of mm-Waves to the pilot contamination problem.

3.6 Antenna Diversity/Selection Procedures Less Effective

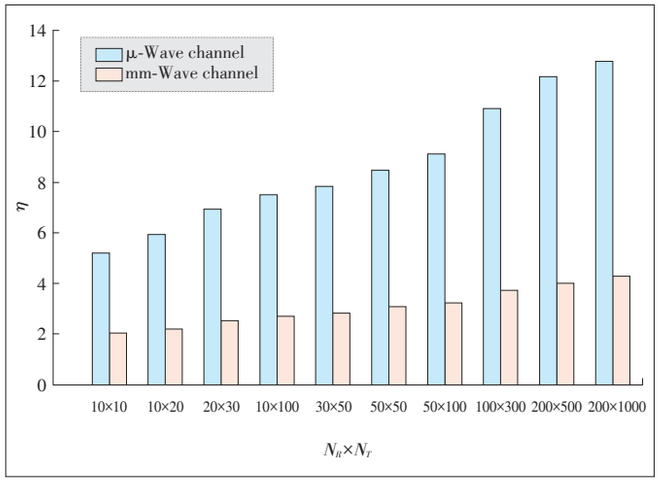
The i.i.d. nature of the fast fading component in the MIMO channel matrix at μ-Waves in (1) leads to a monotonic increase with the number of antennas, of the diversity order that can be attained. In particular, an $N_R \times N_T$ channel brings a diversity order equal to $N_R \times N_T$, thus implying that the average error probability decreases to a zero, in the limit of large Signal-to-Noise Ratio (SNR), as $SNR^{-N_R N_T}$. Such a diversity order can be attained through a simple antenna selection procedure by picking the transmit and receive antennas corresponding to the entry with the largest magnitude in the channel matrix H. Looking at this fact from a different perspective, we can recall the well-known probability result stating that the maximum of a set of positive i.i.d. random variables taking value in the interval $[0,+\infty)$, becomes unbounded as the cardinality of the set diverges. As a consequence, for increasing number of antennas, the probability of observing a very large entry in the channel matrix rapidly increases. The open literature is rich of studies exploiting this peculiarity of μ-Wave MIMO channels and proposing diversity techniques based on antenna selection procedures (e.g. [32] and [33]).

At mm-Waves, instead, given the parametric channel model of (5), a different behavior is observed. In particular, the entries of the matrix channel have no longer an i.i.d. component, and this implies that the maximum of the magnitudes of the entries of $\boldsymbol{H}$ grows at a much reduced pace. As a consequence, diversity techniques using antenna selection procedures are less effective.

As an experimental evidence of this fact, **Fig. 6** compares the parameter η in (9), for different values of $N_R \times N_T$, and for both the μ-Wave and mm-Wave channel models.

$$\eta = \frac{\max_{i,j} \left| \boldsymbol{H}_{i,j} \right|^2}{\mathrm{tr}\left(\boldsymbol{H}^H \boldsymbol{H}\right)/N_T N_R}. \tag{9}$$

The quantity η is the ratio between the largest squared magnitude among the entries of $\boldsymbol{H}$, and the average squared magnitude. The larger η is, the more unbalanced are the magnitudes of the entries of the channel matrix, since η basically measures how far is the largest entry in $\boldsymbol{H}$ from the average magnitude. Fig.6, shows that the parameter η is in general an increasing function of the number of antenna elements, but it



▲**Figure 6. Values of the performance measure η defined in (9) for several antenna array sizes for the mm-Wave and μ-Wave channels.**

grows much more rapidly in the case of μ-Wave channels.

4 Conclusions

This paper outlined a critical comparison between massive MIMO systems at mm-Waves and at μ-Waves. Six key differences were outlined, and their implications on the transceiver architecture and on the attainable performance were discussed and validated also through the result of computer simulations. Among the discussed differences, we believe that the most disruptive one is the first difference, i.e. the fact that MIMO systems may be doubly massive at mm-Waves. Indeed, while it has been shown that the use of large-scale antenna arrays does not have an as beneficial impact on the system multiplexing capabilities as it has at μ-Wave frequencies, the availability of doubly massive MIMO wireless links will enable the generation of very narrow beams, resulting in reduced co-channel interference to other users using the same time-frequency resources. Another key advantage of doubly massive MIMO systems at mm-Waves is the fact that the computational complexity of channel estimation weakly depends on the number of antennas, especially for the case in which analog (beam-steering) beamforming strategies are used. While massive MIMO at μ-Wave frequencies is gradually entering in 3GPP standards, mm-Waves and in particular massive mm-Wave MIMO systems are still under heavy investigation, both in academia and industry. It is however anticipated that sooner or later a technology readiness level will be reached such that they will be included in 3GPP standards. The authors of this paper hope that this article will help to move us forward along this road.

References

[1] J. G. 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] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems, " *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, Feb. 2014. doi: 10.1109/MCOM.2014.6736761.
[3] T. S. 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.
[4] Huawei. (2016, Nov. 16). *Huawei and DOCOMO conduct world's first 5G large scale field trial in the 4.5 GHz band* [Online]. Available: http://www.huawei.com/en/news/2016/11/World-First-5G-Large-Scale-Field-Trial
[5] Ericsson. (2016, Feb.18). *Ericsson 5G field trial gear achieves peak downlink throughput over 25 Gbps with MU-MIMO* [Online]. Available: https://www.ericsson.com/news/1987136
[6] L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive MIMO: the next wireless revolution? " *IEEE Communications Magazine*, vol. 52, no. 9, pp. 56–62, Sept. 2014. doi: 10.1109/MCOM.2014.6894453.
[7] S. Buzzi and C. D'Andrea, "Doubly massive mmWave MIMO systems: Using very large antenna arrays at both transmitter and receiver," in *IEEE Global Communications Conference (GLOBECOM)*, Washington DC, USA, 2016, pp. 1–6. doi: 10.1109/GLOCOM.2016.7841750.
[8] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol.6, no.3, pp. 311–335, 1998. doi:10.1023/A:1008889222784
[9] S. Buzzi and C. D'Andrea, "On clustered statistical MIMO millimeter wave channel simulation," arXiv preprint, arXiv:1604.00648, May 2016.
[10] Aalto University, Nokia, AT&T, et al. (2016, May). *5G channel model for bands up to 100 GHz* [Online]. Available: http://www.5gworkshops.com/5GCM.html, 2015
[11] M. R. Akdeniz, Y. Liu, M. K. Samimi, et al., "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014. doi: 10.1109/JSAC.2014.2328154.
[12] T. L. 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.
[13] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, Mar. 2017. doi: 10.1109/TWC.2017.2655515.
[14] A. Tang, J. Sun, and K. Gong, "Mobile propagation loss with a low base station antenna for NLOS street microcells in urban area," in *Proc. IEEE VTS 53rd Vehicular Technology Conference, Spring 2001*. Rhodes, Greece, 2001, pp. 333–336. doi: 10.1109/VETECS.2001.944859.
[15] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Transaction on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014. doi: 10.1109/TWC.2014.011714.130846.
[16] S. Buzzi, C. D'Andrea, T. Foggi, A. Ugolini, and G. Colavolpe, "Spectral efficiency of MIMO millimeter-wave links with single-carrier modulation for 5G networks," in *Proc. 20th International ITG Workshop on Smart Antennas (WSA 2016)*, Munich, Germany, Mar. 2016.
[17] J. Lee, G. T. Gil, and Y. H. Lee, "Exploiting spatial sparsity for estimating channels of hybrid MIMO systems in millimeter wave communications," in *2014 IEEE Global Communications Conference*, Austin, USA, Dec. 2014, pp. 3326–3331. doi: 10.1109/GLOCOM.2014.7037320.
[18] S. Buzzi and C. D'Andrea, "Are mmWave low-complexity beamforming structures energy-efficient? Analysis of the downlink MU-MIMO," in *Proc. International Workshop on Emerging Technologies for 5G Wireless Cellular Networks*, in conjunction with *2016 IEEE GLOBECOM*, Washington DC, USA, Dec. 2016. doi: 10.1109/GLOCOMW.2016.7848841.
[19] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," arXiv preprint, arXiv:1404.7736v1, Apr. 2014.
[20] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE Access*, vol. 4, pp. 247–267, Jan. 2016. doi: 10.1109/ACCESS.2015.2514261.
[21] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 501–513, Apr. 2016. doi: 10.1109/JSTSP.2016.2520912.
[22] M. D. Migliore, "On the role of the number of degrees of freedom of the field in MIMO channels," *IEEE Transactions on Antennas and Propagation*, vol. 54, no. 2, pp. 620–628, Feb. 2006. doi: 10.1109/TAP.2005.863108.
[23] S. Haghighatshoar and G. Caire, "Massive MIMO channel subspace estimation

from low-dimensional projections," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 303–318, Jan. 2017. doi: 10.1109/TSP.2016.2616336.
[24] H. Ghauch, T. Kim, M. Bengtsson, and M. Skoglund, "Subspace estimation and decomposition for large millimeter-wave MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 528–542, Apr. 2016. doi: 10.1109/JSTSP.2016.2538178.
[25] S. Buzzi and C. D'Andrea, "MIMO channel subspace estimation at mmWave frequencies," presented at the 21st International ITG Workshop on Smart Antennas (WSA 2017), Berlin, Germany, Mar. 2017.
[26] K. Abed-Meraim, S. Attallah, A. Chkeif, and Y- Hua, "Orthogonal Oja algorithm," *IEEE Signal Processing Letters*, vol.7, no. 5, May 2000. doi: 10.1109/97.841157.
[27] M. Biguesh and Alex Gershman, "Training-based MIMO channel estimation: A study of estimator tradeoffs and optimal training signals," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, Mar. 2006. doi: 10.1109/TSP.2005.863008.
[28] R. R. Müller, L. Cottatellucci, and M. Vehkaperä, "Blind pilot decontamination," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 773–786, Oct. 2014. doi: 10.1109/JSTSP.2014.2310053.
[29] N. Krishnan, R. D. Yates, and N. B. Mandayam, "Uplink linear receivers for multi-cell multiuser MIMO with pilot contamination: Large system analysis," *IEEE Transactions on Wireless Communications*, vol. 13, no. 8, pp. 4360–4373, Aug. 2014. doi: 10.1109/TWC.2014.2320914.
[30] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016. doi: 10.1109/TWC.2015.2488634.
[31] S. A. R. Naqvi, S. A. Hassan, and Z. ul Mulk, "Pilot reuse and sum rate analysis of mmWave and UHF-based massive MIMO systems," in *IEEE 83rd Vehicular Technology Conference (VTC Spring)*, Nanjing, China, May 2016, pp. 1–5. doi: 10.1109/VTCSpring.2016.7504425.
[32] S. Sanayei and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 68–73, Oct. 2004. doi: 10.1109/MCOM.2004.1341263.
[33] X. Zhang, Z. Lv, and W. Wang, "Performance analysis of multiuser diversity in MIMO systems with antenna selection," *IEEE Transactions on Wireless Communications*, vol. 7, no. 1, pp. 15–21, Jan. 2008. doi: 10.1109/TWC.2008.060441.

Manuscript received: 2016-11-18

Biographies

Stefano Buzzi (buzzi@unicas.it) is currently an associate professor at the University of Cassino and Lazio Meridionale, Italy. He received the Ph.D. degree in electrical and computer engineering from the University of Naples Federico II, Italy in 1999, and had short-term research appointments at Princeton University, USA in 1999, 2000, 2001 and 2006. He is a former associate editor of *IEEE Signal Processing Letters* and of *IEEE Communications Letters*, while is currently serving as an editor for *IEEE Transactions on Wireless Communications*. Dr. Buzzi's research interests are in the broad field of communications and signal processing, with emphasis on wireless communications. He has co-authored about 150 technical peer-reviewed journal and conference papers, including a highly-cited survey paper "What will 5G be?" (IEEE JSAC, June 2014) on 5G wireless networks.

Carmen D'Andrea (carmen.dandrea@unicas.it) received the B.S. and M.S. degrees, both with honors, in telecommunications engineering from University of Cassino and Lazio Meridionale, Italy in 2013 and 2015. She is currently with the Department of Electrical and Information Engineering at the University of Cassino and Lazio Meridionale, pursuing the Ph.D. degree in electrical and information engineering. Her research interests include wireless communications and signal processing and her current focus is on mm-Wave communications and massive MIMO systems.

DOI: 10.3969/j. issn. 1673-5188. 2017. S1. 006
http://kns.cnki.net/kcms/detail/34.1294.TN.20170623.0824.002.html, published online June 23, 2017

Novel MAC Layer Proposal for URLLC in Industrial Wireless Sensor Networks

Mohsin Raza[1], Sajjad Hussain[2], Hoa Le-Minh[1], and Nauman Aslam[1]
(1. Northumbria University, Newcastle, NE18ST, UK;
2. University of Glasgow, Glasgow, G128QQ, UK)

Abstract

Ultra-reliable and low-latency communications (URLLC) has become a fundamental focus of future industrial wireless sensor networks (IWSNs). With the evolution of automation and process control in industrial environments, the need for increased reliability and reduced latencies in wireless communications is even pronounced. Furthermore, the 5G systems specifically target the URLLC in selected areas and industrial automation might turn into a suitable venue for future IWSNs, running 5G as a high speed inter-process linking technology. In this paper, a hybrid multi-channel scheme for performance and throughput enhancement of IWSNs is proposed. The scheme utilizes the multiple frequency channels to increase the overall throughput of the system along with the increase in reliability. A special purpose frequency channel is defined, which facilitates the failed communications by retransmissions where the retransmission slots are allocated according to the priority level of failed communications of different nodes. A scheduler is used to formulate priority based scheduling for retransmission in TDMA based communication slots of this channel. Furthermore, in carrier-sense multiple access with collision avoidance (CSMA/CA) based slots, a frequency polling is introduced to limit the collisions. Mathematical modelling for performance metrics is also presented. The performance of the proposed scheme is compared with that of IEEE802.15.4e, where the performance is evaluated on the basis of throughput, reliability and the number of nodes accommodated in a cluster. The proposed scheme offers a notable increase in the reliability and throughput over the existing IEEE802.15.4e Low Latency Deterministic Networks (LLDN) standard.

Keywords

industrial wireless sensor network (IWSN); IEEE802.15.4e; Low Latency Deterministic Network (LLDN); low latency communications (LLC); ultra-reliable low latency communication (URLLC)

1 Introduction

The past couple of decades have witnessed a relatively perpetual rise in industrialization and automation of the processes [1]. In the competitive industrial world, automation is the key to cost reduction whether it is a production, nuclear power, oil refinement or chemical plant [2]. These industrial plants can greatly benefit from technological advancements and can implement successful process control with efficient and effective formation of a close loop control system. However, to introduce a suitable process control, a reliable communication infrastructure is needed which should also offer scalable architecture for future enhancements and permits infrastructural extension in the ever-changing industrial plants [3]-[6]. To cope with the first objective (reliability), wired networks offer suitable solutions but some intrinsic properties of these networks do not sit very well with the present-day industries. The lack of scalable architecture and flexibility of industrial wired networks poses serious limitations for dynamic industrial environments. On top of that, the high price tag in the wired networks also comes as a setback. As an alternate to the wired feedback systems in the industrial infrastructure, industrial wireless sensor networks (IWSNs) are also sometimes considered as a more cost-effective approach. However, the less predictable wireless communication links in IWSNs appear to be a major challenge [7]. Therefore, in order to establish a wireless feedback network, the reliability and real-time data delivery must be ensured in such networks.

The IWSNs offer a suitable reduction in the deployment cost as well as the maintenance cost of the feedback communication loop and in some cases these wireless networks offer a cost reduction by a factor of ten or even more [5], [8]. The IWSNs also offer other significant benefits over traditional wired networks. These benefits include scalability, cost efficiency, self-healing ability, reduced planning overload for network formation, and relatively less time for new installations of IWSNs [7]. All these benefits have encouraged the use of IWSNs in in-

dustries, eventually leading to an extensive increase in research and development activities to ensure suitable IWSN solutions for enabling it for wider variety of applications in the industries.

In the past couple of decades, the use of IWSNs has exponentially increased and can primarily be credited to the improvements in Micro Electro Mechanical Systems (MEMS) technology, which enabled the cost and size reduction of the sensor nodes and an increase in processing capabilities and memory capacity as well. The present IWSNs are capable of taking in account the channel conditions, sensor readings, network specifications and suitable responses to the sampled sensor data. These abilities if properly utilized can also serve as a tool to overcome uncertainty in wireless links and timely delivery of the information. All these improvements in IWSNs have encouraged their use in industrial environments. In past few years many industrial protocols and IEEE standards surfaced, which include Zigbee, WirelessHart, 6LoWPAN, WiaPA, ISA100.11a, IEEE802.15.4, and IEEE802.15.4e [9]–[16]. The time-division multiple access (TDMA) based channel access was introduced in IWSNs over traditional carrier-sense multiple access with collision avoidance (CSMA/CA) based access for guaranteed channel access and improved reliability. The research community also contributed in many research oriented solutions targeting improved reliability, network lifetime enhancement and real-time data delivery. A few priority based schemes were also defined to offer hierarchical access to the wireless channel resources. In some cases, the use of multiple channels for enhanced data rates was also considered as well.

Despite these benefits and the recent improvements, the researchers are still struggling to offer substantial solutions for the improved reliability and real-time data delivery in IWSNs to match the strict deadlines as needed for the close loop process control and ultra-reliable and low-latency communications (URLLC) [17], [18]. URLLC is mandatory for IWSNs, especially when dealing with emergency communications and regulatory and supervisory control feedback systems. To improve the overall acceptability of IWSNs, and to cope with fast paced improvements in industry and protocol stack developments, restructuring and procedural changes for URLLC in IWSNs are very important. Furthermore, the inclusion of Machine-to-Machine (M2M) communications in 5G offers potentially benefitting framework, targeting three main aspects of IWSNs: 1) supporting for large number of low-rate network devices, 2) sustaining a minimal data rate in all circumstances to satisfy the feedback control requirements, and 3) enabling very low-latency data transfer [19]. Further architectural improvements and procedure restructuring are necessary for 5G M2M infrastructure to address such requirements in IWSNs.

In this paper, a multi-channel TDMA based hybrid scheme is proposed, which benefits from the use of multiple-channels and short frame communication for the communication of time critical data. The proposed scheme targets real-time data delivery with improved data reliability. The effectiveness of the scheme is demonstrated with a test case with two frequency channels: one is used for the slotted access of the medium of communication for low latency networks, and the other channel used for the communication of the urgently required or critically needed information to be delivered to the control center within a specified time deadline. The second channel is also used for the retransmission of the failed communications to improve reliability of the communication taking place in the network. The time deadline enforced by the control society is taken into consideration to offer a reliable solution for close loop control systems in industrial plants.

The performance of the proposed work is also presented in this paper in comparison with the IEEE802.15.4e LLDN, specifically defined for industrial automation. Since the proposed scheme considers typical IWSNs, the work can further be extended to certain specific 5G network application areas. Communications for ubiquitous machine type devices, Moving Networks (MNs) and Ultra Reliable Communication (URC) [20] are some of the examples. Besides, as demonstrated in the results, the proposed protocol is able to use multi-channel diversity to incorporate a large number of nodes. This ability enables it to be considered for applications in 5G ultra dense networks (UDNs) [21].

The rest of the paper is organized as follows. Section 2 presents literature review. The proposed system model is presented in Section 3. Section 4 discusses the results and presents performance analysis. Finally, Section 5 gives conclusions and future directions.

2 Literature Review

The IWSNs are now widely used in various industrial processes. Different industrial wireless protocols are also defined to facilitate certain industrial applications and to encourage the use of IWSNs in these applications. Most of the industrial protocols presently used in the industry are CSMA/CA based and the core functionalities of physical and media access control (MAC) layers are inherited from IEEE802.15.4. Zigbee and 6LoWPAN are examples of such protocols. Although the specified protocols offer flexibility of operation and can be used to establish ad-hoc on-demand networks with the ability of active network formation and handling runtime changes, these protocols are more suitable for monitoring traditional wireless sensor networks (WSNs) applications. Since the suitability of WSNs in time insensitive applications is widely accepted, this paper is focused on time sensitive and reliable IWSNs.

In 2012, the IEEE 802.15.4e standard was launched, which mainly targets the critical applications of WSNs and primarily focuses on industrial environments where time-sensitive and information-critical data are to be routed [15]. It uses TDMA based channel access to ensure collision free access to the

wireless resources on pre-specified and dedicated time slots. This standard also takes into consideration the low latency demands of the industrial processes and so a special framework, Low Latency Deterministic Network (LLDN) is introduced to meet the critical time deadlines for the emergency and close loop control applications. Some widely used industrial protocols (ISA100.11a, WirelessHART, etc.) also use TDMA based channel access.

Some recent researches have also targeted the priority based communications in IWSNs. In [22], the authors established priority levels based on the critical nature of information. The entire traffic in the network is divided in four levels where the highest priority nodes get instant channel access and its communication is facilitated by allocating the channel bandwidth of low priority nodes to it. In other words, based on the assigned priority level, a high priority node can hijack the low priority traffic bandwidth. The protocol offers an improved QoS for high priority nodes at the price of the low priority node's communication sacrifice. In [23], the authors present an arbitration based protocol where each node is assigned with a unique frequency. The frequency assignment is linked to the priority of the node and hence, on the basis of these preassigned frequencies a priority-wise schedule of transmission is created. The protocol executes in two phases, an arbitration decision period and an arbitration execution period. In the arbitration decision period, each node that wants to communicate broadcasts its preassigned frequency to determine a deterministic channel access order. This allows each node participated in the arbitration decision period to know how many time slots it has to wait before its transmission can take place, thus, the node assigned with arbitration frequency of the highest priority communicates instantly, while the node with the lowest priority arbitration frequency waits until all the other nodes have communicated. The scheme allows multilevel priority system, however, it requires a special coordinator to identify the received arbitration frequencies and respond to them accordingly. The frequencies are also preassigned on the basis of the pre-specified priority system which implements static priority system. Furthermore, with the increase in number of nodes, the problem in defining orthogonal frequencies also surfaces.

The multi-channel schemes in IWSNs for MAC optimization offer improved medium utilization. Many schemes are presented in literature which use multi-channels to offer improvements in the existing scenarios. In [24], the authors demonstrated the effectiveness of the multi-channel schemes in improving throughput over other schemes. In [25], the authors took into account the benefit of the availability of multiple channels, defined a scalable media access and considered the limitation of the presently available sensor nodes. However, this scheme requires frequent channel hopping and has relatively high scheduling overhead. In [26], the authors used TDMA based channel access in a multi-channel scenario. The scheme is relatively static and does not exploit the available resources. In [27], a pseudo random scheduling was introduced where each node randomly decides two factors, the wakeup time and the channel sequence. The primary aim of the protocol is to distribute the traffic in the available communication resources. However, the scheme fails to offer an efficient traffic scheduling and resource sharing mechanism. In [28], the authors used multiple channels to increase the network throughput. The proposed algorithm eliminates collisions by establishing coordinated transmissions. The scheme schedules both periodic and event based traffic using reinforcement learning to establish collision free transmissions on parallel data streams using multiple channels. However, the scheme fails to offer a differentiated treatment for different datasets of different priorities. It also fails to suggest a suitable alternate in case of communication failure. In [29], the authors proposed a multi-channel scheme where the network is divided into sub-trees. Once divided, each subtree is allocated a unique channel. In [30], the authors present a multi-channel scheme for the static networks. The scheme benefits from the TDMA based source aware scheduling. However, the scheme fails to give satisfactory assurance on reliability of the scheme. In [31], the multichannel overhead reduction was achieved using the regret matching based algorithm. For the evaluation of the proposed scheme, both software and hardware based analyses were presented. In [32], the authors proposed a hybrid scheme that uses both TDMA and CSMA/CA for communication purposes. The proposed work offers a mechanism to switch between the two access schemes based on the traffic density. The proposed protocol also considers multi-channel scenario. However, in this proposed scheme suitable reliability and QoS can only be achieved using much higher delays acceptable in critical industrial processes, making the scheme unsuitable for time critical and information sensitive industrial processes.

The discussed schemes though offer suitable improvements in the existing scenarios, almost all of the encountered schemes fail to offer suitable plans for retransmission of failed communications. In most of the cases, the importance of retransmission is ignored, which results in extended delay and failure in deterministic behavior of the network. The proposed scheme in this paper focuses on how to ensure the retransmission of failed communications up to certain desirable extent within the superframe and time deadline, which helps in improving both the overall delay and communication reliability.

3 Proposed System Model

Feedback control systems play a very important role in automation and process control. In such applications, the IWSNs serve as the feedback path for the sensory information. For better control of the processes, the reliability of the feedback link is very important and can be termed as an integral ingredient for smooth running of control processes. The deadline in the discrete feedback systems also alters the performance of the

implemented process control, hence, making the processes more time sensitive.

The proposed scheme uses TDMA based channel access to minimize interference and to ensure URLLC. Apart from this, the scheme also considers a multi-channel scenario where channels are effectively used to offer improved throughput, reliability and timely delivery. The proposed scheme focuses on short burst communications where a dedicated channel is used to facilitate the retransmissions and urgently required data. A detailed description of the network topology, superframe structure, channel specifications and system modelling is presented as follows. Before this, the frequently used parameters in the discussion are first listed in **Table 1**.

3.1 Superframe Structure

The proposed scheme targets improvement in MAC layer architecture by taking in account the availability of multiple channels. More specifically, the frequency and time division multiple access is utilized so as to improve the QoS and meet the time and reliability requirements of critical industrial processes. From the available frequency channels, a Special Purpose (SP) channel is specified which is dedicated for the short frame communications for highly time sensitive or erroneous packet communications. Any failures in transmissions from Regular Communication (RC) channels which require urgent retransmission due to sensitive nature of the data are facilitated by the SP channel. Superframe structure for the proposed system is presented in **Fig. 1**, where the superframe structure for RC and SP channels are presented. Except for the SP channel, all the channels use TDMA based access scheme for collision free communications. The SP channel, however, is implemented using CSMA/CA based as well as TDMA based access.

3.2 Network Architecture and Multi-Channel Scenario

In the proposed scenario, a star topology is considered, where the coordinator (cluster-head) considers the nodes' suitability for association and disassociation with a cluster (**Fig. 2**). Each node affiliated to a cluster is assigned a local id starting from 1 to w, where W is the maximum number of nodes in the cluster. All the TDMA based channels maintain a uniform superframe duration (T) in which the synchronization takes place at the start of every frame with a synchronization beacon transmitted by the cluster-head. In each superframe, the communication takes place in n time slots, each of duration t. These time slots are preassigned to the nodes in the cluster in sequence of the highest priority node to lowest priority node. Each time slot is further divided into data transmission and acknowledgement section. Note that in Fig. 2, the cluster represented is only with respect to a single frequency channel and there can exist more than n nodes in a cluster. The incorporation of multiple channels can be used to either improve the data rate of the individual nodes in a cluster or increase the number of affiliated nodes to the cluster head. Time slotted channel hopping (TSCH) is also considered for low latency and lossy networks by imposing a maximum limit on number of channels to C_u^s, where u is the total number of channels available and s is the number of channels selected for communications at certain time.

The SP channel is used to offer both contention based and TDMA based channel access, where the initially first k-slots are used for contention based access and the remaining $(n-k)$ slots are used for TDMA based communications. In the contention based time slots, the nodes in the cluster get the flexibility to transmit time sensitive information by using CSMA/CA. Since the communication failures from the same time frame (with possibly different channels) are used so at the start, the number of contenders trying to access the channel in first couple of time slots, using a CSMA/CA based channel access mechanism, is relatively low and hence most of the nodes get access to one of the first k-slots. The scheme also allows retransmission of failed communications in RC channels using the SP channel by implementing instant retransmission. The delayed synchronization beacon in SP channel serves to synchronize the upcoming TDMA based communications in this channel. Along with synchronization the transmission schedule of urgently required or failed communications in last k time slots is also broadcasted. In other words, the TDMA based time slots of SP channel are used for rescheduling failed communications in the other RC channels. In **Fig. 3**, a broadcasted schedule for retransmission of selected nodes' information is presented. The schedule is part of SYNP as represented in Fig. 2. The schedule consists of a sequence of 0's and 1's, where one bit is specified for each node. The position of the bit from left to right is assigned as per the nodes' id. In this sequence, the left most bit is for node 1, next for node 2, and so on until the rightmost bit specified for node w. The total 1's in the sequence cannot exceed $n-k$, the total number of TDMA based slots in the current superframe of SP channel.

In the proposed scheme a hierarchical architecture is used to offer suitable scalability features. The number of RC channels (H) are also limited to a maximum of 10 with one SP chan-

▼Table 1. Description of frequently used variables in the paper

Parameters	Variables
Total nodes	W
High priority nodes	K
Time slots in a superframe	N
Total RC channels	H
Total high priority nodes communicating in a single superframe	w
Frequency bands for RC channels	$f_1, f_2, \ldots, f_H$
TDMA based time slots in SP channel	$n-k$
Probability of successful communication of a node	P

RC: Regular Communication SP: Special Purpose TDMA: time-division multiple access

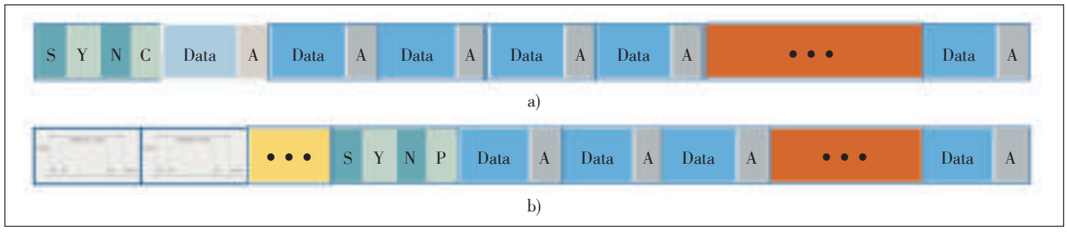


Figure 1. Superframe structure: a) TDMA based (RC) channels and b) SP channel .

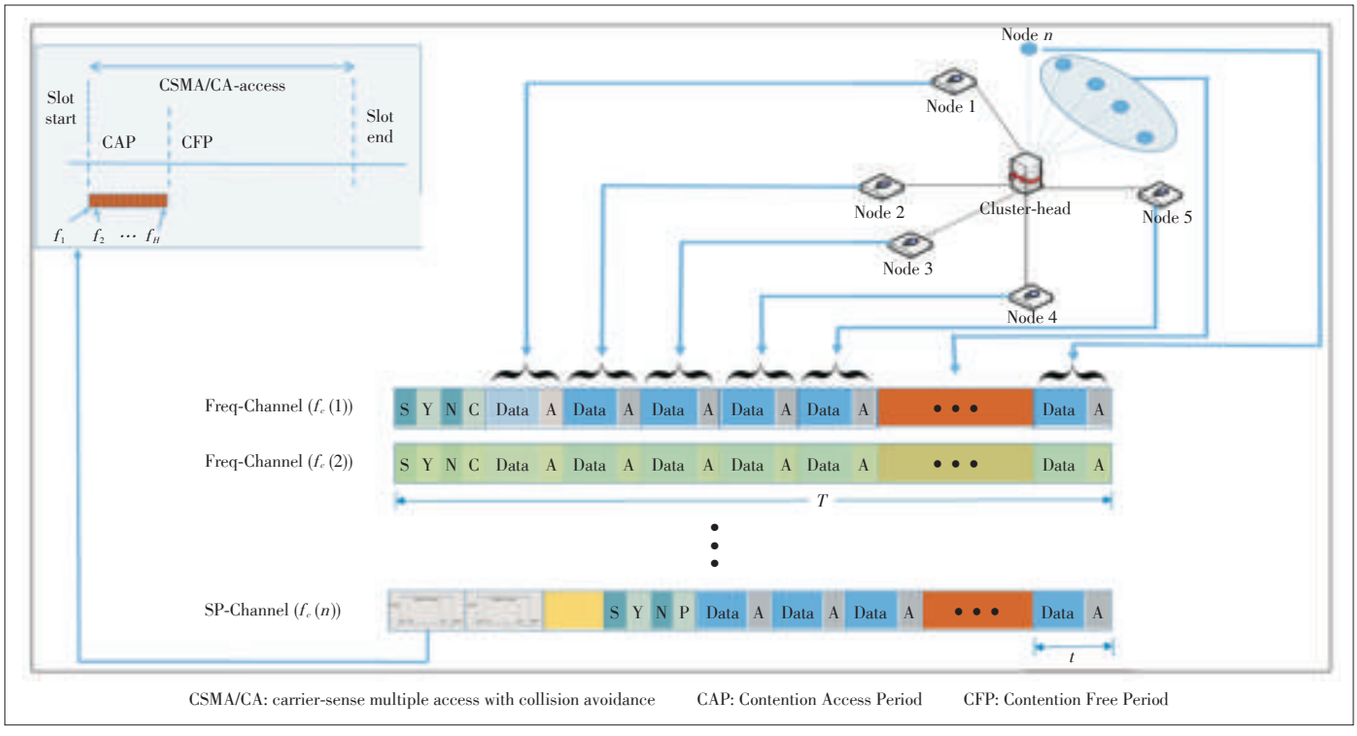


Figure 2. Superframe Structure, channel distribution and cluster representation.

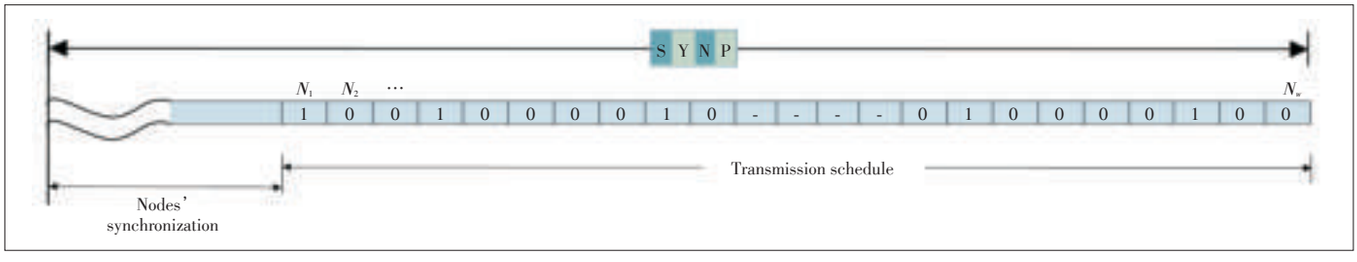
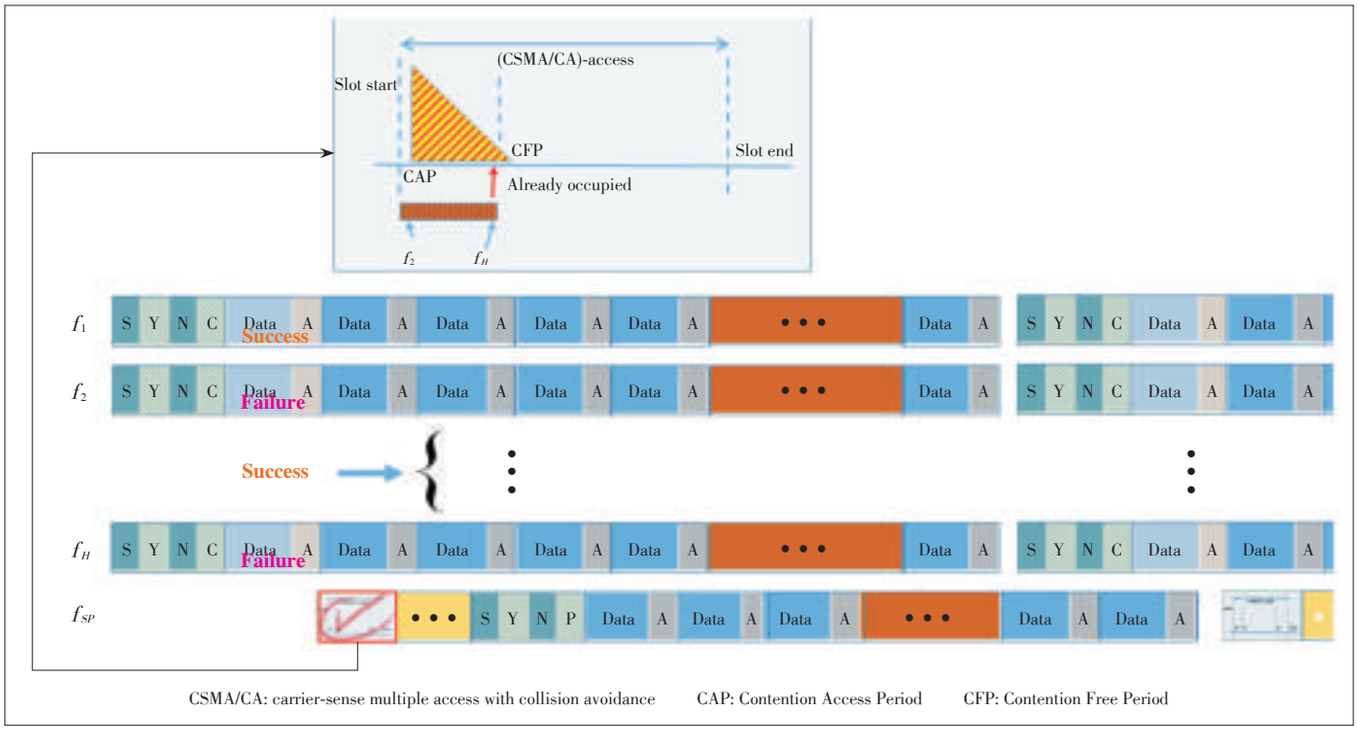


Figure 3. Schedule of the nodes transmitting in TDMA slots of SP channel.

nel. Each superframe in RC channels is divided in 20 time slots and the time duration of the superframe is limited to 10 milliseconds (ms) to establish low latency and deterministic networks. Apart from this, the communication from all the RC channels and the SP channel are aligned as represented in **Fig. 4** and all superframes are of the same duration. Note that the start of the superframe in case of SP channel is shifted by exactly one time slot (t duration) after the beacon of all the RC channels. In this way every node in the cluster is synchronized with SP channel. For the evaluation purposes, the scope of this paper is limited to the cases where multiple channels are used for the increase in the number of nodes affiliated to single cluster head. Other multi - channel diversity improvement techniques including throughput enhancement, data replication, etc. and relevant investigations are not presented in this paper.

The superframes at different frequency channels are synchronized in a manner represented in Fig. 4, which allows the allocation of first three time slots in superframes of all RC channels to highest priority nodes in the cluster. Furthermore, the priority of these nodes is also distinguished by affiliating a



▲**Figure 4. Synchronized superframe structure for NC and SP channel and priority based access.**

priority factor based on frequency of the channels. To further elaborate, please consider the scenario where, after the beacon synchronization, the communication of first time slot in superframe of all RC channels takes place. In this time slot, all the nodes to whom this time slot is allocated try to communicate. Out of these communications, one or more communications can fail. As represented in Fig. 4, the nodes with unsuccessful communications (communicating at f_2 and f_H) will try to access the CSMA/CA based time slot of the SP channel (marked with $\surd$) and try to get hold of it during Contention Access Period (CAP). A magnified view of this slot is also presented in the figure, where based on frequency, the access to the slot is divided. Each node based on its channel frequency will sense the channel first and if vacant will initiate its access beacon giving a signal for the rest of the nodes that the Contention Free Period (CFP) of this slot is reserved for its communication. In the presented case, the node operating at frequency channel f_2 will sense the channel and finds no other access beacons, so its beacon is broadcasted. The node communicating at frequency f_H, due to higher frequency is allowed to access the channel later in CAP and finds beacon of node operating at f_2 and hence withdraws its access till the next time slot. Any node, which fails to access CSMA/CA based communication slots for retransmission of its information, is scheduled for transmission by the coordinator and its transmission schedule is included in the SYNP for retransmission on TDMA based time slots in SP channel. Doing so improves the communication reliability and timely delivery of information to the coordinator. Since this scheme tries to improve the reliability and real time data delivery of high priority nodes, the total number of nodes benefitting from this scheme are limited to w, where $w = k \times H$.

3.3 Mathematical Formulation

The proposed scheme considers the impact of multiple channels compared to single channel schemes and evaluates improvements in number of nodes per cluster, communication reliability and overall throughput. To evaluate the performance of the proposed scheme, a mathematical formulation of the possible scenarios for typical IWSNs as well as for the proposed scheme in IWSNs is presented as follows.

The communication from all the affiliated nodes in a cluster periodically originates in a specified time slot of every superframe. The success of each individual communication is dependent on the channel conditions and probability of success of an individual communication, which is represented with p whereas the total successes in every time slot are modelled as binomial (p, w) distribution.

In a typical IEEE802.15.4e system, using a single frequency channel, the frame error rate depends on several factors including the number of high priority nodes, multipath fading, dispersion, reflection, refraction, jitter, interference, distance, congestion, transmission power restrictions, and receiver sensitivity. To demonstrate the frame error rate for various number of high priority nodes, a mathematical formulation is presented in (1), where k is the number of high priority nodes and changes from 1 to 10. The total communications in a single frame are limited

to n.

$$P(frame\_error\_rate|H=1)=\sum_{x=1}^{k}\binom{k}{x}(1-p)^{x}p^{k-x}. \quad (1)$$

The use of multiple channels introduces a significant improvement in the throughput of the system but the reliability in such cases is dependent on the number of RC channels and ratio of RC and SP channels. A mathematical expression for the frame error rate in case of multiple channel scenarios is given in (2) to (4). The maximum number of high priority nodes are limited to k per RC channel whose maximum numbers are limited to H. Based on the total number of the high priority nodes permitted (k) and total communications in a single frame (n), limits can be generalized to w_1 and w_2 for (2) to (4), where $w=k\times H$, $w_1=n/2$ and $w_2=n$.

$$P\left(frame\_error\_rate\ |(H>1)\&(w\le w_1)\right)=\frac{\left[\left(\sum_{x=1}^{w}\binom{w}{x}(1-p)^{x}p^{w-x}\times\left(\sum_{y=1}^{x}\binom{x}{y}(1-p)^{y}p^{x-y}\right)^{2}\right)\right]}{H}, \quad (2)$$

$$P(frame\_error\_rate\ |(H>1)\&(w_1<w\le w_2))=\left[\left(\sum_{x=1}^{w_1}\binom{w}{x}(1-p)^{x}p^{w-x}\times\left(\sum_{y=1}^{x}\binom{x}{y}(1-p)^{y}p^{x-y}\right)^{2}\right)+\left\{\sum_{x=w_1+1}^{w}\binom{w}{x}(1-p)^{x}p^{w-x}\times\left\{\left(\sum_{y=1}^{x+w_1-w}\binom{x}{y}(1-p)^{y}p^{x-y}\right)^{2}+\left(\sum_{y=x+w_1-w+1}^{x}\binom{x}{y}(1-p)^{y}p^{x-y}\right)\right\}\right\}\right]\Big/H, \quad (3)$$

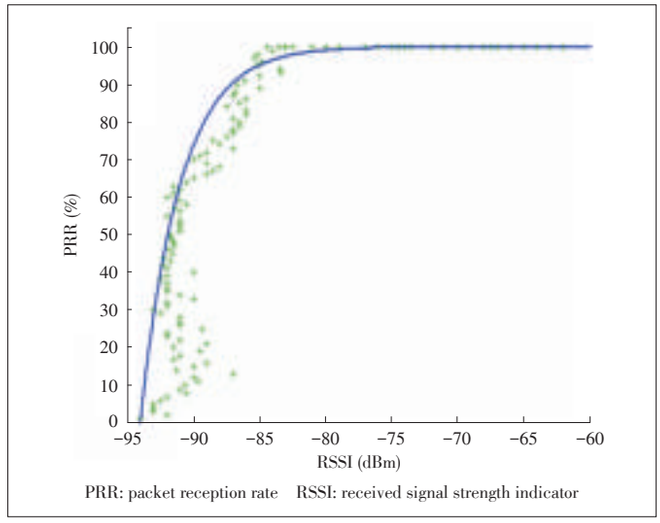
$$P(frame\_error\_rate\ |(H>1)\&(w>w_2))=\left[\left(\sum_{x=1}^{w_1}\binom{w}{x}(1-p)^{x}p^{w-x}\times\left(\sum_{y=1}^{x}\binom{x}{y}(1-p)^{y}p^{x-y}\right)^{2}\right)+\left\{\sum_{x=w_1+1}^{w_2}\binom{w}{x}(1-p)^{x}p^{w-x}\times\left\{\left(\sum_{y=1}^{x+w_1-w}\binom{x}{y}(1-p)^{y}p^{x-y}\right)^{2}+\left(\sum_{y=x+w_1-w+1}^{x}\binom{x}{y}(1-p)^{y}p^{x-y}\right)\right\}\right\}+\left\{\sum_{x=w_2+1}^{w}\binom{w}{x}(1-p)^{x}p^{w-x}\right\}\right]\Big/H. \quad (4)$$

4 Results and Discussion

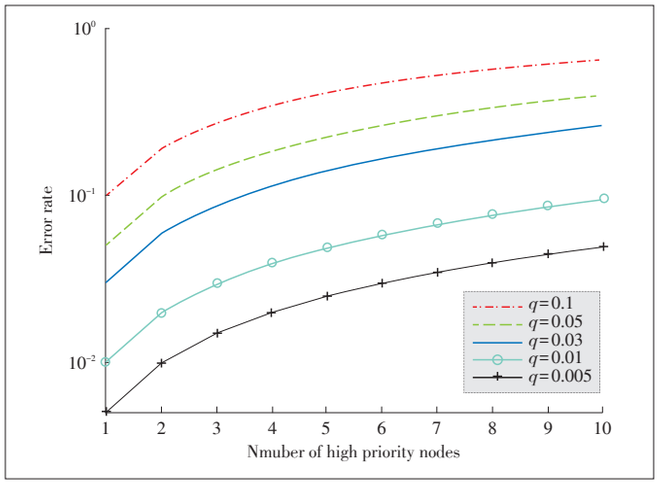
The performance of the proposed multiple channel scheme with a dedicated retransmission channel (SP channel) is evaluated as a function of probability of successful communication of an individual node, total number of parallel data streams, i.e. the number of communication channels (RC channels) and the number of high priority nodes (k) trying to communicate in a single superframe duration. For the evaluation purposes, the maximum number of RC channels (H) and k are limited to 10. Twenty transmissions in a single superframe are used (n = 20), so w_1 and w_2 are set to 10 and 20 respectively.

In **Fig. 5**, the probability of successful communication of a node is presented as a function of received signal strength indicator (RSSI), and the packet reception rate (PRR) is plotted against the RSSI. The plot is acquired using a communication established between the SunSPOT sensor node and SunSPOT base-station. The nodes use CC2420 radio, which operates at 2.4 GHz and uses offset quadrature phase-shift keying (OQPSK) modulation with a chip rate of 2 Mchips/s. The plot in Fig. 5 represents the percentage of successfully received packets for different values of RSSI, with blue line representing polynomial curve fitting of scatter plot. As can be seen in the figure, if the received RSSI is maintained above −87 dBm, 90% or more successful transmissions are expected. To counter the effect of uncertainty of the wireless channel, 10 dBm margin is suggested when establishing a link between the coordinator/cluster-head and sensor nodes. For communication, a superframe duration of 10 ms is used. The maximum number of parallel data streams is limited to 10 and synchronized in time domain. To evaluate the performance of the proposed scheme, the performance of typical IEEE802.15.4e system with single a channel is presented as a reference in **Fig. 6**. The performance is evaluated, based on the number of high priority nodes (k) communicating within the superframe where the total number of nodes trying to attempt a communication in a single superframe is limited to 20. The frame error rate is evaluated for different channel conditions under which the probability of communication failure (1-p) is represented by q.

By introducing a SP channel in the typical IEEE 802.15.4e system as expressed in the proposed scheme, a significant error rate reduction can be seen in the communication of high priority nodes. Since the rest of the n-k nodes communicating



▲**Figure 5. Packet reception rate as a function of RSSI using CC2420 [33].**

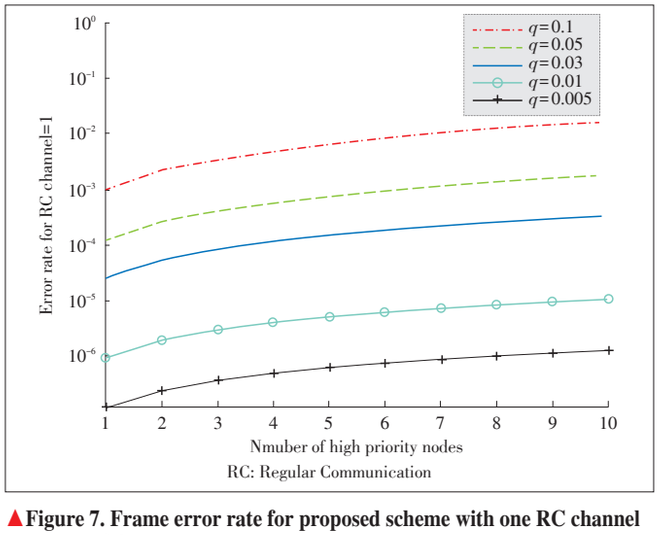


▲Figure 6. Frame error rate for typical IEEE802.15.4e (LLDN) with one RC channel and no SP channel.

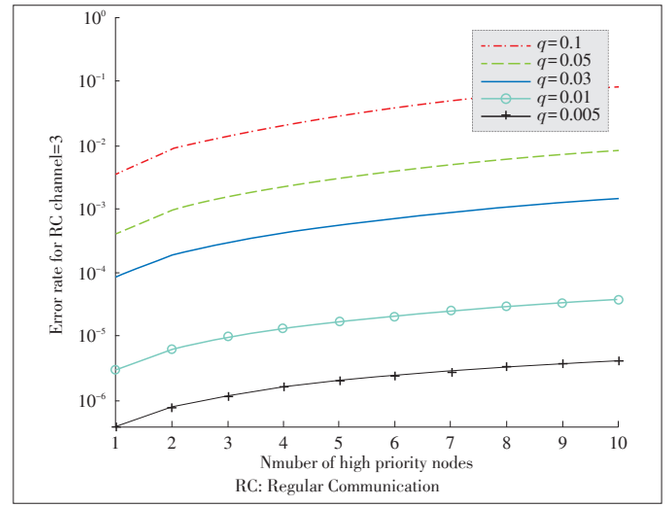
in the network are considered as low priority nodes, so the failure in communication of these nodes is not critical and will not affect the performance of feedback control systems. With the retransmission of failed communication in RC channel through SP channel, a significant improvement in the reliability of communication can be seen. Similar conclusion can be deducted by comparing the error - rate of IEEE802.15.4e presented in Fig. 6, and proposed multi-channel scheme with one RC channel and one SP channel, presented in **Fig. 7**. To further evaluate the effect of using multi-channel scheme for a higher number of parallel data streams, the error rate is evaluated for the cases with one SP channel and 3, 5 and 10 RC channels, as presented in **Figs 8**, **9** and **10** respectively.

Due to the introduction of multi-channel scheme, the overall throughput of the network greatly increases along with the potential rise in the total number of nodes which can affiliate to a single cluster. In a scenario where the single channel scheme is compared with the two-channel scheme (one SP and one RC channel), the throughput is relatively the same as in the case the single channel scheme because only one channel is used for communication. However, a notable improvement in the communication reliability can be seen. The use of more than one RC channels, however, strongly influences the overall throughput and with the increase in these communication channels the throughput is increased several times. The overall throughput for different number of frequency channels is represented in **Fig. 11**. As represented in this figure the overall throughput of the network can increase up to 900 percent with additional ten frequency channels in use. Apart from the throughput, as discussed earlier the reliability of the communication also improves along with the throughput.

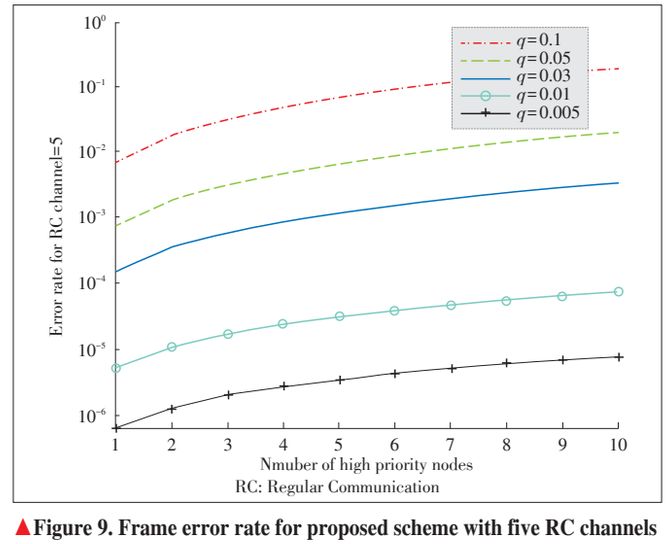
The total number of nodes that can communicate to the cluster-head in time T with different number of frequency channels



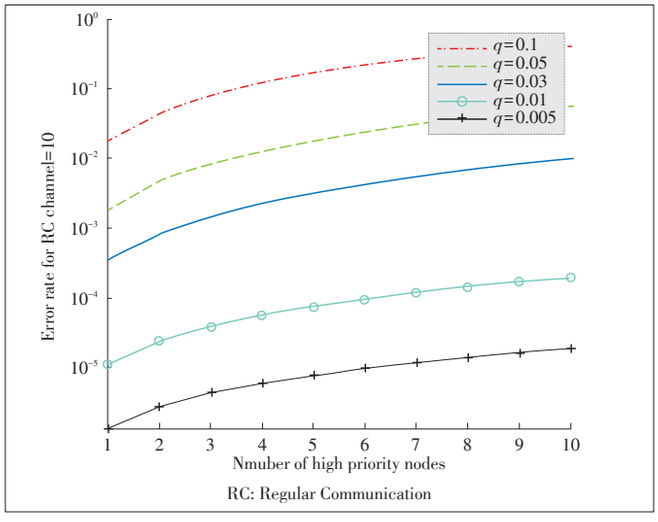
▲Figure 7. Frame error rate for proposed scheme with one RC channel and one SP channel.



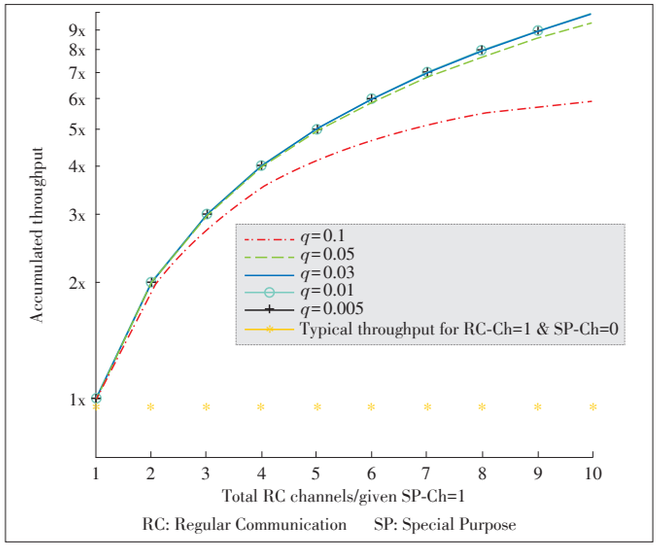
▲Figure 8. Frame error rate for proposed scheme with three RC channels and one SP channel.



▲Figure 9. Frame error rate for proposed scheme with five RC channels and one SP channel.



▲Figure 10. Frame error rate for proposed scheme with ten RC channels and one SP channel.

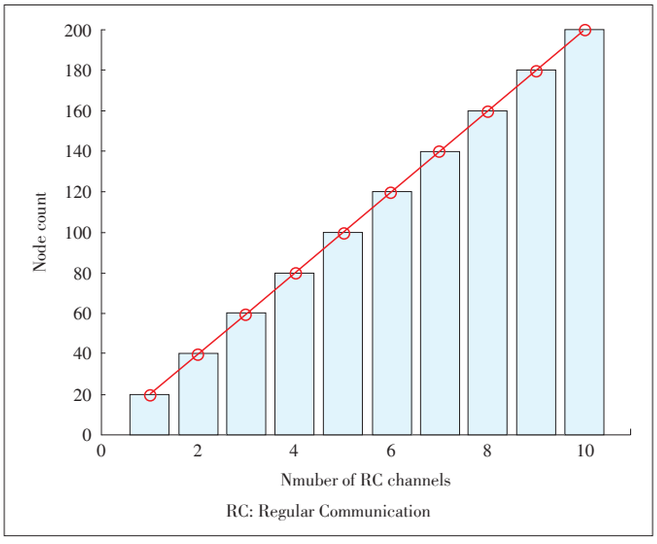


▲Figure 11. Accumulated throughput of the proposed multi-channel scheme with reference to single channel low data-rate WPAN.

is presented in **Fig. 12**. Since the short burst communication is used for urgent delivery of information, the time duration of superframe T is limited to 10 ms. A communication overhead of 3.84 ms is considered, leaving the connected nodes a relatively short duration for communication. Each node in a cluster is assigned at least one time-slot in the 10 ms window enabling fast, frequent, and reliable communication to the central control unit, well within the specified times.

5 Conclusions and Future Directions

The paper presents a multi-channel performance and throughput enhancement scheme for IWSNs. The primary objective of the scheme is to enhance the reliability of the communication between the sensor nodes and the cluster-head. A SP channel is also defined which ensures suitable reliability enhancement in the communication over the traditional single channel schemes. In the scheme, the performance is evaluated using throughput, reliability and the number of nodes accommodated in a cluster. The scheme offers a notable increase in the reliability and throughput over the existing IEEE802.15.4e standard. The overall improvement in reliability is directly dependent on the SP channels to RC channels ratio. The throughput however is more dependent on the number of RC channels and the probability of successful communication.



▲Figure 12. Total number of affiliated nodes to a cluster-head.

For the evaluation purposes, the scheme considers one SP channel and can further be realized for multiple SP channels and give a more suitable venue for performance improvement. In this paper the SP/RC channels ratio is limited to the cases ranging from 1:1 to 1:10. However, a more generic approach may be considered, where ratios 1:1 to 10:1 are evaluated for study purposes. Investigating the whole range of channel ratios enables the use of the scheme in different challenging scenarios and a predefined projected or predicted output can also be formulated to better meet the desired requirements.

References

[1] Y.-C. Chou, C.-H. Lu, and P.-L. Chang, "Using theory of constraints to find the problem about high level inventory in the aerospace industry," in *Technology Management for Global Economic Growth (PICMET)*, Phuket, Thailand, Oct. 2010, pp. 1–10.
[2] A. Z. Xu. (2014). *2020: future automation* [Online]. Available: http://www.controleng.com/single-article/2020-future-automation/c33ed4679973dcb1ed2f534115200 88d.html
[3] T. M. Chiwewe and G. P. Hancke, "A distributed topology control technique for low interference and energy efficiency in wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 11–19, 2012. doi: 10.1109/TII.2011.2166778.
[4] M. Magno, D. Boyle, D. Brunelli, et al., "Extended wireless monitoring through intelligent hybrid energy supply," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 4, pp. 1871–1881, Apr. 2014. doi: 10.1109/TIE.2013.2267694.
[5] L. Palopoli, R. Passerone, and T. Rizano, "Scalable offline optimization of indus-

trial wireless sensor networks, " *IEEE Transactions on Industrial Informatics*, vol. 7, no.2, pp. 328–339, May 2011. doi: 10.1109/TII.2011.2123904.
[6] S. Shin, T. Kwon, G.-Y. Jo, Y. Park, and H. Rhy, "An experimental study of hierarchical intrusion detection for wireless industrial sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 6, no. 4, pp. 744–757, Nov. 2010. doi: 10.1109/TII.2010.2051556.
[7] V. C. Gungor and G. P. Hancke, "Industrial wireless sensor networks: challenges, design principles, and technical approaches, " *IEEE Transactions on Industrial Electronics*, vol. 56, no. 10, pp. 4258–4265, Oct. 2009. doi: 10.1109/TIE.2009.2015754.
[8] B. C. Villaverde, S. Rea, and D. Pesch, "InRout—a QoS aware route selection algorithm for industrial wireless sensor networks," *Ad Hoc Networks*, vol. 10, no. 3, pp. 458–478, May 2012. doi: 10.1016/j.adhoc.2011.07.015.
[9] Tang, Z. Mei, P. Zeng, and H. Wang, "Industrial wireless communication protocol WIA-PA and its interoperation with Foundation Fieldbus," in *International Conference on Computer Design and Applications (ICCDA)*, Qinhuangdao, China, 2010, pp. 370–374. doi: 10.1109/ICCDA.2010.5541074.
[10] K. A. Agha, M.-H. Bertin, T. Dang, et al., "Which wireless technology for industrial wireless sensor networks? The development of OCARI technology," *IEEE Transactions on Industrial Electronics*, vol. 56, no. 10, pp. 4266–4278, Oct. 2009. doi: 10.1109/TIE.2009.2027253.
[11] H. C. Foundation. (2016). *WirelessHART overview* [Online]. Available: http://en.hartcomm.org/hcp/tech/wihart/wireless_overview.html
[12] IETF datatracker 6loWPAN. (2016). *6loWPAN active drafts* [Online]. Available: https://datatracker.ietf.org/doc/search/?name=6loWPAN&activeDrafts=on&rfcs=on
[13] The ZigBee Alliance. (2016). *Utility industry* [Online]. Available: http://www.zigbee.org/what-is-zigbee/utility-industry
[14] ISA-100 Wireless Compliance Institute (2016). *ISA-100 Wireless compliance institute—official site of ISA100 wireless standard* [Online]. Available: http://www.isa100wci.org
[15] *IEEE Standard for Local and Metropolitan Area Networks—Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 1: MAC Sublayer*, IEEE Std 802.15.4e-2012 (Amendment to IEEE Std 802.15.4-2011), 2012.
[16] *IEEE Standard for Information Technology—Local and Metropolitan Area Networks—Specific Requirements, Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low Rate Wireless Personal Area Networks (WPANs)*, IEEE Std 802.15.4 - 2006 (Revision of IEEE Std 802.15.4-2003), 2006.
[17] P. Suriyachai, U. Roedig, and A. Scott, "A survey of MAC protocols for mission-critical applications in wireless sensor networks," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 2, pp. 240–264, 2012. doi: 10.1109/SURV.2011.020211.00036.
[18] P. T. A. Quang and D.-S. Kim, "Enhancing real-time delivery of gradient routing for industrial wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 61–68, Feb. 2012. doi: 10.1109/TII.2011.2174249.
[19] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, Feb. 2014. doi: 10.1109/MCOM.2014.6736746.
[20] M. Fallgren and B. Timus, "Scenarios, requirements and KPIs for 5G mobile and wireless system," METIS deliverable D, vol. 1, p. 1, 2013.
[21] A. Osseiran, F. Boccardi, V. Braun, et al., "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, 2014. doi: 10.1109/MCOM.2014.6815890.
[22] S. Wei, Z. Tingting, F. Barac, and M. Gidlund, "PriorityMAC: a priority-enhanced MAC protocol for critical traffic in industrial wireless sensor and actuator networks," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 824–835, 2014. doi: 10.1109/TII.2013.2280081.
[23] T. Zheng, M. Gidlund, and J. Akerberg, "WirArb: a new MAC protocol for time critical industrial wireless sensor network applications," *IEEE Sensors Journal*, vol. 16, no. 7, pp. 2127–2139, 2016. doi: 10.1109/JSEN.2015.2504948.
[24] O. D. Incel, A. Ghosh, B. Krishnamachari, and K. Chintalapudi, "Fast data collection in tree-based wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 11, no. 1, pp. 86–99, 2012. doi: 10.1109/TMC.2011.22.
[25] G. Zhou, C. Huang, T. Yan, et al., "MMSN: multi-frequency media access control for wireless sensor networks," in *Proc. 25th IEEE International Conference on Computer Communications (INFOCOM)*, Barcelona, Spain, 2006, pp. 1–13. doi: 10.1109/INFOCOM.2006.250.
[26] O. D. Incel, L. Van Hoesel, P. Jansen, and P. Havinga, "MC-LMAC: a multi-channel MAC protocol for wireless sensor networks," *Ad Hoc Networks*, vol. 9, no. 1, pp. 73–94, 2011. doi: 10.1016/j.adhoc.2010.05.003.
[27] L. Tang, Y. Sun, O. Gurewitz, and D. B. Johnson, "EM-MAC: a dynamic multi-channel energy-efficient mac protocol for wireless sensor networks," in *Proc. 12th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Paris, France, 2011, no. 23. doi: 10.1145/2107502.2107533.
[28] K.-H. Phung, B. Lemmens, M. Goossens, et al., "Schedule-based multi-channel communication in wireless sensor networks: a complete design and performance evaluation," *Ad Hoc Networks*, vol. 26, no. pp. 88–102, 2015. doi:10.1016/j.adhoc.2014.11.008.
[29] Y. Wu, J. A. Stankovic, T. He, and S. Lin, "Realistic and efficient multi-channel communications in wireless sensor networks," in *Proc. 27th IEEE International Conference on Computer Communications (INFOCOM)*, Phoenix, USA, 2008. doi: 10.1109/INFOCOM.2008.175.
[30] J. Zhao, Y. Qin, D. Yang, and Y. Rao, "A source aware scheduling algorithm for time-optimal convergecast," *International Journal of Distributed Sensor Networks*, vol. 2014, article 251218, 2014. doi: 10.1155/2014/251218.
[31] C. Jiming, Y. Qing, C. Bo, et al., "Dynamic channel assignment for wireless sensor networks: a regret matching based approach," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 1, pp. 95–106, 2015. doi: 10.1109/TPDS.2014.2307868.
[32] S. Zhuo, Z. Wang, Y. Q. Song, Z. Wang, and L. Almeida, "A traffic adaptive multi-channel MAC protocol with dynamic slot allocation for WSNs," *IEEE Transactions on Mobile Computing*, vol. 15, no. 7, pp. 1600–1613, 2016.
[33] M. Raza, G. Ahmed, and N. M. Khan, "Experimental evaluation of transmission power control strategies in wireless sensor networks," in *2012 International Conference on Emerging Technologies (ICET)*, Islamabad, Pakistan, 2012, pp. 1–4. doi: 10.1109/ICET.2012.6375498.

Manuscript received: 2016-11-30

Biographies

Mohsin Raza (mohsinraza119@gmail.com) received his B.S. and M.S. degrees in electronic engineering from Mohammad Ali Jinnah University, Pakistan. Currently he is a Ph.D. student in math, physics and electrical engineering at Northumbria University, UK. Prior to this he worked as a lecturer in Department of Electronic Engineering, Mohammad Ali Jinnah University from 2010 to 2015 and prior to that as a hardware support engineer at USS in 2009 and 2010. His research interests include wireless sensor networks, mobile communications, smart grids and autonomous transportation & vehicular networks.

Sajjad Hussain (sajjad.hussain@glasgow.ac.uk) received his M.S. degree from SUPELEC, France and Ph.D. degree from University of Rennes 1, France, both in wireless communication and signal processing. He is currently a lecturer at University of Glasgow, UK. Prior to joining University of Glasgow, he was an associate professor at Capital University of Science and Technology, Pakistan and prior to that, an assistant professor at National university of Science and Technology, Pakistan. His main research interests include spectrum sensing, security, and cross layer optimization in cognitive radios and wireless networks.

Hoa Le-Minh (hoa.le-minh@northumbria.ac.uk) received his B.E. degree in telecommunications from Ho Chi Minh University of Technology, Vietnam in 1999, M. S. in communications engineering from Munich University of Technology, Germany in 2003, and Ph.D. in optical communications from Northumbria University, UK in 2007. Prior to joining Northumbria University as a senior lecturer in 2010 and subsequently the programme leader of BEng (Hons) and MEng Electrical and Electronic Engineering (2013), he was a research fellow of the Department of Engineering Science and a tutor of St Edmund Hall College, University of Oxford, UK (2007–2010). He also worked at R&D Siemens AG, Munich, Germany (2002–2004). His research interests include ad-hoc and wireless networks, visible light communication and free space optics.

Nauman Aslam (nauman.aslam@northumbria.ac.uk) joined Northumbria University, UK in August 2011 and is a senior lecturer in Department of Computer Science and Digital Technologies there. Dr. Aslam received his Ph.D. in engineering mathematics from Dalhousie University, Canada in 2008. He was awarded M.E. in internetworking from Dalhousie University in 2003 and B.S. in mechanical engineering from University of Engineering and Technology, Pakistan in 1993. Prior to joining Northumbria University, he worked as an assistant professor at Dalhousie University from 2008 to 2011. Currently, he also holds an adjunct assistant professor position at Dalhousie University. His research interests include wireless ad-hoc and computer networks, process optimization and artificial intelligence.

DOI: 10.3969/j. issn. 1673-5188. 2017. S1. 007
http://kns.cnki.net/kcms/detail/34.1294.TN.20170619.1441.002.html, published online June 19, 2017

Device-to-Device Based Cooperative Relaying for 5G Network: A Comparative Review

JIANG Wei[1,2]
(1. Intelligent Networking Group, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern 67663, Germany;
2. Department of Electrical and Computer Engineering (EIT), University of Kaiserslautern, Kaiserslautern 67663, Germany)

Abstract

Due to the proliferation of mobile internet access, the cellular traffic is envisaged to experience a 1000-fold growth in the second decade of the 21st century. To meet such a huge traffic demand, the Fifth Generation (5G) network have to adopt new techniques to substantially increase spectral efficiency and reliability. At the base station side, available resources (power supply, equipment size, processing capability, etc.) are far more sufficient than that of the terminal side, which imposes a high challenge on the uplink transmission. The concept of cooperative communications opens a possibility of using multiple terminals to cooperatively achieve spatial diversity that is typically obtained by means of multiple antennas in the base station. The application of Device-to-Device (D2D) communications in the 3GPP LTE system further pushes the collaboration of terminals from the theory to the practice. The utilization of D2D-based cooperative relaying is promising in the era of 5G. In this paper, we comparatively study several cooperative multi-relay schemes, including the proposed opportunistic space-time coding, in the presence of imperfect channel state information. The numerical results reveal that the proposed scheme is the best cooperative solution until now from the perspective of multiplexing-diversity tradeoff.

Keywords

cooperative communications; Device-to-Device (D2D); Distributed Space-Time Coding (DSTC); outdated chanel state information; opportunistic relaying

1 Introduction

Due to the proliferation of mobile internet access, the cellular traffic experienced an exponential growth in the past years, which imposes a high pressure on the current networks and urges the industry to develop a new generation wireless system [1]. A consensus for the 5G system is that mobile traffic will increase 1000 times in the second decade of the 21th century [2], [3]. In addition to other cutting-edge paradigms, protocols and architectures (Licensed-Assisted Access [4], Device-to-Device (D2D) communications [5], software-defined networking [6], network function virtualization [7], etc.), the 5G system has to adopt New Radio (NR) transmission techniques to substantially increase spectral efficiency and reliability so as to meet such a huge capacity demand. However, a multi-path channel suffers from a severe fading due to constructive and destructive interferences of received signals in wireless communications [8]. At a high data transmission rate, it is challenging for a receiver to correctly detect a signal without some form of diversity. Since the time and frequency resources in a wireless system are tightly limited, the exploitation of spatial resource is of great importance. A particularly appealing approach is the utilization of antenna array, such as Multiple-Input Multiple-Output (MIMO) [9] and massive MIMO [10], which can achieve higher diversity by means of simply installing additional antennas.

Until now, the technical discussions on the application of antenna array in a wireless system are mainly focused on the base station side. Due to the limitations of power supply, cost and hardware size, a mobile terminal is traditionally hard to be equipped with multiple antennas. Recently, the millimeter wave (mmWave) band [11] has been intensively investigated and the size of antenna at this band is small enough to be massively integrated in a mobile terminal. However, due to the severe propagation characteristics of radio signals at the higher frequencies, an mmWave-based system cannot independently form a wide-area-covered network in a cellular manner. Complementary to macro-cell coverage, it suits to provide ultra-high data access at hot spots [12]. Taking into account the requirement of ubiquitous signal coverage, especially the provision of control signaling and system information broadcasting [13], the carrier frequencies below 6 GHz is still the mainstream for macro-cell transmission and plays a vital role from the perspective of a holistic wireless system. In a nutshell, the working assumption of a signal antenna at the mobile terminal for macro-cell transmission is practically meaningful.

With a single antenna, it is infeasible to exploit multi-antenna diversity for mobile terminals in a cellular system. In this context, the concept of cooperative communications [14] has been proposed to solve this problem by means of making full use of the broadcast nature of wireless signals in a relay chan-

nel [15], where multiple single-antenna terminals can form a virtual antenna array to collaboratively transmit their signals. Once a terminal sends a signal to its destination (a base station), its neighboring terminals that overhear this signal are capable of decoding and retransmitting. By means of combining multiple copied versions of the original signal at the receiver, an inherent spatial diversity referred to as cooperative diversity [16] can be achieved.

Currently, it is still impossible to commercially implement a full-duplex [17] mobile terminal to simultaneously transmit and receive signals at the same frequency. A terminal has to operate in a half-duplex mode where Time- or Frequency-Division Multiplexing (TDD/FDD) is applied. Without loss of generality, we are allowed to use TDD as an example to analyze the cooperative relaying. Basically, an end-to-end signal transmission happens in two phases [18]. In the broadcast phase, a terminal (the source) transmits its signal in the source-relay channels while all neighboring terminals listen. Those neighboring terminals who have overheard and successfully decoded this signal can act as the relays. In the relaying phase, all or a subset of the relays retransmit this signal in the relay-destination channels. However, a scheduling problem occurs in the scenario of multi-relay cooperative transmission. That is which relays should be selected and how the regenerated signals should be transmitted by the selected relays. In the literature, several cooperative multi-relay schemes have been proposed. Generalized Selection Combining (GSC) [19] choosing multiple relays to orthogonally retransmit suffers from a substantial loss of spectral efficiency. To avoid this penalty, the distributed Beam-Forming (BF) [20] based on simultaneous transmission has been taken into account. Given a perfect channel knowledge, the relays adjust the phases of their transmit signals for coherently combining at the receiver. As we know, BF is very sensitive to phase noise [21]. That is why co-located antennas in an MIMO system must apply an antenna calibration scheme to align phase distortions on different radio-frequency (RF) chains. However, the RF-chain calibration schemes designed for co-located antennas cannot be applied among spatially-distributed terminals. In practice, the performance degradation from the phase distortions overwhelms the expected BF gain. In [22], an approach called Distributed Space-Time Coding (DSTC) has been proposed to transmit space-time-coded signals by multiple relays. Although a full diversity can be achieved, designing such a code is infeasible since the number of distributed antennas is unknown and randomly varying. Moreover, the synchronization among simultaneously transmitting relays becomes challenging when the number of relays is large. In a nutshell, the aforementioned multi-relay transmission methods are hard to be applied for practical systems.

In [23], Bletsas et al. revealed that the multi-relay synchronization problem can be avoided while keeping full cooperative diversity by opportunistically selecting the best relay to retransmit. They proposed an approach referred to as Opportunistic Relaying System (ORS), which has been extensively verified as a simple but efficient cooperative relaying scheme. Although only a single terminal with the best channel (in accordance to a given selection criterion) is selected to serve as a relay, a full spatial diversity with the number of all cooperative relays can be available. Its achieved performance is about the same as that of the DSTC scheme, which uses an all-participating strategy [24]. From the perspective of multiplexing-diversity tradeoff, the ORS scheme provides no performance loss in comparison with the DSTC scheme, while avoiding the complicated implementation.

From the practical point of view, the channel state information (CSI) at the time instant of relay selection may substantially differ from the CSI at the instant of using the selected relay to retransmit owing to the channel fading and feedback delay. The imperfect CSI imposes a possibility of wrongly selecting the best relay in the ORS scheme, which drastically deteriorates its performance. The impact of the outdated CSI on the performance of opportunistic relaying has been extensively analyzed in the previous works. In [25], a closed-form expression of outage probability for Decode-and-Forward (DF) ORS has been derived. Seyfi et al. [26] investigated the impact of feedback delay and channel estimation error on the relaying selection. Kim et al. evaluated the performance degradation in terms of symbol error probabilities in [27]. Regarding Amplify-and-Forward (AF) ORS, Torabi et al. presented a lot of results through [28]–[30]. The impact of the outdated CSI on partial relay selection has also been reported in [31] and [32]. The error probabilities of ORS considering channel estimation errors have been derived in [33]. Based on the outcomes presented in the literature, the following conclusions can be drawn:

1) The relay selection is very vulnerable to the imperfect channel quality, where its achieved diversity is limited to one (no diversity).
2) Regardless of the number of relays participating in a cooperation, there is no diversity, even if correlation coefficient of the actual and outdated CSI tends to one $(\rho \to 1)$.
3) From a practical point of view, it is worth designing a robust cooperative strategy to combat the outdated CSI.

To the best knowledge of the author, a few cooperative schemes to tackle the outdated CSI problem have been proposed until now. Taking advantage of Geo-location information, a scheme has been proposed in [34]. However, it makes sense only in a fixed wireless system, where the relays' locations do not change, rather than a mobile network. Another scheme taking into account the statistical knowledge of channel has been given in [35]. In spite of a remarkable increase of the implementation complexity, this scheme achieves merely a marginal performance gain and its diversity is always limited to one. Generalized selection combining [19] and its enhanced version called N plus Normalized Threshold Opportunistic Relay Selection (N+NT-ORS) [36] have also been applied. However, they require at least N orthogonal channels to retransmit, re-

sulting in a large loss of spectral efficiency.

In this context, we proposed a simple but effective scheme called Opportunistic Space-Time Coding (OSTC) [37] to alleviate the effect of the outdated CSI while avoiding an unnecessary loss of spectral efficiency. A predefined number N of relays, rather than a single relay in the conventional ORS, are opportunistically selected from K cooperating relays according to instantaneous CSIs of the relay-destination channels. At these selected relays, N-dimensional orthogonal space-time block coding (OSTBC) [38] is employed to encode the regenerated signals. N branches space-time coded signals are simultaneously transmitted from the selected relays to the destination, followed by a simple maximum-likelihood decoding based only on linear processing at the receiver. In contrast to DSTC where all relays participate in the signal's retransmission without a process of relay selection, only a subset of relays are opportunistically activated. Therefore, opportunistic space-time coding can be regarded as a combination of opportunistic relay selection and distributed space-time coding. Our research outcomes [39], [40] further recommend that the optimal number of relays to be selected is N=2. A pair of relays with the strongest and second strongest CSI at the relay-destination channels are selected, and the Alamouti scheme [41] is applied to encode the original signal. Since the Alamouti scheme is a unique space-time code achieving both full-rate and full-diversity with complex signal constellations, there is no spectral efficiency loss in comparison with the case of N>2. Another consideration of using N=2 is that the less number of relays can simplify the distributed synchronization. The practical timing and frequency synchronization schemes [42], [43] proposed for cooperative systems can help two relays to achieve a satisfied level of synchronization, whereas the time and frequency offsets become unacceptable with the increased number of relays.

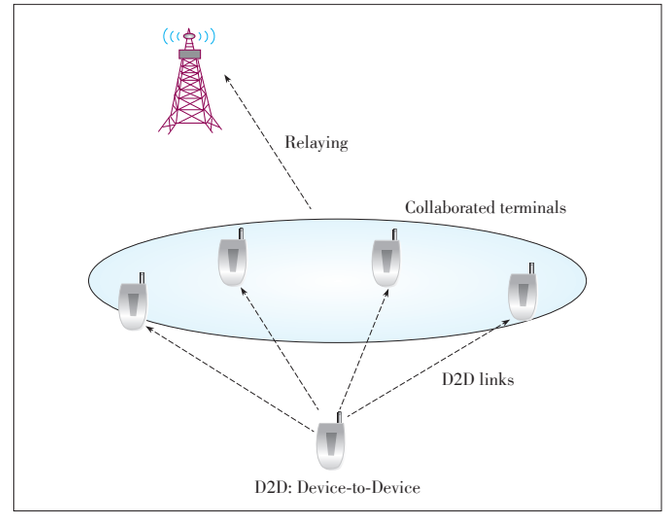
This paper gives a comparative review of the cooperative multi-relay transmission schemes. The rationale of different schemes, as well as their performance in terms of outage probability and channel capacity, are presented. The rest of this paper is organized into the following four sections. Section 2 introduces the system model of DF cooperative system. Section 3 illustrates the aforementioned cooperative schemes. In Section 4, simulation results are given. Finally, Section 5 concludes this paper.

2 Sytem Model

2.1 Multi-Relay Cooperative Network

A dual-hop decode-and-forward cooperative network is considered, where a terminal (the source) communicates with a base station (the destination) with the help of neighboring terminals (relays). Due to the line-of-sight blockage, a direct link between the source and the destination is assumed to be neglected. Because of severe signal attenuations in radio channels, a strong self-interference will be generated if a relay simultaneously transmits and receives signals at the same frequency. Currently, it is impractical to commercially implement a full-duplex mobile terminal. Hence, the relays have to operate in a half-duplex mode and the TDD scheme can be applied. It is generally assumed that the relays are equipped with a single antenna due to the limitations of cost, power supply and hardware size on mobile terminals. Although the spatially-distributed terminals with independent oscillators give rise to multiple timing offset and multiple carrier frequency offset, this multi-relay synchronization problem has been well-addressed and reported in the literature [42]–[45]. For simplicity, as most of papers in this field, we assume perfect synchronization throughout the rest of this paper.

As illustrated in **Fig. 1**, a base station provides cell coverage to a number of mobile terminals. A terminal may suffer from an out-of-coverage problem due to the blocking of buildings or a weak received signal when it locates at the cell edge. Besides, in the scenarios of disaster relief or emergency events as investigated in Aerial Base Stations with Opportunistic Links for Unexpected & Temporary Events (ABSOLUTE) project [46], this base station might be rapidly deployed without any network planning and optimization. In this case, the coverage is not good enough while the requirement of link reliability and system robustness is quite high. To improve the spectral efficiency at the cell edge, extend the signal coverage and improve the link reliability, cooperative communications can be applied by exploiting the broadcast nature of wireless signals. The mobile terminals cooperate with one another to communicate using the cooperative relaying. As shown in Fig. 1, the signals are first transmitted from the source terminal outside the coverage area of base station to its neighboring terminals through D2D communications. Those neighboring terminals that overheard this signal are capable of decoding and retransmitting. The distributed antennas at the terminals form a virtu-



▲**Figure 1. Principle of cooperative communications.**

al antenna array. By means of combining multiple copied versions of the original signal at the receiver, an inherent spatial diversity referred to as cooperative diversity can be achieved without any need of physical antenna array.

2.2 Outdated CSI

From a practical point of view, the channel information is imperfect due to the feedback delay and channel estimation error. In a traditional system such as the MIMO system, this imperfect CSI has a neglect effect mainly on the performance of signal detection. However, the relay selection scheme is far more vulnerable since the CSI is applied to not only detect a received signals but also select the best relay(s). The CSI at the time instant of relay selection denoted by h may substantially differ from the actual CSI $\hat{h}$ at the instant of data retransmission. Using the relays selected according to the outdated version of CSI rather than the actual CSI may make the wrong selection decision. To quantify the impact of imperfect CSI on the system, the envelop of correlation coefficient is defined as

$$\rho = \frac{\left|\mathrm{cov}(h,\hat{h})\right|}{\mu_h \mu_{\hat{h}}}, \tag{1}$$

where $\mathrm{cov}(\cdot)$ and μ stand for the covariance of two random variables and the standard deviation, respectively. The detail modeling of the outdated CSI and its statistics can be found in [47] and [48].

3 Cooperative Multi-Relaying Schemes

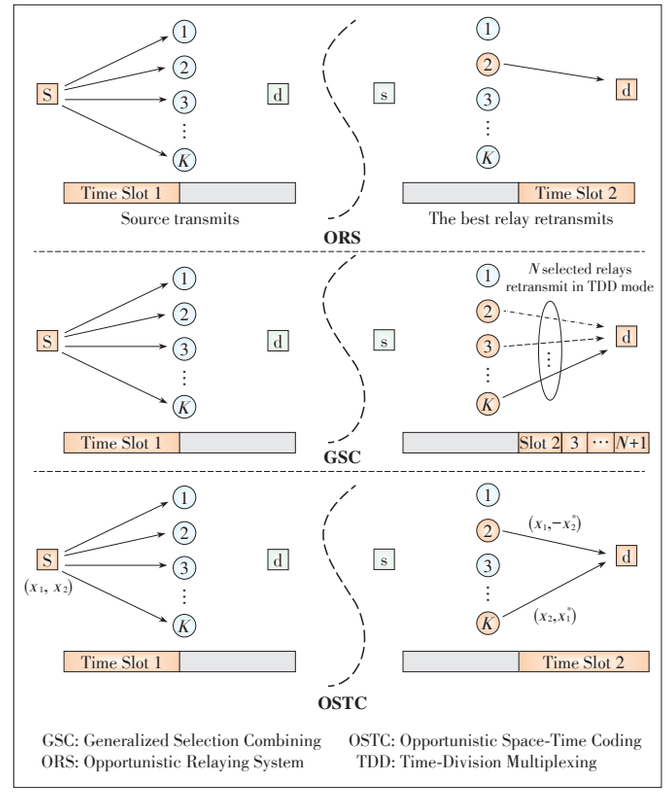
This section introduces mechanisms of different cooperative schemes, including ORS, GSC and OSTC. Because of the severe attenuation of radio signals, a single-antenna relay has to operate in a half-duplex mode to avoid harmful self-interference between the transmitter and receiver. Without loss of generality, the end-to-end signal transmission can be divided into two time slots: the broadcast and relaying phases. In the broadcast phase (Time Slot 1 indicated in **Fig. 2**), the source transmits and those relays that can correctly decode the original signal constitute a decoding subset:

$$DS \triangleq \left\{k : \frac{1}{2}\log_2(1+\hat{\gamma}_{SK}) \geq R\right\}, \tag{2}$$

where $\hat{\gamma}_{SK}$ is the Signal-to-Noise Ratio (SNR) of source-relay channel, and R stands for an end-to-end target data rate for the dual-hop relaying. Note that the required rate for each hop doubles to $2R$ owing to the half-duplex mode.

In the ORS scheme, the relay having the strongest SNR (interchangeable with CSI if the given transmit power for each relay is equal) in the relay-destination channels is selected from the decoding subset to serve as the best relay, i.e.,

$$\dot{k} = \arg\max_{k \in DS} \gamma_{kd}, \tag{3}$$



▲**Figure 2. Schematic diagrams of ORS, GSC and OSTC.**

where γ_{kd} denotes the instantaneous SNR of relay-destination channel at the instant of selecting relay, which may be outdated in comparison with the actual SNR $\hat{\gamma}_{kd}$ at the instant of using the selected relay to retransmit.

Instead of only a single relay, the GSC scheme selects N relays with the largest SNRs to retransmit the original signal in the second phase. In the relaying phase, as shown in Fig. 2, the time resource is divided into N sub-slots. Following the time-division multiplexing, each selected relay occupies one different slot to orthogonally retransmit the original signal. Equivalently, the frequency-division multiplexing can be applied to orthogonally retransmit the signal over relay-destination links, i.e., N subcarriers or sub-channels are used by N selected relays at the same time. Its enhanced version, the N+NT-ORS scheme, introduces a normalized threshold to further select qualified relays from the remaining $K-N$ relays. Although the number of selected relays may be different, the N+NT-ORS scheme still relays the signal in an orthogonal manner. These two schemes require at least N orthogonal channels to retransmit, resulting in a large loss of spectral efficiency.

In the DSTC scheme, no relay selection process is performed, but all relays within the current decoding subset are used to simultaneously retransmit by means of space-time coding. In this case, the number of participating relays is unknown and randomly varying since the decoding subset dynamically changes with the fluctuation of radio channels. Neither select-

ing a single relay in ORS nor all-participating in DSTC, the OSTC scheme chooses a predefined number N of relays. In the relaying phase, an N-dimensional orthogonal space-time block code is applied to encode the regenerated signals at the selected relays in a distributed manner. N branches coded signals are simultaneously transmitted by the selected relays at the same frequency, followed by a simple maximum-likelihood decoding based only on linear processing at the receiver. If the number of relays in the current decoding subset is denoted by L, we have $0 \leqslant L \leqslant K$. It is possible that $L < N$. In this case, all of L relays participate in the signal retransmission directly in combination with an L-dimensional orthogonal space-time block code.

For illustration purposes, we select N=2 and use the Alamouti scheme to clarify the OSTC scheme. In the broadcast phase, as illustrated in Fig. 2, the source sends a pair of symbols (x_1, x_2) to all relays within two consecutive symbol periods. Those relays that correctly decode the original signal constitute a decoding subset. In accordance to instantaneous SNRs of the relay-destination channels, a pair of best relays are opportunistically selected. In the relaying phase, the regenerated symbols are space-time encoded as:

$$(x_1, x_2) \rightarrow \begin{pmatrix} x_1 & -x_2^* \\ x_2 & x_1^* \end{pmatrix}, \tag{4}$$

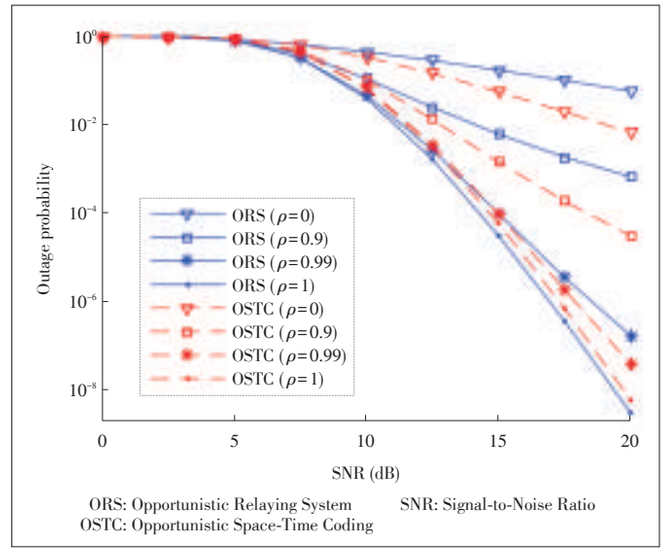
where the superscript * denotes the complex conjugate. Then, a relay transmits the first branch of coded symbols $(x_1, -x_2^*)$, while another relay sends (x_2, x_1^*) simultaneously at the same frequency, analogous to the Alamouti scheme applying for two co-located antennas in the MIMO system.

4 Performance Comparisons

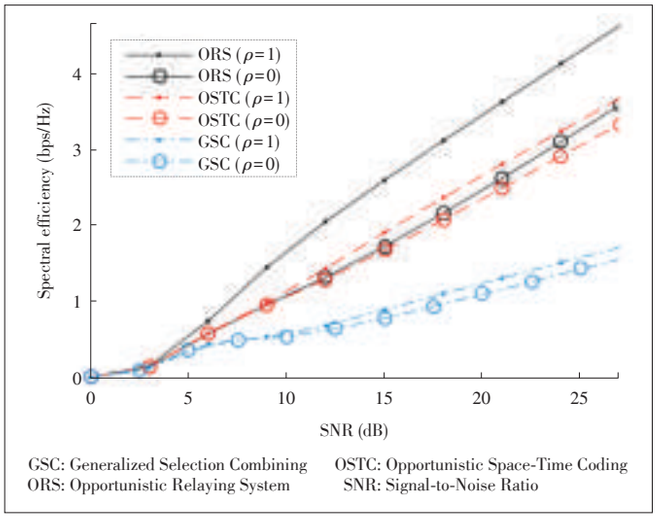
The Monte-Carlo simulations are setup in order to comparatively get the performance results in terms of outage probability and ergodic capacity. Given i.i.d. Rayleigh channels with a normalized gain, performance comparisons of ORS, GSC and OSTC in the absence and presence of imperfect channel quality are carried out. The numerical results are obtained by iterating 10^6 channel realizations into Monte-Carlo simulations, and the target rate is set to R=1 bps/Hz.

We first investigate the impact of the outdated CSI on a cooperative network with K=9 decode-and-forward relays. A single relay is selected for ORS, whereas N=2 relays are selected for OSTC in combination with the Alamouti scheme. During the simulation, it can be observed that GSC and OSTC achieve the same outage probability in any value of ρ. It is also theoretically proved in the literature that GSC and OSTC have the same performance in terms of outage probability. Hence, only the curves of OSTC is given in **Fig. 3** for simplicity. As shown in the figure, OSTC suffers from a little bit performance loss compared to ORS at the perfect channel quality. This is because a single relay with the strongest SNR transmits the regenerated signal in ORS, while a pair of relays with the strongest and second strongest SNR are utilized in OSTC. The relay with the second strongest SNR causes this performance gap. In the case of $\rho = 1$, the curve of OSTC is in parallel with its counterpart in ORS that has a full diversity. It can be therefore concluded that OSTC also achieves the diversity of $d = 9$, namely its outage probability decays at a rate of $1/\bar{\gamma}^9$ in the high SNR. In addition, OSTC's curves in the cases of $\rho = 0$, $\rho = 0.9$, and $\rho = 0.99$ are also provided. These curves are parallel among others in the high SNR with the diversity of 2, while the curves of ORS have the diversity of 1. That is to say, the diversity of ORS is one in the presence of outdated CSI, whereas an order of 2 is still kept by OSTC thanks to using N=2 selected relays. It can be observed that even in the case of 0.99, the outdated CSI brings an obvious performance degradation compared to the case of perfect CSI. When the correlation coefficient is reduced to 0.9, at the SNR of 20 dB, the outage probability is increased from 10^{-8} to 10^{-3}, which proves that vulnerability of the relay selection.

To shed light on the channel capacity of different schemes in the absence and presence of outdated CSI, their capacities as a function of the average SNR in the cases of $\rho = 0$ and $\rho = 1$ are shown in **Fig. 4**. The number of cooperating relays is assumed to be K=8 and the number of selected relays for OSTC and GSC is N=4. At the perfect CSI, OSTC suffers from a small capacity loss since the applied 4-dimensional OSTBC supports a maximal rate of only 3/4 in relay-destination link. The capacity loss of GSC is more severe due to the use of 4 orthogonal channels, equivalent to a rate of 1/4 in comparison with that of the ORS scheme. For example, ORS, OSTC and GSC achieve the spectral efficiencies of 4.3 bps/Hz, 3.4 bps/



▲**Figure 3. Outage probabilities of ORS and OSTC as a function of the average SNR. GSC and OSTC have the same performance in terms of outage probability.**



▲**Figure 4. Spectral efficiencies of ORS, GSC and OSTC as a function of the average SNR.**

Hz and 1.6 bps/Hz, respectively, at a given SNR of 25 dB. In the case of $\rho=0$, the spectral efficiency of OSTC closes to that of ORS with a loss of less than 0.2 bps/Hz. That is to say, despite a diversity gain of $d=4$ achieved by OSTC in the presence of outdated CSI, the price on the spectral efficiency loss is negligible. In comparison, GSC's spectral efficiency is around 1.5 bps/Hz at a given SNR of 25 dB, less than a half of ORS and OSTC. On the other hand, ORS is vulnerable to the outdated CSI because its spectral efficiency is reduced from 4.3 bps/Hz to 3.3 bps/Hz when the correlation coefficient is decreased to 0 from 1, equivalent to a loss of 1.0 bps/Hz. In contrast, the spectral efficiency loss is less than 0.3 bps/Hz for OSTC and GSC, implying their effectiveness of combatting the outdated CSI and their robustness feature.

5 Conclusions

The exponential growth of mobile traffic imposed a high pressure on the 5G system, where NR technologies have to be applied to substantially improve transmission performance, especially in the uplink. Taking advantage of D2D links, the cooperative communication can provide a remarkable performance gain in terms of spectral efficiency and reliability by means of collaborating the neighboring terminals. In this paper, we comparatively reviewed several cooperative multi-relay schemes in the presence of imperfect channel state information. The ORS scheme is easy to implement, and can achieve the full diversity, i.e., the number of all cooperating relays, at the perfect CSI. But in the presence of outdated CSI, the outage probability of ORS drastically deteriorates and its diversity degrades to one, i.e., no diversity. The GSC scheme is robust to the outdated CSI, while its capacity loss is large due to the orthogonal transmission. The proposed OSTC scheme opportunistically selects multiple relays, rather than a single relay, to decode and simultaneously retransmit the original signal by means of space-time coding. When the knowledge of CSI is perfect, it can achieve the full diversity. In the presence of outdated CSI, the diversity of N can still be kept. Besides, OSTC has a negligible capacity loss in comparison with that of ORS. Compared to DSTC, a fixed number of relays is used, instead of a random number, which makes sense for the practical systems. From the perspective of both performance and complexity, the OSTC scheme has been considered as the best solution until now.

References

[1] R. E. Hattachi and J. Erfanian, "NGMN 5G white paper," NGMN Alliance, Feb. 2015.
[2] A. Osseiran, F. Boccardi, V. Braun, et al., "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, May 2014. doi: 10.1109/MCOM.2014.6815890.
[3] J. G. Andrews, S. Buzzi, W. Choi, et al., "What will 5G be?" *IEEE Journal on Selected Areas Communications*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014. doi: 10.1109/JSAC.2014.2328098.
[4] S. Han, Y. C. Liang, Q. Chen, and B. H. Soong, "Licensed-assisted access for LTE in unlicensed spectrum: a MAC protocol design," *IEEE Journal on Selected Areas Communications*, vol. 34, no. 10, pp. 2550–2561, Oct. 2016. doi: 10.1109/JSAC.2016.2605959.
[5] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device communication in LTE-advanced networks: a survey," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 4, pp. 1923–1940, 2015. doi: 10.1109/COMST.2014.2375934.
[6] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: past, present, and future of programmable networks," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 3, pp. 1617–1634, 2014. doi: 10.1109/SURV.2014.012214.00180.
[7] R. Mijumbi, J. Serrat, J.-L. Gorricho, et al., "Network function virtualization: state-of-the-art and research challenges," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 1, pp. 236–262, 2016. doi: 10.1109/COMST.2015.2477041.
[8] D. Tse and P. Viswanath, "The wireless channe," in *Fundamentals of Wireless Communication*, 1st ed. Cambridge, UK: Cambridge University Press, 2005, ch. 2, sec. 1, pp. 21–31.
[9] G. J. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, pp. 311–335, Mar. 1998. doi: 10.1023/A:1008889222784.
[10] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE Journal on Selected Areas Communications*, vol. 31, no. 2, pp. 160–171, Feb. 2013. doi: 10.1109/JSAC.2013.130205.
[11] 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.
[12] H. Ishii, Y. Kishiyama, and H. Takahashi, "A novel architecture for LTE-B: C-plane/U-plane split and Phantom Cell concept," in *Proc. IEEE Globecom Workshops*, Anaheim, USA, 2012, pp. 624–630. doi: 10.1109/GLOCOMW.2012.6477646.
[13] T. Nakamura, S. Nagata, A. Benjebbour, et al., "Trends in small cell enhancements in LTE advanced," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 98–105, Feb. 2013. doi: 10.1109/MCOM.2013.6461192.
[14] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity—Part I and II," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1948, Nov. 2003. doi: 10.1109/TCOMM.2003.818096.
[15] T. M. Cover and A. A. E. Gamal, "Capacity theorems for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, Sept. 1979. doi: 10.1109/TIT.1979.1056084.
[16] J. N. Laneman, D. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behaviour," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004. doi: 10.1109/TIT.2004.838089.
[17] Z. Zhang, K. Long, A. V. Vasilakos, and L. Hanzo, "Full-duplex wireless communications: challenges, solutions, and future research directions," *Proceed-*

ings of the IEEE, vol. 104, no. 7, pp. 1369–1409, Jul. 2016. doi: 10.1109/JPROC.2015.2497203.
[18] M. Torabi, D. Haccoun, and J.-F. Frigon, "Relay selection in AF cooperative systems: an overview," *IEEE Vehicular Technology Magazine*, vol. 7, no. 4, pp. 104–113, Dec. 2012. doi: 10.1109/MVT.2012.2216751.
[19] L. Xiao and X. Dong, "Unified analysis of generalized selection combining with normalized threshold test per branch," *IEEE Transactions on Wireless Communications*, vol. 5, no. 8, pp. 2153–2163, Aug. 2006. doi: 10.1109/TWC.2006.1687731.
[20] Y. Jing and H. Jafarkhani, "Network beamforming using relays with perfect channel information," *IEEE Transactions on Information Theory*, vol. 55, no. 6, pp. 2499–2517, Jun. 2009. doi: 10.1109/TIT.2009.2018175.
[21] X. Luo, "Multiuser massive MIMO performance with calibration errors," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4521–4534, Jul. 2016. doi: 10.1109/TWC.2016.2542135.
[22] J. N. Laneman and G. W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2415–2425, Oct. 2003. doi: 10.1109/TIT.2003.817829.
[23] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE Journal on Selected Areas Communications*, vol. 24, no. 3, pp. 659–672, Mar. 2006. doi: 10.1109/JSAC.2005.862417.
[24] A. Bletsas, H. Shin, and M. Z. Win, "Cooperative communications with outage-optimal opportunistic relaying," *IEEE Transactions on Wireless Communications*, vol. 6, no. 9, pp. 3450–3460, Sept. 2007. doi: 10.1109/TWC.2007.06020050.
[25] J. L. Vicario, A. Bel, J. A. Lopez-Salcedo, and G. Seco, "Opportunistic relay selection with outdated CSI: outage probability and diversity analysis," *IEEE Transactions on Wireless Communications*, vol. 8, no. 6, pp. 2872–2876, Jun. 2009. doi: 10.1109/TWC.2009.081561.
[26] M. Seyfi, S. Muhaidat, J. Liang, and M. Dianati, "Effect of feedback delay on the performance of cooperative networks with relay selection," *IEEE Transactions on Wireless Communications*, vol. 10, no. 12, pp. 4161–4171, Dec. 2011. doi: 10.1109/TWC.2011.101711.100901.
[27] S. Kim, S. Park, and D. Hong, "Performance analysis of opportunistic relaying scheme with outdated channel information," *IEEE Transactions on Wireless Communications*, vol. 12, no. 2, pp. 538–549, Feb. 2013. doi: 10.1109/TWC.2012.122212.111556.
[28] M. Torabi, D. Haccoun, and J.-F. Frigon, "Impact of outdated relay selection on the capacity of AF opportunistic relaying systems with adaptive transmission over non-identically distributed links," *IEEE Transactions on Wireless Communications*, vol. 10, no. 11, pp. 3626–3631, Nov. 2011. doi: 10.1109/TWC.2011.092711.110136.
[29] M. Torabi and D. Haccoun, "Capacity of amplify-and-forward selective relaying with adaptive transmission under outdated channel information," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 5, pp. 2416–2422, Jun. 2011. doi: 10.1109/TVT.2011.2139232.
[30] M. Torabi and D. Haccoun, "Capacity analysis of opportunistic relaying in cooperative systems with outdated channel information," *IEEE Communications Letters*, vol. 14, no. 12, pp. 1137–1139, Dec. 2010. doi: 10.1109/LCOMM.2010.12.101179.
[31] N. S. Ferdinand, N. Rajatheva, and M. Latva-aho, "Effects of feedback delay in partial relay selection over Nakagami-m fading channels," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 4, pp. 1620–1634, May 2012. doi: 10.1109/TVT.2012.2187691.
[32] M. Soysa, H. A. Suraweera, C. Tellambura, and H. K. Garg, "Partial and opportunistic relay selection with outdated channel estimates," *IEEE Transactions on Communications*, vol. 60, no. 3, pp. 840–850, Mar. 2012. doi: 10.1109/TCOMM.2012.12.100671.
[33] O. Amin, S. S. Ikki, and M. Uysal, "On the performance analysis of multirelay cooperative diversity systems with channel estimation errors," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 5, pp. 2050–2059, Jun. 2011. doi: 10.1109/TVT.2011.2121926.
[34] B. Zhao and M. C. Valenti, "Practical relay networks: a generalization of hybrid-ARQ," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 7–18, Jan. 2005. doi: 10.1109/JSAC.2004.837352.
[35] Y. Li, Q. Yin, W. Xu, and H.-M. Wang, "On the design of relay selection strategies in regenerative cooperative networks with outdated CSI," *IEEE Transactions on Wireless Communications*, vol. 10, no. 9, pp. 3086–3097, Sept. 2011. doi: 10.1109/TWC.2011.072511.110077.
[36] M. Chen, T. C.-K. Liu, and X. Dong, "Opportunistic multiple relay selection with outdated channel state information," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 3, pp. 1333–1345, Mar. 2012. doi: 10.1109/TVT.2011.2182001.
[37] W. Jiang, T. Kaiser, and A. J. H. Vinck, "A robust opportunistic relaying strategy for co-operative wireless communications," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2642–2655, Apr. 2016. doi: 10.1109/TWC.2015.2506574.
[38] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1456–1467, Jul. 1999. doi: 10.1109/18.771146.
[39] W. Jiang, H. Cao, and T. Kaiser, "Opportunistic space-time coding to exploit cooperative diversity in fast-fading channels," in *Proc. IEEE ICC' 2014*, Sydney, Australia, Jun. 2014, pp. 4814–4819. doi: 10.1109/ICC.2014.6884082.
[40] W. Jiang, H. Cao, M. Wiemeler, and T. Kaiser, "Achieving high reliability in Aerial-Terrestrial networks: Opportunistic space-time coding," in *Proc. 2014 European Conference on Networks and Communications (EuCNC)*, Bologna, Itay, Jun. 2014, pp. 1–5. doi: 10.1109/EuCNC.2014.6882624.
[41] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas Communications*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998. doi: 10.1109/49.730453.
[42] A. A. Nasir, H. Mehrpouyan, S. Durrani, et al., "Transceiver design for distributed STBC based AF cooperative networks in the presence of Timing and Frequency offsets," *IEEE Transactions on Signal Processing*, vol. 61, no. 12, pp. 3143–3158, Jun. 15, 2013. doi: 10.1109/TSP.2013.2258015.
[43] Q. Huang, M. Ghogho, J. Wei, and P. Ciblat, "Practical timing and frequency synchronization for OFDM-based cooperative systems," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3706–3716, Jul. 2010. doi: 10.1109/TSP.2010.2046898.
[44] A. A. Nasir, H. Mehrpouyan, S. D. Blostein, S. Durrani, and R. A. Kennedy, "Timing and carrier synchronization with channel estimation in multi-relay cooperative networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 2, pp. 793–811, Feb. 2012. doi: 10.1109/TSP.2011.2174792.
[45] H. Mehrpouyan and S. D. Blostein, "Bounds and algorithms for multiple frequency offset estimation in cooperative networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 4, pp. 1300–1311, Apr. 2011. doi: 10.1109/TWC.2011.030311.101184.
[46] EU FP7. (2016, Oct.). *ABSOLUTE project* [Online]. Available: http://www.absolute-project.eu
[47] W. Jiang, H. Cao, and T. Kaiser, "Power optimal allocation in decode-and-forward opportunistic relaying," in *Proc. IEEE WCNC' 2014*, Istanbul, Turkey, Apr. 2014, pp. 1001–1006. doi: 10.1109/WCNC.2014.6952245.
[48] T. R. Ramya and S. Bhashyam, "Using delayed feedback for antenna selection in MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 12, pp. 6059–6067, Dec. 2009. doi: 10.1109/TWC.2009.12.090304.

Manuscript received: 2016-11-18

Biography

JIANG Wei (wei.jiang@dfki.de) received his Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), China. From March 2008 to June 2012, he worked in Central Research Institute of Huawei Technologies. In September 2012, he joined University of Duisburg-Essen, Germany, where he worked for EU FP7 ABSOLUTE project and H2020 5G-PPP COHERENT project. Since October 2015, he has joined German Research Center for Artificial Intelligence (DFKI) as a senior researcher and worked for H2020 5G-PPP SELFNET project. Meanwhile, he works for the University of Kaiserslautern, Germany as a senior lecturer and teaches wireless communications. He served as a vice chair of IEEE TCCN Special Interest Group "Cognitive Radio in 5G". He is the author of more than 30 papers and holds about 30 patent applications. He wrote a chapter "From OFDM to FBMC: Principles and Comparisons" for the book "Signal Processing for 5G" (Wiley and IEEE Press 2016).

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:

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

ZTE COMMUNICATIONS
Vol. 15 No. S1 (Issue 57)
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

---

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: June 25, 2017

Publication Licenses:
ISSN 1673-5188
CN 34-1294/ TN

Annual Subscription: RMB 80