Power Domain Non-Orthogonal Transmission for Cellular Mobile Broadcasting: Basic Scheme, System Design, and Coverage Performance

Zhengquan Zhang, Zheng Ma, Xianfu Lei, Ming Xiao, Cheng-Xiang Wang, and Pingzhi Fan

Abstract

Power domain non-orthogonal transmission is a promising technology for 5G wireless networks and beyond, as it can achieve higher spectrum efficiency than the orthogonal kind by multiplexing multiple users in the power domain. This article studies power domain non-orthogonal transmission for cellular mobile broadcasting to satisfy the increasing demands on multimedia communications in 5G and beyond. We first present two schemes for non-orthogonal transmission-based cellular mobile broadcasting: multi-rate and multi-service superposition transmissions, and then discuss their information-theoretical perspectives. Furthermore, we provide system designs for virtualized network architecture and physical layer processing, and discuss the key elements. We present a general superposition transmission framework to integrate three schemes developed by the 3GPP and to reduce the complexity of implementation, and then study constellation rotation to improve the BER performance of superposition transmission. Finally, we evaluate the SINR coverage performance of the presented schemes, followed by the main challenges and future research directions.

Introduction

Mobile data traffic is experiencing explosive growth and will be up to 1000 times higher by 2020 and beyond, of which more than 70 percent will be videos according to the Cisco forecast. This challenge activates researchers and standards bodies to develop fifth generation (5G) wireless networks and beyond. The International Telecommunication Union — Radiocommunication Standardization Sector (ITU-R) presents three usage scenarios for 5G networks — enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine type communications (mMTC) — and requires that 5G networks can provide superior capabilities, especially in data transmission rate and spectrum efficiency. To satisfy these challenging requirements, several fundamental technologies were discussed in [1], such as massive multiple-input multiple-output (MIMO), millimeter-wave (mmWave) communications, and non-orthogonal multiple access (NOMA). Power domain NOMA [2–4] will be a promising technology for 5G and beyond, due to its high spectrum efficiency, low latency, and the support of massive connectivity. The main idea of NOMA is to multiplex multiple users in the power domain on the same radio resource blocks by superposition coding [5] and decode the desired message from the superposed signal by successive interference cancellation (SIC) [5]. Both single-cell NOMA [3] and multi-cell NOMA [4] were studied to clarify the performance gain of NOMA. In addition, the combination of NOMA with other promising technologies were studied, such as MIMO-NOMA for mmWave communications [6]. Power domain NOMA has already been adopted by some practical systems. For example, the Third Generation Partnership Project (3GPP) LTE-Advanced Pro networks employed multi-user superposition transmission (MUST) [3] (i.e., NOMA) to enhance downlink transmission, and the Advanced Television Standards Committee (ATSC) developed the layer-division multiplexing (LDM) scheme [7] as a variation of NOMA to the next-generation digital television system, ATSC3.0.

On the other hand, broadcasting/multicasting [8] over cellular networks (i.e., cellular mobile broadcasting) will be an important spectrum-efficient technology for multimedia communications in 5G and beyond. This technology adopts the point-to-multipoint (PTM) transmission mechanism to distribute the media content to all interested users on the same channel. Currently, cellular mobile broadcasting can use either a multimedia broadcast/multicast service single-frequency network (MBSFN) transmission (i.e., multi-cell PTM) [9] or single-cell PTM (SC-PTM) transmission [10] to transmit multimedia broadcast/multicast services (MBMS). With the ever increasing demands for multimedia communications, for example, 4/8K ultra-high definition (UHD) videos, hybrid mobile and fixed TV services, and emerging broadcast applications such as broadcast-like [10] and group-oriented machine-type services [11], the research on highly spectrum-efficient 5G cellular mobile broadcasting has already attracted increasing attention. The Next Generation Mobile Networks (NGMN) alliance presented the following requirements: 200 Mb/s downlink user experienced data rate and < 200 ms end-to-end latency [9]. The challenges and key technologies for 5G cellular mobile broadcasting were also discussed in [9, 11].

Recently, some efforts have been devoted to

Digital Object Identifier: 10.1109/MWC.2018.1700125

Zhengquan Zhang, Zheng Ma, Xianfu Lei, and Pingzhi Fan are with Southwest Jiaotong University; Xianfu Lei is also with Southeast University; Ming Xiao is with KTH Royal Institute of Technology; Cheng-Xiang Wang is with Heriot-Watt University.

1536-1284/18/$25.00 © 2018 IEEE

studying the application of power domain NOMA to cellular mobile broadcasting in order to further improve the spectrum efficiency. Minimum transmit power multicast beamforming with superposition coding [12] was studied for NOMA systems. The superposition transmission of multicast and unicast streams [13] by power domain NOMA was studied. The system modeling and performance analysis of power domain NOMA-based MBMS transmission was also provided in [14]. In this article, we systematically present the power domain non-orthogonal 5G cellular mobile broadcasting in terms of fundamental principle, transmission schemes, system design, physical layer processing, key elements, and performance evaluation. First, we present two schemes for non-orthogonal cellular mobile broadcasting: multi-rate superposition transmission (MRST) and multi-service superposition transmission (MSST). The former explores the difference in channel conditions among users to increase the data transmission rate and improve strong users' quality of service (QoS), while the latter utilizes this channel difference to deliver multiple services to different user groups on the same channel simultaneously. Then we discuss the information-theoretical perspectives of the presented schemes. Next, we provide system designs for network architecture and physical layer processing together with key elements. Software defined networking (SDN) and network functions virtualization (NFV) together with cloud/edge/fog computing [15] and caching technologies are introduced to the design of virtualized network architecture. A general framework for superposition coded modulation (SCM) [5] is presented to integrate three superposition transmission schemes developed by 3GPP and reduce the complexity of implementation, and then the bit error rate (BER) performance enhancement by using constellation rotation is investigated. A joint iterative SCM demodulation and channel decoding receiver scheme is also studied. Furthermore, we evaluate the signal-to-interference-plus-noise ratio (SINR) coverage performance of the presented schemes. Finally, we identify the main challenges and future research directions, followed by the conclusions.

Schemes for Non-Orthogonal Cellular Mobile Broadcasting

Two schemes for non-orthogonal cellular mobile broadcasting are studied: MRST and MSST. For N-layer power domain non-orthogonal transmission, the power P_n with power ratio $\alpha_{P,n}$ is allocated to the nth layer, and then all layers are superposed to form a signal by superposition coding, while each layer can be decoded from the superposed signal by SIC. Without loss of generality, two-layer non-orthogonal transmission is considered, where the two layers are called the *primary layer* and the *secondary layer*, respectively. The primary layer refers to the first layer of two-layer non-orthogonal transmission, while the secondary layer refers to the second layer.

Multi-Rate Superposition Transmission

Multi-rate multicasting [9] delivers multimedia contents with different data rates, such that users can decode the corresponding content according to their channel conditions. This can utilize the difference in channel conditions among users to improve strong users' QoS, as well as guarantee the QoS of weak users. With power domain non-orthogonal transmission, MRST can be achieved by splitting the power domain to multiplex multiple data streams with different data rates on the same channel. The multi-rate superposition transmission for scalable and non-scalable multimedia are presented as follows.

Scalable Multimedia Transmission: Scalable multimedia[1] is coded into one basic data layer and one or more enhanced data layers through source layered coding.[2] Then these data are processed by channel coding and modulation, and form a superposed signal by allocating different power levels. Since the basic data achieve the basic service quality, it is carried by the primary layer with high priority and more power, while the enhanced data are carried by the secondary layer with the rest of the power. When users receive this superposed signal, they directly decode the primary layer to obtain the basic data, and then try to decode the secondary layer through successive interference cancellation (SIC). Therefore, weak users can only decode the primary layer to obtain the basic service quality, while strong users can decode both the primary and secondary layers to obtain better services.

Non-Scalable Multimedia Transmission: The non-scalable multimedia data is first coded and modulated with low data rate as the primary layer. The same multimedia data is also encoded and modulated with high data rate as the secondary one. Then these two modulated symbols are superposed with different power levels. After receiving this superposed signal, weak users decode the primary layer to obtain the low-rate data stream, while strong users decode the secondary layer to obtain the high-rate data stream to achieve better service quality.

Multi-Service Superposition Transmission

Multi-service superposition transmission provides different services for different users or user groups on the same radio resources, by multiplexing these services in the power domain. For example, one mobile HDTV service to handheld devices and one 4k-UHDTV service to fixed TV reception terminals are carried on the same channel [7]. More specifically, take two-service superposition transmission for an example: the primary layer carries the high-priority service with more power, while the remaining power is allocated to the secondary layer to carry the low-priority service. After receiving the superposed signal, high-priority users directly decode the primary layer to obtain the high-priority service, while low-priority users decode the secondary layer through SIC to recover the low-priority service.

In cellular networks, both unicast and broadcast/multicast services can be supported. Therefore, MSST can fall into three categories: unicast and unicast, unicast and broadcast/multicast, and broadcast/multicast and broadcast/multicast superposition.

Unicast and Unicast Superposition Transmission: It co-schedules multiple users on the same radio resources without spatial separation, for example, MUST [3]. In general, it is recommended that one strong user and one

Multi-service superposition transmission provides different services for different users or user groups on the same radio resources, by multiplexing these services in the power domain. For example, one mobile HDTV service to handheld devices and one 4k-UHDTV service to fixed TV reception terminals are carried on the same channel.

[1] Scalable multimedia refers to multimedia that can be represented in multiple layers with different quality and frame rates.

[2] Source layered coding is a technique used to divide media into multiple data layers with different quality and frame rates.

Non-orthogonal transmission eliminates the orthogonality of conventional transmission schemes adopted by 2/3/4G networks in order to achieve high spectrum efficiency, although this orthogonality can alleviate the aggregate interference at users by allocating distinct radio resources in certain resource domains to multiple users.

weak user should be selected as a user pair. For this category, optimal user pairing and power allocation are the key to improving network capacity. With larger channel gain differences between the users, higher network capacity gain can be obtained.

Unicast and Broadcast/Multicast Superposition Transmission: In current cellular networks, broadcast/multicast services are delivered by sharing radio resources with unicast services based on time-division multiplexing (TDM), which limits the maximum broadcast throughput. Through superposition transmission, unicast and broadcast/multicast services can be superposed in the power domain [13], which enables broadcast/multicast services to occupy all of the time-frequency resources together with unicast services. This can improve the throughput of cellular mobile broadcasting in a shared network.

Broadcast/Multicast and Broadcast/Multicast Superposition Transmission: It multiplexes multiple broadcast/multicast services on the same radio resources by allocating different power levels, which can increase the system throughput and provide mixed services for different user groups simultaneously, for example, mobile users and fixed TV reception terminals.

Information-Theoretical Perspective on Non-Orthogonal Cellular Mobile Broadcasting

Non-orthogonal transmission eliminates the orthogonality of conventional transmission schemes adopted by 2/3/4G networks in order to achieve high spectrum efficiency, although this orthogonality can alleviate the aggregate interference at users by allocating distinct radio resources in certain resource domains to multiple users. In this section, we discuss the information-theoretical perspective on non-orthogonal cellular mobile broadcasting and also give a brief discussion from the viewpoint of degree of freedom [6].3 The two-layer non-orthogonal transmission employs power split $P_1 + P_2 = P$, where $P_i \in (0, P)$ is the power allocated to the signal of the ith layer and P is the average power constraint of the transmit signal. The power ratio can be expressed as $\alpha_P = P_1/P$. τ_i ($i = 1,2$) is the normalized broadcast rate for the ith layer by the bandwidth W, which relies on the specific broadcast schemes. It is assumed that if a user's channel capacity is no smaller than the fixed broadcast rate, the user can decode the broadcast data successfully.

Multi-Rate Superposition Transmission

For a multi-rate broadcast area consisting of N BSs, M users who are interested in the broadcast service can be categorized into a weak user set with $M_1 \in (0, M)$ users and a strong user set with $M_2 = M - M_1$ users. Note that users in the strong user set can decode the primary and secondary layers to obtain both the basic and enhanced data, while users in the weak user set can only decode the primary layer to obtain the basic data. We first study the data rate for individual users in the broadcast area. Without loss of generality, one arbitrary user (i.e., user $m1$) from the weak user set and one arbitrary user (i.e., user $m2$) from the strong user set are discussed. The corresponding channel gains of users $m1$ and $m2$ are h_{m1} and h_{m2}, and satisfy $|h_{m1}|^2 \leq |h_{m2}|^2$. As a result, the data rate for users $m1$ and $m2$ are $R_{m1} = \tau_1 \mathbb{I}(E^1_{m1,1})$ and $R_{m2} = \tau_1 \mathbb{I}(E^1_{m2,1}) + \tau_2 \mathbf{I}(E^1_{m2,1} \,\&\&\, E^1_{m2,2})$, respectively, where

$$E^1_{mi,1} = \left\{ C\left(\frac{|h_{mi}|^2 P_1}{|h_{mi}|^2 P_2 + N_0} \right) \geq \tau_1 \right\}$$

in the event that user m_i's channel capacity decoding the message x_1 is no smaller than the fixed broadcast rate τ_1,

$$E^1_{m2,2} = \left\{ C\left(\frac{|h_{m2}|^2 P_2}{N_0} \right) \geq \tau_2 \right\}$$

in the event that user $m2$'s channel capacity decoding the message x_2 after SIC is no smaller than the fixed broadcast rate τ_2, $C(x) \triangleq \log_2(1 + x)$ for the channel capacity, $\mathbb{I}(.)$ for the indicator function, and N_0 for the power spectral density (PSD) of white Gaussian noise. Therefore, the average broadcast throughput per BS can be expressed as

$$R = \left(\sum_{m1=1}^{M_1} R_{m1} + \sum_{m2=1}^{M_2} R_{m2} \right) N^{-1}. \tag{1}$$

For single-rate broadcasting with fixed data rate τ_1, the average broadcast throughput per BS is

$$R = \left(\sum_{m=1}^{M} \tau_1 \mathbb{I}\left(E^2_m\right) \right) N^{-1},$$

where the event

$$E^2_m = \left\{ C\left(\frac{|h_m|^2 P}{N_0} \right) \geq \tau_1 \right\}$$

means that user m's channel capacity decoding message x_1 is no smaller than τ_1. Obviously, for single-rate broadcasting, the strong users cannot fully utilize the degree of freedom, as media contents are distributed with low data rate in order to ensure that interested users within the broadcast coverage can decode them successfully. However,the MRST scheme can enable strong users to overcome this shortage to obtain better QoS, at the cost of some SINR loss and the increase of processing complexity.

Multi-Service Superposition Transmission

For a multi-service broadcast area with N BSs, there are M users in the broadcast service, where $M_1 \in (0, M)$ high-priority users are served by the high-priority service, and $M_2 = M - M_1$ low-priority users are served by the low-priority service. τ_1 is the fixed broadcast rate for the high-priority service, while τ_2 is for the low-priority service. For MSST, high-priority users directly decode the desired service by regarding low-priority service as noise, while low-priority users decode their corresponding service from the superposed signal through SIC. Therefore, the average broadcast throughput per BS of multi-service broadcast can be expressed as in Eq. 1, where $R_{m1} = \tau_1 \mathbb{I}(E^1_{m1,1})$ and $R_{m2} = \tau_2 \mathbb{I}(E^1_{m2,1} \,\&\&\, E^1_{m2,2})$ are the data rates for user $m1$ in the high-priority service area and user $m2$ in the low-priority service area, respec-

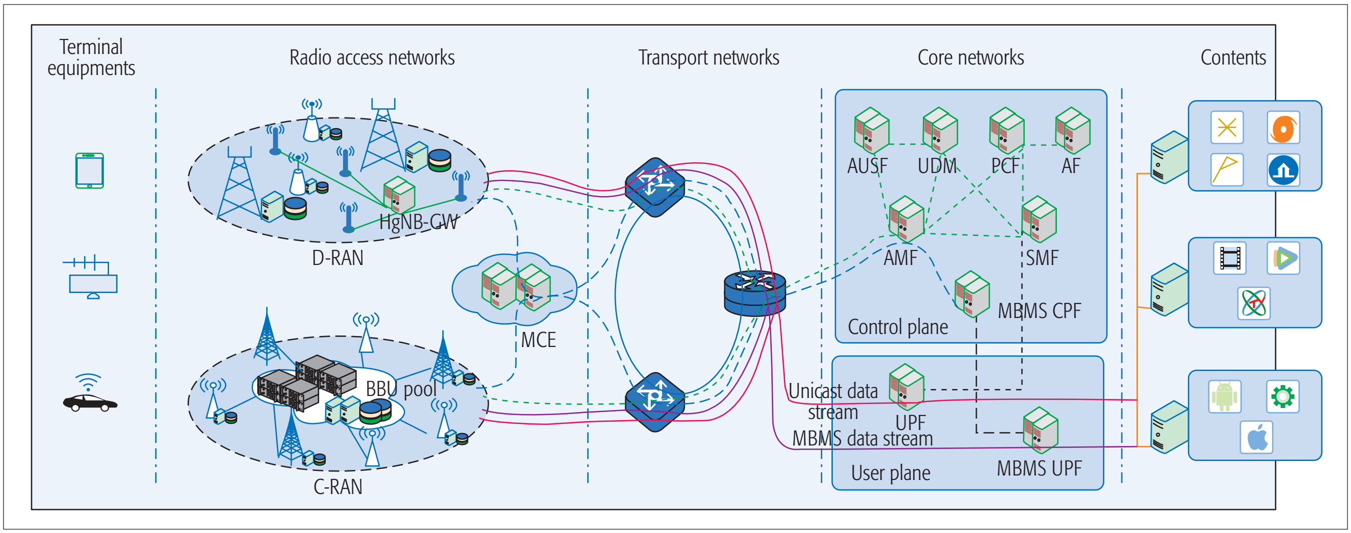


FIGURE 1. Network architecture for non-orthogonal cellular mobile broadcasting.

tively. For orthogonal cell mobile broadcasting, two services are transmitted by allocating two orthogonal radio resources with degree-of-freedom split β ∈ (0, 1). Therefore, the average broadcast throughput per BS can expressed as in Eq. 1, where the data rate for user $m1$ is $R_{m1} = \tau_1 \mathbb{I}(E_{m1}^3)$ with the event

$$E_{m1}^3 = \left\{ \beta C\left(\frac{|h_1|^2 P_1}{\beta N_0} \right) \geq \tau_1 \right\},$$

while the data rate for the user $m2$ is $R_{m2} = \tau_2 \mathbb{I}(E_{m2}^3)$ with the event

$$E_{m2}^3 = \left\{ (1-\beta) C\left(\frac{|h_2|^2 P_2}{(1-\beta) N_0} \right) \geq \tau_2 \right\}.$$

Regarding the degree of freedom, MSST can obtain full degrees of freedom at the cost of some SINR loss of high-priority users and increased processing complexity of low-priority users, while the orthogonal one suffers a certain loss of degree of freedom.

System Design for Non-Orthogonal Cellular Mobile Broadcasting

This section discusses overall system design for non-orthogonal cellular mobile broadcasting systems including network architecture and physical layer processing. The innovative network architecture based on SDN/NFV and cloud/edge/fog computing together with enhanced physical layer processing based on non-orthogonal transmission is presented to satisfy the challenging requirements for multimedia communications in 5G and beyond. The presented system design for non-orthogonal cellular mobile broadcasting can also be backward compatible with the orthogonal one by allocating all power to the primary layer.

Network Architecture

SDN and NFV [15] are the two most important technologies to design the flexible, efficient, and scalable next-generation network architecture, where the key idea of SDN is to separate the control and user planes of networks, while NFV virtualizes the functions of network nodes into building blocks and provides the logical network slices such as broadcast/multicast or unicast. Based on SDN and NFV, 3GPP TS 23.501 presented an innovative system architecture for 5G systems with unicast transmission, in which the control plane consists of core access and mobility management function (AMF), session management function (SMF), and so on, while user plane function (UPF) is for user data forwarding. Figure 1 illustrates the network architecture for non-orthogonal cellular mobile broadcasting, which can enable multimedia contents to be transmitted by unicast or broadcast/multicast. The MBMS-control plane function (MBMS-CPF) entity hosts the control functions of the conventional broadcast/multicast service center (BMSC), the MBMS-gateway (MBMS-GW), and provides MBMS user service provisioning and session management. The MBMS-user plane function (MBMS-UPF) entity, which consists of the data forwarding functions of the conventional BMSC and MBMS-GW, achieves the delivery of MBMS packets to each BS. The multi-cell coordination entity (MCE) provides not only the synchronization among the cells in the MBSFN area, but also radio resource management and radio configuration. Radio access networks (RANs) with heterogeneous and ultra-dense properties achieve radio transmission to users, which shall be deployed in the form of mixed cloud and fog RANs. Cloud RANs (C-RANs) utilize centralized cloud computing [15] to efficiently provide such applications as regional and national broadcast, while fog RANs benefitting from the distributed network deployment utilize edge and fog computing [15] residing at the network edge to efficiently achieve such applications as local broadcast. Furthermore, cloud-fog computing coordination intelligently distributes user computing tasks to cloud or fog computing units, based on the decision results of intelligent algorithms that consider factors including computing task attributions and computing resource status comprehensively. Finally, local caches are deployed at the network edge, and then the popular contents are prefetched and put in these

SDN and NFV are the two most important technologies to design the flexible, efficient, and scalable next-generation network architecture, where the key idea of SDN is to separate the control and user planes of networks, while NFV virtualizes the functions of network nodes into building blocks and provides the logical network slices such as broadcast/multicast or unicast.

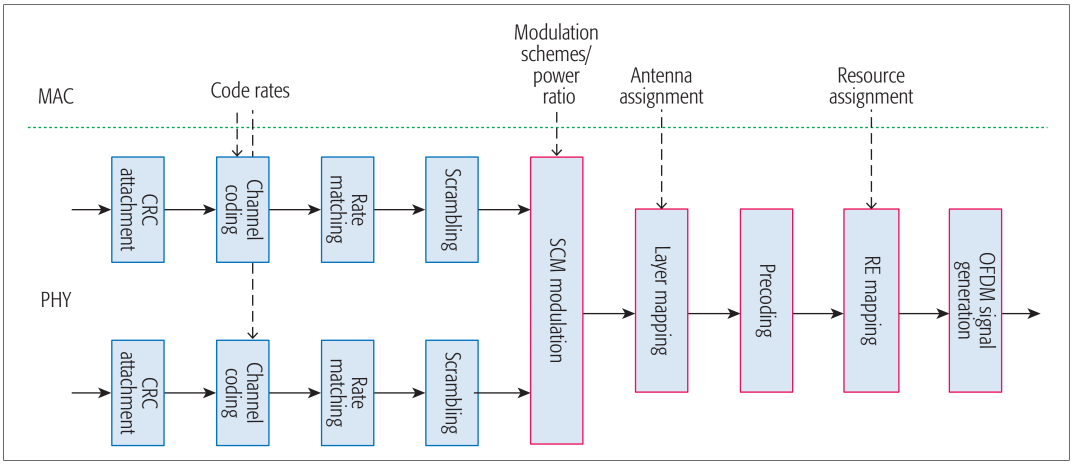


FIGURE 2. Basic physical layer processing of two-layer non-orthogonal cellular mobile broadcasting.

local caches according to the popularity distributions and user preferences.

Physical Layer Processing

Based on the physical layer processing specified in 3GPP TS 38.201, we present the physical layer processing of two-layer non-orthogonal cellular mobile broadcasting illustrated in Fig. 2. The data blocks from the medium access control (MAC) layer are processed by the steps of cyclic redundancy check (CRC) attachment, channel coding, rate matching, scrambling, superposition coded modulation, layer mapping, precoding, resource element mapping, and orthogonal frequency-division multiplexing (OFDM) signal generation sequentially, and then are transmitted over the radio links. At the receiver side, the data streams can be recovered after the inverse processing. The functions of corresponding processing steps are as follows.

CRC Attachment: Provide error check for each data block by attaching a parity sequence.

Channel Coding: Achieve reliable data communications through forward error correction (FEC) codes such as low-density parity check (LDPC) code. LDPC encodes the K information bits based on a sparse parity matrix $H_{K \times N}$ to obtain the N coded bits, which leads to superior performance approaching the Shannon capacity limit. Further, the concatenation of two FEC codes such as Bose-Chaudhuri-Hocquenghem (BCH) as outer code and LDPC as inner code can further improve the FEC performance.

Rate matching: Interleave the coded bits in each code block and adjust its size to match the number of bits contained in the allocated radio resources through bit selection and pruning. After interleaving, the bit order is disrupted at the transmitter, while burst errors can be distributed after de-interleaving at the receiver. Therefore, the ability of FEC codes to correct burst errors can be significantly improved.

Scrambling: Scramble the bits of each codeword by using broadcast-service-area-specific scrambling sequence to randomize the interference.

Superposition Coded Modulation: Transform codewords to complex-valued modulation symbols suitable for channel transmission, and combine these modulation symbols with adaptive power ratio to form a superposed symbol. In general, the SCM composite constellation, which is generated by the superposition of multiple component constellations with different power levels, is non-uniform.

Layer Mapping: Achieve the complex-valued symbol mapping onto the given transmission layers.

Precoding: Precode the complex-valued symbols on each layer for transmission on the antenna ports.

Resource Element Mapping: Map the precoded complex-valued symbols for each antenna port to resource elements.

OFDM Signal Generation: Use inverse fast Fourier transform (IFFT) to generate complex-valued time-domain OFDM signal for each antenna port.

Point-to-Multipoint Transmission Mode

Single-Cell Point-to-Multipoint (SC-PTM): It enables a single cell to serve a group of users within its coverage by PTM transmission, which can efficiently provide services for localized demands. SC-PTM can be used for critical communications such as public safety [10] and other commercial use cases, for example, mobile advertising. In such use cases, users are located within a local geographical area and have a common interest in the service/content. In addition, the broadcast area is dynamic and may be a single cell.

Multi-Cell Point-to-Multipoint (MC-PTM): Unlike use cases for SC-PTM, in certain use cases such as TV services, users are distributed in a large pre-planned and rather static area. MC-PTM transmission is more efficient to support these use cases, by enabling all cells in the broadcast area to deliver the same media content to all interested users on the same channel synchronously [9]. With synchronous transmission, users, especially at the cell edge, can receive and combine multiple signals from multiple cells to improve the quality of signals rather than contribute to the aggregate interference. In addition, users on the edge of the broadcast area also receive inter-cell interference from BSs belonging to other broadcast areas. To reduce the processing complexity of users, inter-cell interference is simply treated as noise.

Key Elements for Non-Orthogonal Cellular Mobile Broadcasting

Superposition Coded Modulation

Superposition coded modulation is used to achieve non-orthogonal transmission in the power domain, by mapping the coded bits in multiple codewords to the complex-valued symbols that are superposed with adaptive power ratio. To achieve adaptive power ratio, the scheduler in the MAC layer chooses power ratio according to channel state information (CSI) reported by users in each scheduling interval. 3GPP studied three SCM schemes [3], including SCM Category 1 with non-Gray-mapped composite constellation and adaptive power ratio on component constellations, SCM Category 2 with Gray-mapped composite constellation and adaptive power ratio on component constellations, and SCM Category 3 with label-bit assignment on Gray-mapped composite constellation. Note that SCM Category 3 can be considered as a special case of SCM Category 2. To guarantee the Gray property, SCM Category 2 needs more complicated construction rules than SCM Category 1, which just performs direct superposition. However, due to the loss of the Gray property, SCM Category 1 has higher BER than SCM Category 2.

Figure 3 illustrates a general framework for SCM to integrate SCM Categories 1, 2, and 3 developed by 3GPP, which consists of shuffling, modulation, rotation, power allocation, and combining modules. In this framework, constellation rotation is used to improve the modulation performance. The basic process of SCM is: the coded bits of codeword C_1 are mapped to complex-valued symbol, $S_1 = I_1 + jQ_1$, while the coded bits of the codeword C_2 are mapped to complex-valued symbol, $S_2 = e^{j\theta}(I_2 + jQ_2)$, after shuffling, modulation and constellation rotation, where θ is the rotation angle and $\theta = 0$ means no rotation. Note that the modulation module employs the 3GPP standard modulation schemes. Finally, these two complex-valued symbols are superposed with adaptive power ratio to form a superposed symbol S. Therefore, the output complex-valued symbol of SCM can be expressed as

$$S = \left(\sqrt{\alpha_p}(I_1 + jQ_1) + \sqrt{1-\alpha_p e^{j\theta}}(I_2 + jQ_2)\right),$$

where $0 < \alpha_p < 1$ is the power ratio. Shuffling the coded bits of codeword C_2 is to guarantee SCM the composite constellation's Gray property, and its specific rules depend on the coded bits of codeword C_1 and specific composite constellations. Note that there is no shuffling for SCM Category 1 due to its non-Gray mapping. Note that for the number of codewords $N > 2$, N-layer non-orthogonal transmission can be used. Based on specific design rules, optimization problems can be formulated and solved to obtain optimal power ratios and rotation angles.

SCM suffers a certain loss of BER performance due to power split and inter-layer interference caused by superposition operation. As shown in Fig. 3b, with power ratio, α_p, the minimum Euclidean distances for quadrature phase shift keying (QPSK) constellations of codewords C_1 and C_2 are $d_{1,\min} = \sqrt{2(1-\alpha_p)}$ and $d_{2,\min} = \sqrt{2\alpha_p}$, respectively, which are smaller than standard QPSK constellation with $d_{\min} = \sqrt{2}$. After constellation rotation and superposition, the minimum Euclidean distance of the composite constellation reduces to

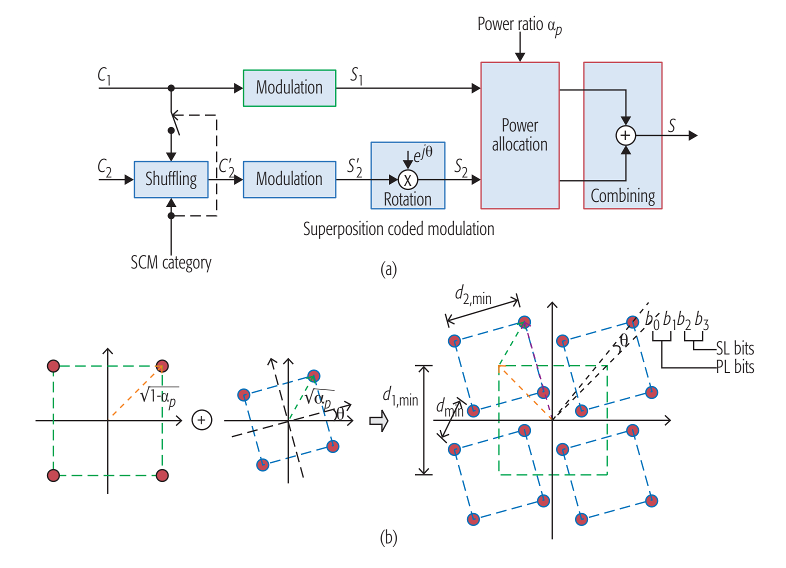


FIGURE 3. Superposition coded modulation: a) general SCM modulation framework; b) QPSK/QPSK example.

$$d_{\min} = \sqrt{2(1-2)\sqrt{\alpha_p(1-\alpha_p)}\cos\theta},$$

which means that inter-layer interference further deteriorates the BER performance of codewords C_1. When $\theta = 0$, there is no rotation. Obviously, the minimum Euclidean distance of superposition transmission with constellation rotation is larger than the one without rotation. Therefore, constellation rotation can improve the performance of superposition transmission.

Joint Iterative SCM Demodulation and Channel Decoding

SIC is employed to decode the desired data from the superposed signal, which follows the processing: weak users decode their own data directly by treating the signal of strong users as noise, while strong users decode the desired data through SIC. Generally, SCM demodulation and channel decoding are performed independently, which cannot achieve the optimal receiving performance. Figure 4 illustrates a joint iterative demodulation and channel decoding receiver scheme for non-orthogonal transmission, which consists of SCM demodulation, de-interleaving, channel decoding, interleaving, and constellation probability calculating modules. The joint iterative SCM demodulation and channel decoding scheme follows these steps:

- The received symbols together with a priori probability information from the outer loop are input to the SCM demodulation module to generate log-likelihood ratios (LLRs).
- These LLRs are deinterleaved, and then the deinterleaved LLRs are input to the channel decoding module.

SIC is employed to decode the desired data from the superposed signal, which follows the processing: weak users decode their own data directly by treating the signal of strong users as noise, while strong users decode the desired data through SIC. Generally, SCM demodulation and channel decoding are performed independently, which cannot achieve the optimal receiving performance.

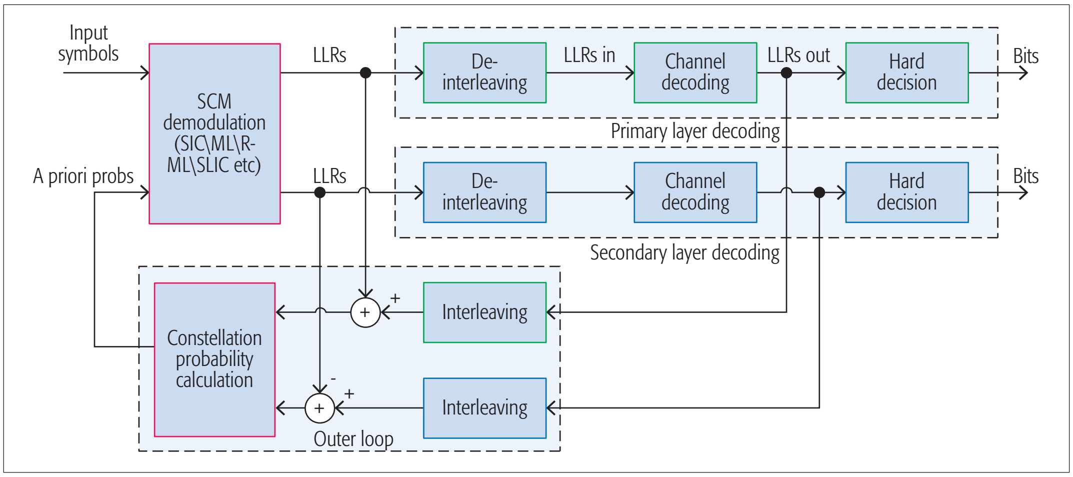


FIGURE 4. Joint iterative demodulation and decoding scheme for superposition transmission.

- The output LLRs after channel decoding are interleaved, and then the interleaved LLRs together with the demodulated LLRs are used to generate extrinsic LLRs.
- The extrinsic LLRs are input to the constellation probability calculating module to generate a priori probability information.
- Finally, the a priori probability information is fed back to the SCM demodulation module. This iterative processing will continue until a certain stop condition is met. During the iterative processing, the EXIT chart tool can be used to help the analysis of the extrinsic information and the optimal algorithm design.

User Association

The user association procedure [14] is used to associate a user with a specific cell before data transmission, which consists of cell selection/reselection and random access. A user first performs cell selection/reselection to select the most suitable cell for camping. The user first measures reference signal receiving power (RSRP) and then selects the most suitable cell according to a certain criterion, such as maximum average receiving power. Furthermore, the user performs a random access procedure to establish radio resource control (RRC) connection with its serving BS through the following four steps:

- Random access preamble. The user randomly selects a random access preamble (RAP) from a RAP pool (e.g., 64 preambles for LTE networks) and then transmits it to the selected BS (i.e., Msg1).
- Random access response. When detecting a RAP, the BS replies with a random access response (i.e., Msg2) carrying RAP identifier, initial uplink grant, and temporary cell-radio network temporary identifier (C-RNTI).
- Scheduled transmission. When the user receives a successful response, it performs the first scheduled uplink transmission (i.e., Msg3), which conveys user identity, RRC Connection Request, and so on.
- Contention resolution. The BS responds with Msg4 conveying user identity, RRC Connection Setup, and so on. After receiving Msg4, the user compares user identities included in Msg3 and Msg4. If they are matched, the contention resolution is considered to be successful, and the temporary C-RNTI is promoted to C-RNTI.

Multi-User Group Scheduling and Power Allocation

Multi-user group scheduling and power allocation are two important functions of the MAC layer in non-orthogonal cellular mobile broadcasting systems, as it hosts radio resource allocation among active user groups in order to satisfy their QoS requirements. For SC-PTM, BSs perform multi-user group scheduling and power allocation, while it is done in MCE for MC-PTM.

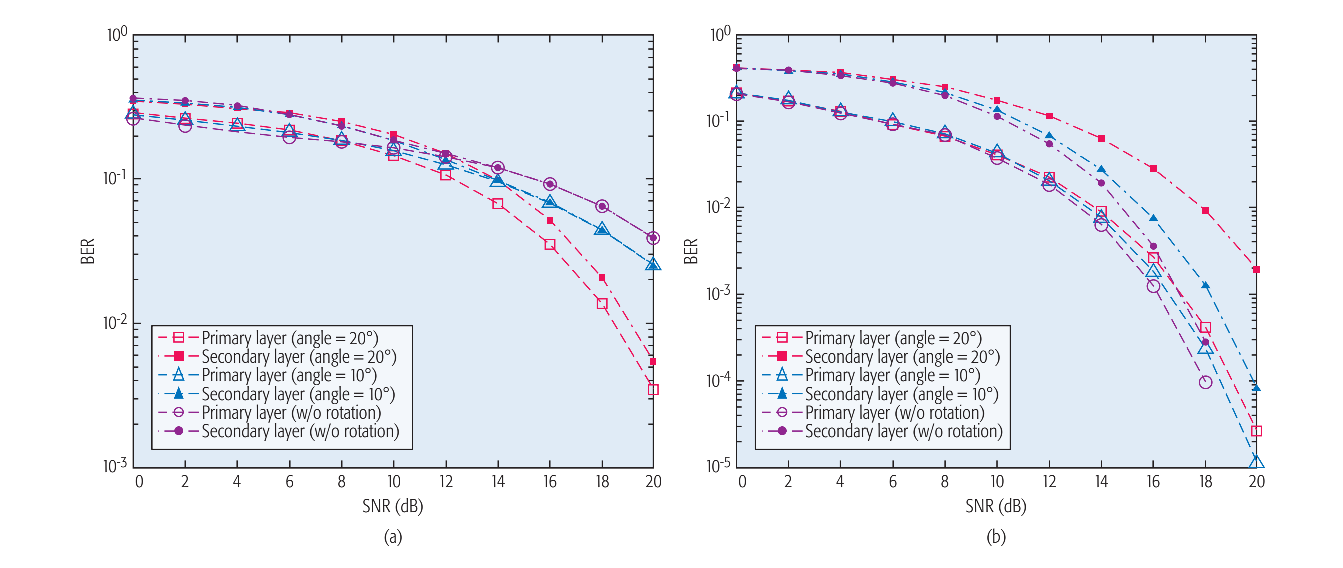
Multi-Rate Superposition Transmission: The scheduler in the MAC layer performs the following steps to achieve multi-user group scheduling and power allocation.

Transmission Power Allocation: The scheduler first performs power allocation and estimates instantaneous throughput of each user group, according to CSIs reported by all users in each user group. There are several available power allocation schemes, such as full search power allocation (FSPA), fractional transmit power allocation (FTPA), or fixed power allocation (FPA). Note that if no CSIs are available, the predefined power ratios are configured.

Multi-User Group Scheduling: The scheduler calculates the scheduling metric based on the estimated instantaneous throughput and selects one user group from the candidate user groups according to a certain scheduling algorithm, such as round-robin (RR), maximum throughput (MT), proportional fairness (PF), or max-min fairness.

Multi-Service Superposition Transmission:

- User group pairing: For multi-service superposition transmission, user group pairing is needed as media contents of multiple user groups are transmitted superposedly on the same channel. The scheduler chooses multiple user groups to form a user group pair, and obtains all candidate user group pairs according to a certain algorithm, such as exhaustive search or a CSI-based one.
- Transmission power allocation: The scheduler obtains the channel gain of each user group in candidate user group pairs, according to CSIs reported by all users in each user group. Then the scheduler performs power allocation for each user group pair to obtain optimal power ratio, and estimates the instantaneous throughput. Note that the predefined power ratios can be configured, if no CSI is obtained.



(a)

(b)

FIGURE 5. BER performance of superposition transmission with constellation rotation under AWGN channel: a) power ratio 0.6; b) power ratio 0.8.

- Multi-user group scheduling: The scheduler calculates the scheduling metric based on the estimated instantaneous throughput of each user group pair, and then selects one user group pair from candidate user group pairs according to certain scheduling algorithms.

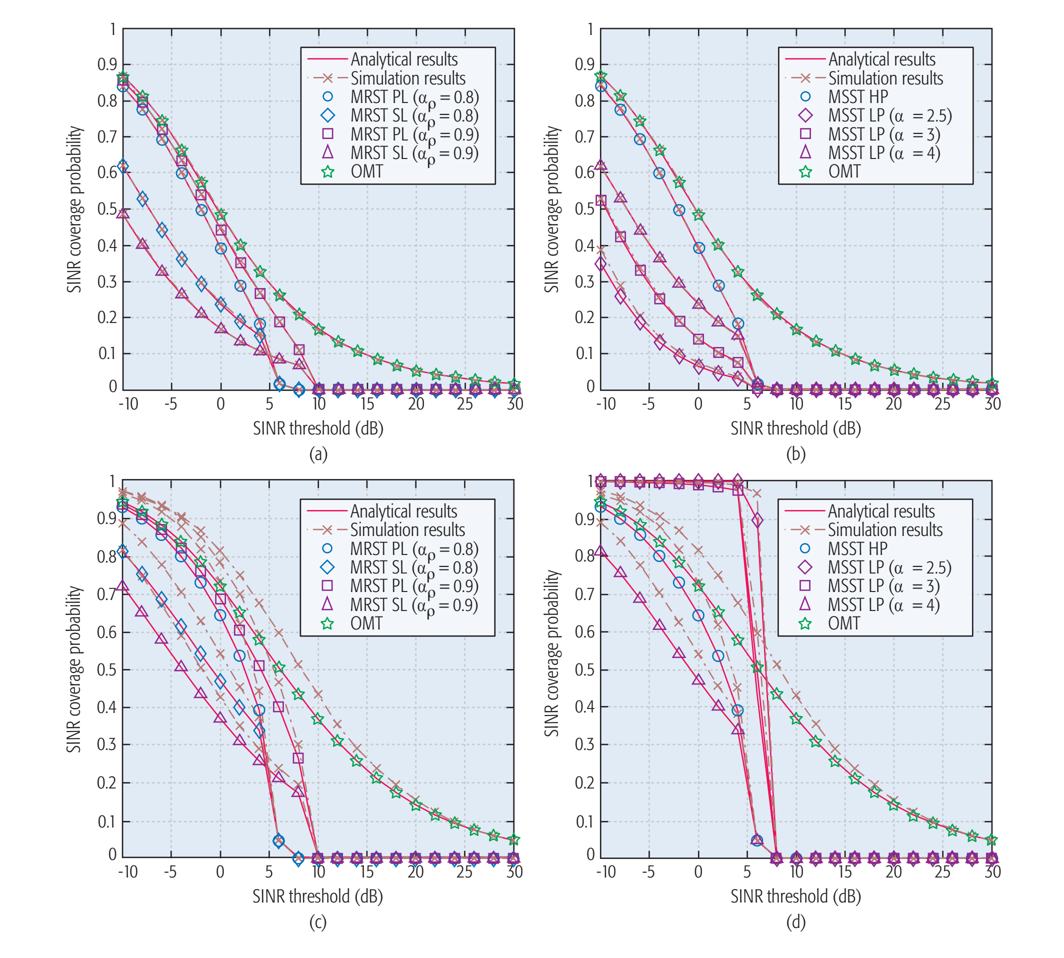
Performance Evaluation

In this section, we evaluate the BER performance of non-orthogonal transmission with constellation rotation and the downlink SINR coverage performance of non-orthogonal cellular mobile broadcasting in a large-scale two-tier single-frequency heterogeneous network (HetNet). The important parameters related to the downlink HetNet model are: the base densities of macro BSs and small cells are $\lambda_{B_1} = 1/(\pi 1000^2)$ and $\lambda_{B_2} = 1/(\pi 200^2)$, respectively, $P_{t,1} = 43$ dBm and $P_{t,2}$ =30 dBm for the corresponding transmit powers of macro BSs and small cells, 10 MHz for the system bandwidth, and 2 GHz for the carrier frequency. The power-law path loss propagation model with path loss exponent a is used to model signal attenuation with distance.

Figure 5 shows the BER performance of QPSK/QPSK superposition transmission with different constellation rotation angles and power ratios under the additive white Gaussian noise (AWGN) channel. We can observe that constellation rotation can obviously improve the BER performance of superposition transmission, because constellation rotation increases the minimum Euclidean distance of the composite constellation. However, the BER performance improvement declines with the increase of power ratio. We can also observe that the optimal degree of rotation angle is varied with the power ratio for superposition transmission, and reduces with the increase of power ratio. Besides, it can be observed that the BER performance of superposition transmission with power ratio $a_p = 0.8$ is better than the one with power ratio $\alpha_p = 0.6$. This is because with the increase of power ratio, the demodulation performance of the primary layer can be improved, which can reduce the negative effect of SIC error propagation on demodulating the secondary layer.

Figure 6 plots the SINR coverage probabilities of two-layer non-orthogonal cellular mobile broadcasting with SC-PTM and MC-PTM modes in a large-scale two-tier single-frequency HetNet. Note that the analytical results are from [14], and the results of MC-PTM mode are obtained approximately, which results in certain gaps between the analytical results and simulations.. The results show that for SC-PTM mode, multi-rate superposition transmission can provide one similar SINR coverage layer as the orthogonal one to transmit the basic data, as well as provide another coverage layer for enhanced data transmission to improve the QoS of strong users. With more power is allocated to the primary layer, its SINR coverage gradually approaches the orthogonal one, while the SINR coverage of the secondary layer decreases gradually. When all power is allocated to the primary layer, non-orthogonal transmission degrades to the orthogonal one. For multi-service superposition transmission, the services for high- and low-priority users (e.g., outdoor mobile users and indoor fixed TV users with rooftop antennas) are multiplexed in the power domain. Different path loss exponents are used to characterize different reception scenarios. The results also show that multi-service superposition transmission can provide similar SINR coverage for high-priority users as the orthogonal one, as well as a coverage layer for low-priority users simultaneously. However, the SINR coverage of the secondary layer with $\alpha = 2.5$ is lower than that of $\alpha = 3, 4$, as the users with smaller path loss exponents will suffer strong inter-cell interference in the SC-PTM mode. Finally, comparing MC-PTM with SC-PTM, we can observe that both MRST and MSST with MC-PTM mode can provide better SINR coverage than with SC-PTM mode. Besides, unlike SC-PTM mode, the SINR coverage of the secondary layer with $\alpha = 2.5$ is better than that of $\alpha = \{3, 4\}$. This is because MC-PTM mode enables all BSs in the broadcast area to deliver the same content on the same channel synchronously, such that users can receive and combine multiple signals from adjacent BSs to improve the quality of signal, rather than contribute to the aggregate interference.

Multi-user group scheduling and power allocation are two important functions of the MAC layer in non-orthogonal cellular mobile broadcasting systems, as it hosts radio resource allocation among active user groups in order to satisfy their QoS requirements. For SC-PTM, BSs perform multi-user group scheduling and power allocation, while it is done in MCE for MC-PTM.



(a) (b) (c) (d)

FIGURE 6. SINR coverage probabilities of non-orthogonal transmission-based cellular mobile broadcasting: a) multi-rate superposition transmission under SC-PTM mode; b) multi-service superposition transmission under SC-PTM mode; c) multirate superposition transmission under MC-PTM mode; d) multi-service superposition transmission under MC-PTM mode.

Research Challenges

Due to superposition transmission, non-orthogonal cellular mobile broadcasting shall encounter the following research challenges.

Hybrid Non-Orthogonal and Orthogonal Transmission: The performance of non-orthogonal transmission mainly relies on the channel gain differences between users. For a case in which users are distributed in certain local areas and have similar channel conditions, orthogonal transmission will be a better choice, considering the trade-off between performance and processing complexity. Consequently, it is important to study hybrid non-orthogonal and orthogonal transmission schemes.

MIMO for Non-Orthogonal Cellular Mobile Broadcasting Systems: MIMO technology [6] is a key enabler to increase system capacity and improve transmission robustness, as it can provide array gain, diversity gain, and multiplexing gain. The application of MIMO to NOMA systems was studied for the unicast case [3, 7]. However, cellular mobile broadcasting systems serve all interested users on the same channel, and generally work in unidirectional transmission without feedback. Therefore, how to exploit the multiplexing gain of MIMO systems to significantly improve the performance of non-orthogonal cellular mobile broadcasting is one great challenge.

Interference Management in Non-Orthogonal Cellular Mobile Broadcasting Systems: Inter-cell interference is one of the main issues in multi-cell NOMA networks. The combination of inter-cell interference management techniques and NOMA, for example, NOMA-joint transmission (NOMA-JT), NOMA-dynamic cell selection (NOMA-DCS), NOMA-coordinated scheduling/beamforming (NOMA-CS/CB), was studied for the unicast case [4]. However, non-orthogonal cellular mobile broadcasting systems transmit the media content to interested users in the broadcast area on the same radio resources, rather than a NOMA user pair consisting of one strong user and one weak user. This inherent distinction results in the fact that the above-mentioned inter-cell interference management techniques for NOMA networks cannot be directly used in the broadcast case. Therefore, the study of inter-cell interference techniques with reasonable complexity and good performance is another great challenge for non-orthogonal cellular mobile broadcasting systems.

Conclusions

In this article, the application of power domain non-orthogonal transmission to cellular mobile broadcasting is studied from transmission schemes to system designs. Two schemes for non-orthogonal cellular mobile broadcasting are presented and the system designs for network architecture and physical layer processing together with key elements are discussed. Some conclusions can be drawn as follows:

- Constellation rotation can dramatically improve the BER performance of non-orthogonal transmission. However, the rotation angle and performance improvements are related to power ratio and reduce with the increase of power ratio.
- Non-orthogonal transmission can achieve superior performance to the orthogonal one by improving degree of freedom rather than SINR.
- Non-orthogonal cellular mobile broadcasting with MC-PTM mode can achieve better coverage performance than the SC-PTM one.

Therefore, non-orthogonal cellular mobile broadcasting will play an important role in future multimedia wireless networks.

Acknowledgment

This work was supported by the NSFC project under Grants 61571373, U1734209, and U1709219, the NSFC China-Swedish Project under Grant 6161101297, the Key International Cooperation Project of Sichuan Province under Grant 2017HH0002, and the 111 Project under Grant 111-2-14. The work of Z. Ma was supported in part by the EU Marie Sklodowska-Curie individual Fellowship (No. 796426). The work of X. Lei was supported in part by the Sichuan International Science and Technology Cooperation Project under Grant 2017HH0035, in part by the NSFC project under Grant 61501382, and in part by the open research fund of the National Mobile Communications Research Laboratory, Southeast University, under Grant 2017D15. The work of C.-X. Wang was supported in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grants EP/P009670/1 and EP/L020009/1 and the EU FP7 QUICK project under Grant PIRSES-GA2013-612652.

References

[1] V. W. Wong *et al.*, *Key Technologies for 5G Wireless Systems*, Cambridge Univ. Press, 2017.

[2] L. Dai *et al.*, "Non-Orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities, and Future Research Trends," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 74–81.

[3] Z. Ding *et al.*, "Application of Non-Orthogonal Multiple Access in LTE and 5G Networks," *IEEE Commun. Mag.*, vol. 55, no. 2, Feb. 2017, pp. 185–91.

[4] W. Shin *et al.*, "Non-Orthogonal Multiple Access in Multi-Cell Networks: Theory, Performance, and Practical Challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, Oct. 2017, pp. 176–83.

[5] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge Univ. Press, 2005.

[6] B. Wang *et al.*, "Spectrum and Energy Efficient Beamspace MIMO-NOMA for Millimeter-Wave Communications Using Lens Antenna Array," *IEEE JSAC*, vol. 35, no. 10, Oct. 2017, pp. 2370–82.

[7] L. Zhang *et. al.*, "Layered-Division-Multiplexing: Theory and Practice," *IEEE Trans. Broadcast.*, vol. 62, no. 1, Mar. 2016, pp. 216–32.

[8] R. O. Afolabi, A. Dadlani, and K. Kim, "Multicast Scheduling and Resource Allocation Algorithms for OFDMA-Based Systems: A Survey," *IEEE Commun. Surveys. & Tutorials*, vol. 15, no. 1, 2013, pp. 240–54.

[9] A. D. L. Fuente, R. P. Leal, and A. G. Armada, "New Technologies and Trends for Next Generation Mobile Broadcasting Services," *IEEE Commun. Mag.*, vol. 54, no. 11, Nov. 2016, pp. 217–23.

[10] J. Kim *et al.*, "Group Communication over LTE: A Radio Access Perspective," *IEEE Commun. Mag.*, vol. 54, no. 4, Apr. 2016, pp. 16–23.

[11] G. Araniti *et al.*, "Multicasting over Emerging 5G Networks: Challenges and Perspectives," *IEEE Network*, vol. 31, no. 2, Mar./Apr. 2017, pp. 80–89.

[12] J. Choi, "Minimum Power Multicast Beamforming with Superposition Coding for Multiresolution Broadcast and Application to NOMA Systems," *IEEE Trans. Commun.*, vol. 63, no. 3, Mar. 2015, pp. 791–800.

[13] Z. Ding *et al.*, "On the Spectral Efficiency and Security Enhancements of NOMA Assisted Multicast-Unicast Streaming," *IEEE Trans. Commun.*, vol. 65, no. 7, July 2017, pp. 3151–63.

[14] Z. Zhang *et al.*, "Modeling and Analysis of Non-Orthogonal MBMS Transmission in Heterogeneous Networks," *IEEE JSAC*, vol. 35, no. 10, Oct. 2017, pp. 2221–37.

[15] A. Manzalini and N. Crespi, "An Edge Operating System Enabling Anything-as-a-Service," *IEEE Commun. Mag.*, vol. 54, no. 3, Mar. 2016, pp. 62–67.

Biographies

Zhengquan Zhang [S'16] (zhang.zhengquan@hotmail.com) received his M.Sc. degree in Communication and Information System from Southwest Jiaotong University, China, in 2008. From 2008 to 2013, he was with the ZTE Corp. as a communication protocol software engineer. Since 2014, he has been pursuing a Ph.D. degree at Southwest Jiaotong University. His research interests include millimeter-wave communications, non-orthogonal multiple access, cooperative communications, and full-duplex communications.

Zheng Ma [M'07] (zma@home.swjtu.edu.cn) is currently a professor at Southwest Jiaotong University, and serves as deputy dean of the School of Information Science and Technology. His research interests include information theory and coding, signal design and applications, FPGA/DSP Implementation, and professional mobile radio (PMR). He has published more than 60 research papers in high-quality journals and conferences. He is currently an Editor for *IEEE Communications Letters*. He is also the Chairman of the Communications Chapter of the IEEE Chengdu section.

Xianfu Lei [M'12] (xflei@home.swjtu.edu.cn) received his Ph.D. degree in communication and information systems from Southwest Jiaotong University in 2012. From 2012 to 2014, he was a research fellow with the Department of Electrical and Computer Engineering, Utah State University. Since 2015, he has been an associate professor with Southwest Jiaotong University. His research interests include 5G communications, cooperative communications, and energy harvesting. He currently serves as an Editor of *IEEE Communications Letters* and *IEEE Access*.

Ming Xiao [S'02, M'07, SM'12] (mingx@kth.se) received his Ph.D. degree from Chalmers University of Technology, Sweden, in 2007. Since 2007, he has been with the Department of Information Science and Engineering, School of Electrical Engineering, KTH Royal Institute of Technology, Sweden, where he is currently an associate professor. He received the best paper awards at the International Conference on Wireless Communications and Signal Processing in 2010 and the International Conference on Computer Communication Networks in 2011.

Cheng-Xiang Wang [S'01, M'05, SM'08, F'17] (cheng-xiang.wang@hw.ac.uk) received his Ph.D. degree from Aalborg University, Denmark, in 2004. He is now a professor at Heriot-Watt University, Edinburgh, United Kingdom. He has authored two books, one book chapter, and over 320 papers in refereed journals and conference proceedings. His current research interests include wireless channel modeling and (B)5G wireless communication networks. He is a Fellow of the IET and recognized as a Web of Science 2017 Highly Cited Researcher.

Pingzhi Fan [M'93-SM'99-F'15] (pzfan@home.swjtu.edu.cn) received his Ph.D. degree in electronic engineering fromHull University, United Kingdom. He is currently a professor and director of the Institute of Mobile Communications, Southwest Jiaotong University. His research interests include high mobility wireless communications, machine learning in wireless networks, signal design, and coding.

Cellular mobile broadcasting systems serve all interested users on the same channel, and generally works in unidirectional transmission without feedback. Therefore, how to exploit the multiplexing gain of MIMO systems to significantly improve the performance of non-orthogonal cellular mobile broadcasting is one great challenge.