Wei Xiang · Kan Zheng
Xuemin (Sherman) Shen *Editors*

# 5G Mobile Communications

Springer

# 5G Mobile Communications

Wei Xiang • Kan Zheng
Xuemin (Sherman) Shen
Editors

# 5G Mobile Communications

Springer

*Editors*
Wei Xiang
James Cook University
Cairns, QLD, Australia

Xuemin (Sherman) Shen
University of Waterloo
Waterloo, ON, Canada

Kan Zheng
Beijing University of Posts
and Telecommunications
Beijing, China

ISBN 978-3-319-34206-1 ISBN 978-3-319-34208-5 (eBook)
DOI 10.1007/978-3-319-34208-5

Library of Congress Control Number: 2016950231



Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

# Preface

Mobile communications have been instrumental in transforming our contemporary societies in the past decades. From the first-generation (1G) of analogue mobile phone system to the newest commercial fourth-generation (4G) long-term evolution (LTE) networks deployed widely across the global, mobile communications have fundamentally changed the ways as to how humans in the modern society access, exchange, and share information with each other. Currently, we are at an era of mobile Internet with explosive big data. The growing demand for mobile data traffic and the proliferation of applications requiring high data rates have raised a significant interest in the definition of new standards in the mobile market. This calls for next-generation mobile communication systems, which should have to respond to an avalanche of traffic, an explosion in the number of connected devices, and the large diversity of use cases and requirements.

Against the above backdrop, the fifth-generation (5G) mobile communications are fast emerging to tackle the challenges brought by an exponential increase in wireless data traffic. On top of the massive increase in data volumes and rates, a formidable challenge for the 5G networks to deal with is how to connect billions of smart devices such as surveillance cameras, smart-home/grid devices, connected sensors, etc. The primary goals of 5G networks are to support a 1000-fold gain in capacity, connections for at least 100 billion devices, and 10 Gb/s delivered to individual users. Furthermore, new 5G networks will be able to provide mass low-latency and ultrareliable connectivity between people, machines, and devices, which will ultimately usher in the era of the Internet of Things (IoT). To meet these enormous challenges, disruptive innovations and drastic improvements need to be made in the mobile network architecture design in both the physical and upper layers.

The International Telecommunications Union (ITU) has stipulated 2020 to be the target year of standardising future 5G mobile networks. Although the detailed technical approaches to implementing 5G mobile networks remain uncertain at the time of this writing, several breakthrough 5G techniques stand out such as massive MIMO and millimetre-wave (mmWave) communications. This book aims to be one of the first comprehensive books to reveal the enabling techniques underpinning

next-generation 5G networks and to address the challenges and opportunities brought by 5G mobile communications. Specifically, the book is divided into three major parts: Part I Overview of 5G Networks, Part II Transmission and Design Techniques for 5G Networks, and Part III Networking Techniques and Applications for 5G Networks.

Part I of this book provides a comprehensive introduction to and overview of 5G networks. It consists of three chapters.

The chapter "An Overview of 5G Requirements" presents an overview of next-generation 5G mobile networks. To facilitate the study of 5G requirements and to provide guidance to 5G technical design, this chapter discusses several typical deployment scenarios including indoor hotspot, dense urban, urban macro, rural, and high-speed scenarios. It also presents high-level key capabilities and detailed technical requirements for 5G networks. Some technical performance metrics of 5G networks are also discussed.

The chapter "Spectrum Analysis and Regulations for 5G" discusses various aspects of 5G spectrum issues. It is expected that 5G requires much more bandwidth as well as more flexibility in spectrum usage and management. The suitable frequency ranges of 5G will include those bands below 6 GHz such as re-farmed 2G/3G spectrum, identified frequency bands for IMT, and also WRC-15 candidate bands. However, due to the scarcity of spectrum below 6 GHz, it is imperative to seek potential frequency ranges above 6 GHz. Controlled spectrum sharing is an important way of reusing spectrum to complement current licensed dedicated spectrum, which is still the foundation for the operation of 5G systems.

The chapter "Spectrum Sharing for 5G" first introduces spectrum sharing for 5G systems, which consists of multiple spectrum types with different scenarios. Then, spectrum sharing techniques mapped into different scenarios are introduced, i.e. coordination protocol, GLDB support, cognitive/DSA, and MAC-based coexistence. Besides, current applications of these techniques in real systems are described. Finally, spectrum sharing directions for 5G systems are analysed for different spectrum sharing techniques. It is concluded that licensed dedicated spectrum will continue to be the dominant spectrum usage method for 5G systems due to the possibility to control interference and guarantee coverage, while other spectrum sharing scenarios will act as complementary spectrum usage methods when beneficial.

Part II of this book presents new transmission and design techniques for 5G networks with a focus on physical-layer enabling techniques. It contains 11 chapters.

The chapter "Massive MIMO Communications" argues that every new network generation needs to make a leap in area data throughput, to manage the growing wireless data traffic. Massive MIMO technology can bring at least tenfold improvements in area throughput by increasing the spectral efficiency (bit/s/Hz/cell) while using the same bandwidth and density of base stations as in current networks. These extraordinary gains are achieved by equipping the base stations with arrays of a hundred antennas to enable spatial multiplexing of tens of user terminals.

This chapter explains the basic motivations and communication theory behind the massive MIMO technology and provides implementation-related design guidelines.

The chapter "Millimeter-Wave Mobile Communications" introduces key technologies of mmWave communications. Channel measurements show that mmWave signals suffer from much larger propagation losses and are suitable for small cell coverage. A hybrid network is presented where mmWave is used for capacity enhancement in hotspots, and a low-frequency network is applied for seamless coverage. Uniform air interface is a consequence to simplify the design between mmWave bands and low-frequency bands. Unified access and backhaul technique not only reduces the cost of backhaul but also can meet the requirement of 1000 times capacity enhancement over LTE systems.

The chapter "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access" introduces state-of-the-art NOMA techniques and evaluates the low density spreading (LDS)-based system, which is a strong candidate for the next generation of mobile networks due to its well-known advantages compared to state-of-the-art techniques based on orthogonal frequency division multiple access (OFDMA). Furthermore, the effect of LDS parameters such as density factor and maximum number of users at each time instance on the sum rate is evaluated. The effect of irregularity on the complexity is also discussed. Moreover, it is shown that the loss of achievable rates which is caused by modulation can be compensated by using a suitable channel coding scheme.

The chapter "New Multicarrier Modulations for 5G" presents recent advances in filter bank multicarrier (FBMC) techniques and compares them with the conventional cyclic prefix (CP)-OFDM approach, in the context of 5G. After a brief description of some adaptations of CP-OFDM, FBMC combined with offset-QAM is considered, pointing out the crucial issue of subchannel equalisation. Then, an alternative approach is proposed, FBMC combined with pulse amplitude modulation (PAM). FBMC-PAM is an attractive option whenever asynchronous access and high level of out-of-band rejection are required. Finally, the case of nonoverlapping emitted symbols is considered, and a CP-less OFDM scheme with frequency domain equaliser in the receiver is included in the performance comparison.

The chapter "Fundamentals of Faster-than-Nyquist Signaling" presents the fundamentals of Faster-than-Nyquist (FTN) signalling. As originally introduced, FTN increases the bit rate in the signalling bandwidth by packing symbols closer in time, at the cost of introducing intersymbol interference (ISI). The chapter begins with the Euclidean distance properties of bandwidth-efficient pulses at FTN rates and describes receivers that mitigate the severe ISI. The FTN achievable information rate is compared with the Nyquist information rate for practical pulses. It then discusses the FTN extension to multicarrier systems with not only time packing but also subcarrier, optimising both the time and frequency packing.

The chapter "Generalized Frequency Division Multiplexing: A Flexible Multi-Carrier Waveform for 5G" aims to develop a unified air interface that can be configured on-the-fly to address emerging 5G applications. Apart from an ever-increasing demand for data rates, 5G is facing new applications such as Tactile

Internet and the Internet of Things. Being aligned with the whole concept of software-defined networking, this chapter introduces the multicarrier waveform termed generalised frequency division multiplexing (GFDM) as the basis for realising such a flexible physical design.

The chapter "Spectrally Efficient Frequency Division Multiplexing for 5G" focuses on novel multicarrier communication techniques, which share the common goal of increasing spectrum efficiency in future communication systems. In particular, a technology termed spectrally efficient frequency division multiplexing (SEFDM) is described in detail outlining its benefits, challenges, and trade-offs when compared to the current state of the art. A decade of research has been devoted to examining SEFDM from different angles: mathematical modelling, algorithm optimisation, hardware implementation, and system experimentation. The aim of this chapter is to therefore give a taste of this technology, and in doing so, the chapter concludes by outlining a number of experimental test beds which have been developed for the purpose of evaluating the performance of SEFDM in practical scenarios.

The chapter "Full-Duplex Wireless Communications for 5G" introduces full-duplex (FD) wireless communications for 5G, which enables simultaneous transmission and reception over the same frequency band. In this way, the spectral efficiency can be improved significantly compared with half-duplex (HD). However, there exists severe self-interference (SI), signal leakage from the local transmitter to its own receiver. Three different classes of SI mitigation techniques are presented in this chapter, i.e. propagation-domain SI suppression, analogue-domain SI cancellation, and digital-domain SI cancellation. Furthermore, the system performance of several FD schemes in several different application scenarios is presented. Theoretically, the spectral efficiency of FD bidirectional and cooperative communications can be doubled, while for cognitive radio networks, the FD-based protocol can achieve much better sensing performance than the traditional HD-based cognitive radio schemes.

The chapter "Device-to-Device Communications over 5G Systems: Standardization, Challenges and Open Issues" introduces one of the key enabling technologies at the heart of the 5G systems, namely, device-to-device (D2D) communications. The potential of D2D communication paradigm holding the promise to overcome the limitations of conventional cellular systems with very high bit rates, low delay, and low power consumption is illustrated. Starting from an overview of D2D communication technology, this chapter will browse through the main aspects that characterise the proximity services, with a view on the standardisation process, the challenges, and the open issues.

The chapter "M2M Communications in 5G" provides an overview of machine-type communications (MTC) within the context of 5G networks. The Internet of Everything foresees a hyperconnected World where humans, things, and machines will need to coexist together. They will be interconnected and Internet-connected via communication networks. In specific, the authors review the key novel challenges of MTC: what is new with regard to human-type traffic (HTC). They then analyse existing communication technologies and how suitable they are for MTC. Finally,

the authors identify key technology enablers being considered for the design of 5G networks and provide an outlook for the future.

The chapter "Design Techniques of 5G Mobile Devices in the Dark Silicon Era" is concerned with the design of the prospected 5G mobile communication system, which needs wide skills in wireless communication, analogue circuit design, embedded system, microwave technology, and so forth. System-level analyses, design space exploration, and performance trade-offs are some key steps that enable the design of low-cost, energy-efficient, ubiquitous, and flexible transceiver. This chapter provides comprehensive design techniques for 5G mobile communication in the dark silicon era using More than Moore technology (MtM).

Part III of this book focuses primarily on the networking and application layer techniques for 5G networks, which includes 12 chapters.

The chapter "Ultra-Dense Network Architecture and Technologies for 5G" presents the ultra-density network (UDN), which is the most promising way to meet the ultrahigh area capacity requirement for 5G. The content of this chapter includes characters of UDN scenarios, network architecture design, and key technologies like flexible networking, wireless backhauling, multi-RAT coordination, mobility management, interference management, and radio resource management.

The chapter "5G RAN Architecture: C-RAN with NGFI" describes cloud radio access networks (C-RAN), which are viewed as one of the key RAN architectures for 5G networks, with evolved architecture based on a newly designed fronthaul interface, dubbed the next-generation fronthaul interface (NGFI). The design principles and the challenges of NGFI are introduced. A prototype is further developed to verify the applicability of NGFI-based C-RAN.

The chapter "User-Centric Wireless Network for 5G" addresses the concept of user-centric wireless network for 5G from the perspective of fulfilling multiple user experience requirements in 5G. Four key technical directions are studied based on a gap analysis between LTE technology and 5G requirements, i.e. user-centric 5G access network architecture design, flexible functionality and deployment, smart user and traffic awareness and management, and high-efficient low-cost network operation. These key technologies work together with cross-layer and end-to-end solutions to provide the user-centric 5G ecosystem.

The chapter "Energy Harvesting Based Green Heterogeneous Wireless Access for 5G" is concerned with the issues of energy harvesting for future 5G cellular systems. A feasible and efficient method to tackle this issue is to let the communication systems harvest energy from renewable energy sources instead of fossil fuels. However, by employing the energy harvesting (EH) technique, the instability of renewable energy resources introduces new challenges on the design of the upcoming 5G systems. This chapter focuses on uplink access schemes and power allocations for EH-based heterogeneous networks. First, a heterogeneous access model incorporating EH-based mobile users is proposed and followed by a throughput maximisation framework. Then, by classifying transmission policies into two main categories (i.e. single-channel vs. multichannel scenarios), the proposed framework is concretised under various practical conditions, including

the availability of central control, causality of harvested energy, channel state information, and others.

The chapter "Resource Management in Sustainable Green HetNets with Renewable Energy Sources" investigates the energy sustainability performance of a green HetNet where the small cell base stations (SBSs) are powered by green energy. Specifically, we first develop an analytical framework to study the energy sustainability of each SBS. Then, we propose a distributed admission control strategy at SBSs striking a balance between resource utilisation and energy sustainability. Extensive simulations validate the analytical framework and demonstrate that relaxing the admission control criteria can improve resource utilisation when the energy is abundant, but may significantly degrade resource utilisation instead when the energy comes short due to poor sustainability performance.

The chapter "Resource Allocation for Cooperative D2D Communication Networks" studies various resource allocation policies for cooperative device-to-device (D2D) communications in systems operating under OFDMA or cognitive radio architectures. A variety of system models are explored, wherein additional features, such as, packet storage, energy harvesting, and cognitive radio capabilities, are incorporated at the user devices so as to enable cooperative D2D communications. Computationally efficient solutions are provided for multiple resource optimisation problems including power allocation, subcarrier allocation, subcarrier pairing, and relay selection. Simulation results demonstrate that the sum-throughput performance can be improved whenever the user devices are equipped with cooperative D2D capabilities.

The chapter "Fog Computing and Its Applications in 5G" explains the emergence of fog computing as a promising, practical, and efficient solution tailored to serving mobile traffics. Fog computing deploys highly virtualised computing and communication facilities at the proximity of mobile users. Dedicated to serving the mobile users, fog computing explores the predictable service demand patterns of mobile users and typically provides desirable localised services accordingly. It can provide mobile users with the demanded services via low-latency and short-distance local connections. The authors introduce the main features of fog computing and describe its concept, architecture, and design goals. Lastly, they discuss the potential research issues from the perspective of 5G networking.

The chapter "A Conceptual 5G Vehicular Networking Architecture" shows how 5G communication systems will help to enable connected future cars to implement automated functions in short term and fully autonomous operation in long term. The authors review the well-known existing communication technologies for connected cars and analyse their shortcomings. Towards this end, they outline the innovation areas that 5G aims to address in order to mitigate the limitations of the current technologies.

The chapter "Communications Protocol Design for 5G Vehicular Networks" provides an overview on existing standards in vehicular networking and highlights new emerging trends towards an integrated infrastructure based on the interworking of heterogeneous technologies. Next-generation mobile vehicular networks are first characterised by providing an insight on relevant stable standards in wireless

communication technologies, with a special focus on heterogeneous vehicular networks. Furthermore, the chapter discusses a general framework supporting opportunistic networking scheme and outlines novel application and use cases based on social- and context-awareness paradigms.

The chapter "Next-Generation High-Efficiency WLAN" centres around the topic of next-generation high-efficiency WLAN technology. With the increasing demands for WLAN and the deployment of carrier-WiFi networks, the number of WiFi public hotspots worldwide is expected to increase dramatically. To face this huge increase in the number of densely deployed WiFi networks, and the massive amount of data to be supported by these networks in indoor and outdoor environments, it is necessary to improve the current WiFi standard and define specifications for high-efficiency wireless local area networks (HEWs). This chapter introduces emerging HEW technology, including its typical use cases, environments, and potential techniques that can be applied for HEWs. The typical HEW use cases are first given, followed by an analysis of the main requirements from these use cases and environments. Then, potential techniques, including enhanced medium access and spatial frequency reuse, are presented and discussed.

The chapter "Shaping 5G for the Tactile Internet" investigates the topic of the Tactile Internet, which is expected to have a massive impact on business and society. It has the potential to revolutionise almost every segment of the society by enabling wireless control and remote operation in a range of scenarios. The next-generation (5G) mobile communication networks will play an important role in realising the Tactile Internet. This chapter investigates the interesting area of 5G and Tactile Internet intersection. Key requirements for the Tactile Internet, from a networking perspective, have been identified, after introducing exciting Tactile Internet applications. The chapter covers several technical issues and challenges in shaping 5G networks for realising the vision of the Tactile Internet. The most important challenge would be to ensure tight and scalable integration of various technological solutions into a single network.

We would like to thank all the authors who submitted their research work to this book. We would also like to acknowledge the contribution of many experts who have participated in the review process and offered comments and suggestions to the authors to improve their works. We would also like to express our sincere appreciation to the editors from Springer for their support and assistance during the development of this book.

Cairns, QLD, Australia — Wei Xiang
Beijing, China — Kan Zheng
Waterloo, ON, Canada — Xuemin (Sherman) Shen

# Contents

# Part I
# Overview of 5G Networks

# An Overview of 5G Requirements

**Dajie Jiang and Guangyi Liu**

**Abstract** Mobile Internet and IoT (Internet of Things) are the two main market drivers for 5G. There will be a massive number of use cases for Mobile Internet and IoT, such as augmented reality, virtual reality, remote computing, eHealth services, automotive driving and so on. All these use cases can be grouped into three usage scenarios, i.e., eMBB (Enhanced mobile broadband), mMTC (Massive machine type communications) and URLLC (Ultra-reliable and low latency communications). Eight key capabilities including peak data rate, latency and connection density, etc., are defined to meet the requirements of usage scenarios. Based on the usage scenarios, several typical deployment scenarios including indoor hotspots, dense urban, urban macro, rural and high-speed scenarios are specified, together with the detailed technical requirements for 5G. Both the deployment scenarios and technical requirements are essential guidance for 5G technical design.

## 1 Introduction

A mobile and connected society is emerging in the near future, which is characterized by a tremendous amount of growth in connectivity, traffic volume and a much broader range of use scenarios [1]. Some typical trends are summarized as follows:

- Explosive growth of data traffic: There will be an explosive growth in traffic. The global data traffic will increase by more than 200 times from 2010 to 2020, and about 20,000 times from 2010 to 2030;
- Great increase in connected devices: While smart phones are expected to remain as the main personal devices, the number of other kinds of devices, including wearable devices and MTC devices will continue to increase;
- Continuous emergence of new services: Different kinds of services, e.g., services from enterprises, from vertical industries and Internet companies, etc. will be exploited.

D. Jiang (✉) • G. Liu
China Mobile Research Institute, Beijing, China
e-mail: jiangdajie@chinamobile.com



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_1

The fifth-generation (5G) mobile communications system will emerge to meet new and unprecedented demands beyond the capability of previous generations of systems.

There are two phases of 5G requirements research by different organizations. Phase 1 focuses on 5G use cases and high-level key capabilities of 5G networks, and can be regarded as the 5G vision stage. In Phase 1, ITU has released the vision recommendation [2] and defined the key capabilities of 5G. 3GPP started the smarter program [4] and studied 5G use cases and requirements. NGMN completed a 5G whitepaper and defined a large number of 5G use cases and requirements [5]. IMT-2020 (5G) Promotion Group released the 5G vision and requirements whitepaper in May 2014 [1], which aims to contribute to the ITU-R work in Phase 1. Phase 2 focuses on 5G deployment scenarios and detailed technical requirements. There are two important reports in Phase 2. One is the IMT-2020 technical performance requirements from ITU-R which will be completed by February 2017 [3], while the other is the scenarios and requirements technical report from 3GPP which will be completed in March 2017 [6]. NGMN has started the relevant work at the beginning of 2015 and drafted several liaisons to 3GPP and ITU by March 2016 [7, 8]. IMT-2020 (5G) PG plans to complete the evaluation scenarios and the KPI report in the first half of 2016, and will have an impact on the work of ITU and 3GPP in Phase 2 (Fig. 1).

The rest of this chapter is organized as follows. The outcomes of Phase 1, i.e., 5G use cases and high-level key capabilities are introduced in Sects. 2 and 3, respectively. The latest status of Phase 2 including deployment scenarios

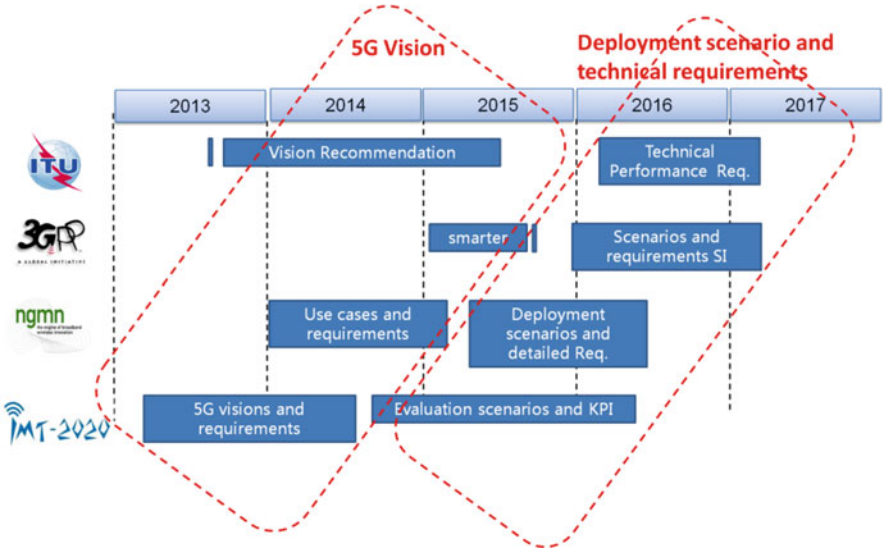


**Fig. 1** Overview of 5G requirements research by different organizations

and detailed technical requirements are presented in Sects. 4 and 5, respectively. Section 6 presents the operational requirements, while Sect. 7 draws concluding remarks.

## 2 Use Cases and Challenges

It is foreseen that there will be a huge number of use cases in the upcoming 5G era. In [1], an overall vision for 5G life is illustrated in Fig. 2. 5G will penetrate into every single element of our future society and create an all-dimensional, user-centered information ecosystem. 5G will break the limitation of time and space to enable an immersive and interactive user experience. 5G will also shorten the distance between human and things, and implement a seamless integration to achieve easy and smart interconnection between people and all things. 5G will enable us to realize the vision—"Information is a finger away, and everything will be kept in touch".

There are a great many use cases proposed by different organizations [1, 9–11]. Mobile Internet and the Internet of Things (IoT) are the two main market drivers in the future development of mobile communications [1], and they will trigger a large range of use cases.


**Fig. 2** Overall vision of 5G

## 2.1 Use Cases of Mobile Internet

Mobile Internet is disrupting the traditional business model of mobile communications, enabling unprecedented user experiences and making a profound impact on every aspect of people's work and life. Looking ahead to 2020 and beyond, mobile Internet will promote the continued evolution of the way human interacting information, and provide users with ultimate experience through more immersive services including but not limited to:

- Video services, such as immersive Ultra High Definition (UHD) and three-dimensional (3D) video
- Augmented reality
- Virtual reality
- Video/photo sharing in stadium/open air gathering
- Online gaming applications
- Mobile cloud/desktop cloud
- Tactile Internet
- Remote computing
- 3D connectivity: aircraft
- 3D connectivity: drones
- Collaborative robots
- Broadcast-like services, like local, regional and national news and information
- Smart office

The future development of Mobile Internet will trigger the growth of mobile traffic by a magnitude of thousands in the future, and promote a new wave of upgrades and a revolution in mobile communications technologies and the telecommunications industry as a whole. Looking ahead to 2020 and beyond, there will be an explosive growth in mobile data traffic. It is estimated by IMT-2020 (5G) Promotion Group that the global mobile data traffic will grow by more than 200 times from 2010 to 2020, and by nearly 20,000 times from 2010 to 2030. In China, the growth rate is projected to be even higher, with mobile data traffic expected to grow by more than 300 times from 2010 to 2020 and by more than 40,000 times from 2010 to 2030. For developed cities and hotspots in China, the growth of mobile data traffic will exceed the average projected. For example, from 2010 to 2020 in Shanghai, the mobile data traffic is projected to grow by 600 times. In Beijing and during this same period, it is estimated that hotspot traffic may grow by up to 1000 times. The above estimation is shown in Fig. 3.

There is some traffic anticipation work by ITU-R and the results are detailed in Report ITU-R M.[IMT.2020BEYOND TRAFFIC] [12]. This report contains global IMT traffic estimates beyond 2020 from several sources. These estimates anticipate that global IMT traffic will grow by 10–100 times from 2020 to 2030. The main drivers behind the anticipated traffic growth include increased video usage, device proliferation and application uptake. These are expected to evolve over time, and this evolution will differ between countries due to social and economic differences.

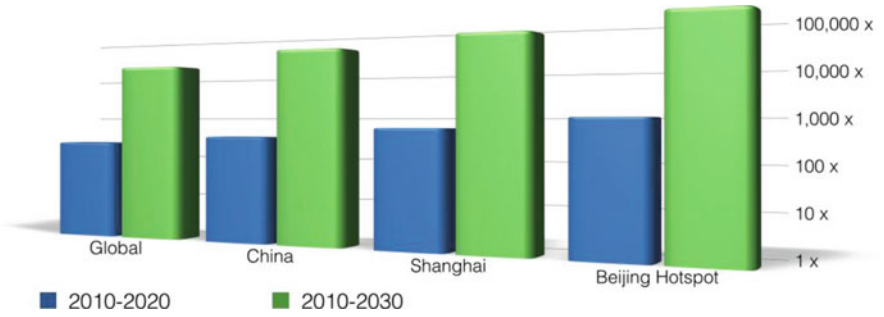


**Fig. 3** 2010–2030 growth of mobile data traffic

Traffic asymmetry aspects for this period are also presented by ITU-R. It is observed that the current average traffic asymmetry ratio of mobile broadband is in favor of the downlink, and this is expected to increase due to a growing demand for audio-visual content.

Mobile Internet is aiming at people-oriented communications with a focus on the user experience. Towards 2020 and beyond, the increasing popularity of ultra-high definition (UHD), 3D and video immersion will significantly drive up the data rates. For example, with a hundredfold compression, the transmission of 8K (3D) video will require a transmission rate close to 1 Gbps. Services, such as augmented reality, desktop cloud, and online gaming will not only pose a challenge to uplink and downlink data transmission rates but also generate a stringent requirement for the so-called "imperceptible latency". In the future, vast amounts of individuals and office data will be stored in the cloud. Such massive data activity will require transmission rates to be comparable to optical fiber communications, which will lead to enormous traffic challenges for mobile communications networks particularly in hotspot areas. Over-the-top (OTT) services, such as social networking, will be counted among leading applications going forward, and the associated frequently-occurring small packets will devour signaling resources. At the same time, consumers will continue to demand better experiences on mobile communications wherever they are. A consistent service experience is expected in all scenarios, including ultra-dense scenarios such as stadiums, open-air gatherings and concerts, and high-speed moving scenarios such as high-speed trains, vehicles and subways.

The total number of devices connected by global mobile communications networks will reach 100 billion in the future. By 2020, it is predicted that the number of mobile terminals around the world will surpass ten billion, of which China will contribute over two billion, as shown in Fig. 4.

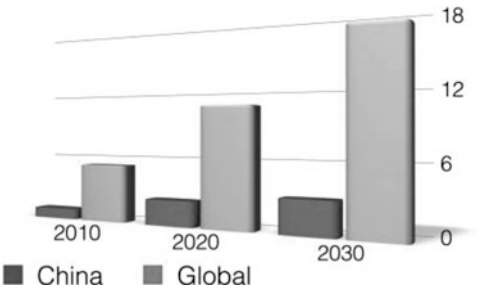


**Fig. 4** 2010–2030 growth of mobile device

## *2.2 Use Cases of Internet of Things*

The IoT has extended the scope of mobile communications services from interpersonal communications to interconnection between tings (smart devices), and between people and things, allowing mobile communications technologies to penetrate into broader industries and fields. By 2020 and beyond, applications such as mobile health, Internet of Vehicles (IoV), smart home, industrial control, and environmental monitoring will drive the explosive growth of IoT applications, facilitating hundreds of billions of devices to connect to a network creating a true "Internet of Everything". This will give rise to emerging industries of a unprecedented scale and instill infinite vitality to mobile communications. Meanwhile, the massive number of interconnected devices and the diversified IoT services will also pose new challenges to mobile communications. The potential IoT use cases include:

- Smart Grid and critical infrastructure monitoring
- Environmental monitoring
- Smart agriculture
- Smart metering
- eHealth services
- Remote object manipulation like remote surgery
- Automotive driving/Internet of vehicles
- Smart wearables, like sports and fitness
- Sensor networks
- Mobile video surveillance
- Smart cities
- Smart transportation
- Smart home
- Industrial control

IMT-2020 (5G) Promotion Group has estimated the number of IoT devices in future years as shown in Fig. 5. It is projected that the total number of devices connected by the global mobile communications network will reach 100 billion in the future. By 2020, the number of mobile terminals around the world will surpass ten billion, of which China will contribute over two billion. The number of IoT

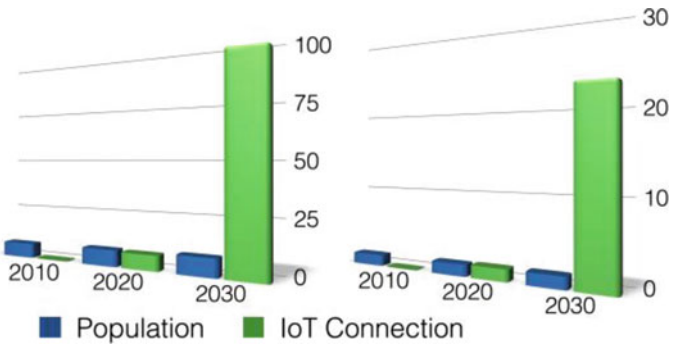


**Fig. 5** 2010–2030 growth of IoT connections

connections will also expand rapidly, reaching the size of global population of seven billion by 2020. By 2030, the number of global IoT connections will reach 100 billion. Among all types of terminals, smart phones will contribute most traffic, and IoT terminals will contribute less, even though the number of devices is much larger.

IoT is focused on communications between things and between things and people, involving not only individual users, but also a large number of various vertical industrial customers. IoT services types and relevant requirements of IoT services are very diverse. For services such as the smart home, smart grid, environmental monitoring, smart agriculture, and smart metering, the network will be required to support a massive amount of device connections and frequently-occurring small data packets. Services like video surveillance and mobile health will have strict requirements on transmission rates, while services such as IOV and industrial control will demand millisecond-level latency and nearly 100 % reliability. In addition, many IoT devices may be deployed in remote, or in areas where transmission losses can be a problem, such as indoor corners, basements and tunnels. Therefore, the coverage of mobile communications networks need to be further enhanced. In order to penetrate into more IoT services, 5G should be more flexible and more scalable, to support massive device connections and meet diverse user requirements.

Users expect better and yet more cost-effective services and experiences with mobile Internet and IoT. In addition to satisfying cost and experience demands, 5G will also need to meet extremely high security requirements, particularly for services such as e-banking, security monitoring, safe driving, and mobile health. 5G will also be able to support lower power consumption to build greener mobile communication networks and to enable much longer terminal battery life, especially for some IoT devices.

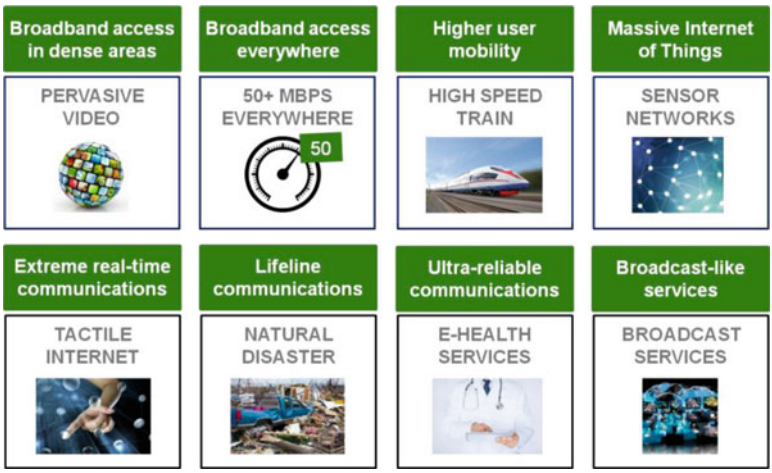


**Fig. 6** 5G use case families and related use case examples

## *2.3 Classification of 5G Use Cases*

5G will support a large variety of use cases which are emerging now or will emerge in the future. Different use cases have varying characteristic and requirements. It is helpful to group countless emerging use cases into several use case families. Use cases in each use case family share similar characteristic and requirements.

NGMN has developed 25 use cases for 5G as representative examples, which are grouped into eight use case families. The following diagram [5] illustrates the eight use case families with one example use case given for each family, and these families and their corresponding use case examples are described in Fig. 6.

ITU-R has concluded three usage scenarios (use case groups) addressing different use case characteristics in Fig. 7 [2]:

– Enhanced mobile broadband: Mobile broadband addresses human-centric use cases for access to multi-media content, services and data. The demand for mobile broadband will continue to increase, leading to enhanced mobile broadband. The enhanced mobile broadband usage scenario will come with new application areas and requirements in addition to existing mobile broadband applications for improved performance and an increasingly seamless user experience.
– Ultra-reliable and low latency communications: This use case has stringent requirements for capabilities such as throughput, latency and availability. Some examples include wireless control of industrial manufacturing or production processes, remote medical surgery, distribution automation in a smart grid, transportation safety, etc.

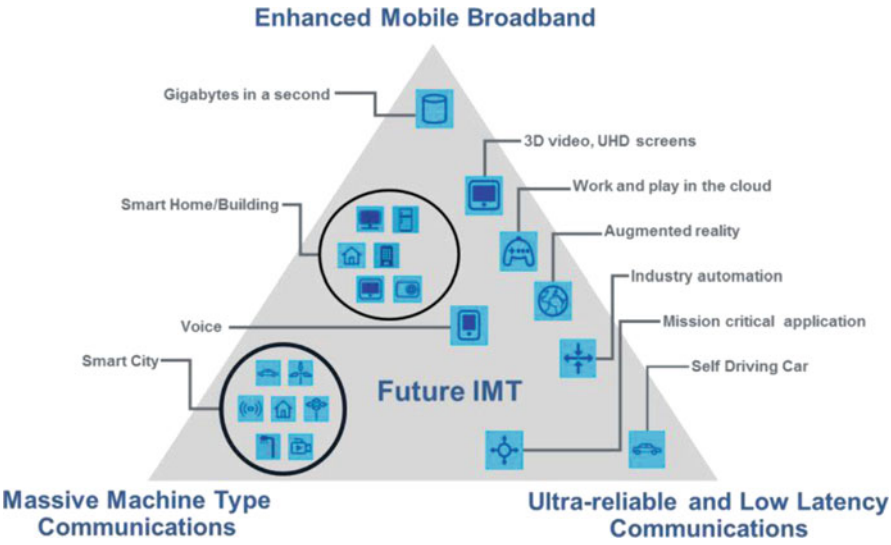


**Fig. 7** 5G usage scenarios

– Massive machine type communications: This use case is characterized by a very large number of connected devices typically transmitting a relatively low volume of non-delay-sensitive data. Devices are required to be of low cost, and have a very long battery life.

Similarly, IMT-2020 (5G) Promotion Group has proposed four technical scenarios for 5G in Fig. 8 which are well in line with the three usage scenarios from ITU-R. The main difference is that the eMBB scenario from ITU-R is divided into two technical scenarios, i.e., the seamless wide-area coverage scenario and high-capacity hot-spot scenario.

For the seamless wide-area coverage scenario, seamless coverage and medium to high mobility are desired, with much improved user data rate compared to existing data rates. However, the data rate requirement may be relaxed compared to the hotspot scenario. For the high-capacity hot-spot scenario, i.e., for an area with high user density, very high traffic capacity is needed, while the requirement for mobility is low and the user data rate is higher than that of wide area coverage. For the wide area coverage case, seamless coverage and medium to high mobility are desired, with much improved user data rate compared to existing data rates. However, the data rate requirement may be relaxed compared with the hotspot.

Figure 9 shows the mapping from the eight use case families proposed by NGMN to the three usage scenarios defined by ITU-R. The eMBB usage scenario consists of broadband access in dense areas, broadband access everywhere, higher user mobility and broadcast-like services. URLLC consists of extreme real-time communications, lifeline communications and ultra-reliable communications. mMTC corresponds to massive Internet of Things.


**Fig. 8** Technical scenarios for 5G

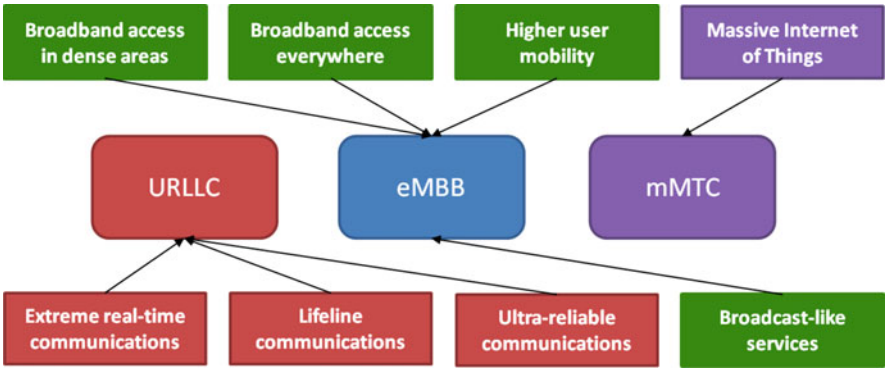


**Fig. 9** Mapping eight use case families to three use scenarios

## 3 High-Level Key Capabilities

IMT-2020 (5G) PG groups the 5G high-level requirements into several performance indicators and efficiency indicators. The key performance indicators for 5G include the user experienced data rate, connection density, end-to-end delay, traffic volume density, mobility, and peak date rate. Their definitions are listed in Table 1. The value for each performance indicator and the relevant scenario are illustrated in Fig. 10.

**Table 1** 5G Performance indicators

| Performance indicators | Definition |
|---|---|
| User experienced data rate (bps) | Minimum achievable data rate for a user in real network environment |
| Connection density (/km$^2$) | Total number of connected devices per unit area |
| End-to-end latency (ms) | Duration between the transmission of a data packet from the source node and the successful reception at destination node |
| Traffic volume density (bps/km$^2$) | Total data rate of all users per unit area |
| Mobility (km/h) | Relative speed between receiver and transmitter under certain performance requirement |
| Peak date rate (bps) | Maximum achievable data rate per user |

Several problems are anticipated if today's networks are used to handle the explosive development of mobile Internet and IoT:

- The energy efficiency level, overall cost per bit and complexity of network deployment and maintenance cannot effectively handle 1000 times traffic growth and the massive number of connected devices in the next decade;
- Co-existence of multiple radio access technologies (RAT) causes increased complexity and degraded user experience;
- Existing networks can not realize accurate monitoring of network resources and effective awareness of services, and therefore they cannot intelligently fulfill the diversified requirements of future users and services;
- Widely distributed and fragmented spectrum will cause interference and co-existence complexity.

To solve these problems, 5G should have the following capabilities to achieve sustainability. In terms of network construction and deployment, 5G networks need to:

- Provide higher network capacity and better coverage, while decreasing the complexity and cost of network deployment, especially the deployment of ultra-dense networks.
- Have a flexible and scalable architecture to adapt to the diverse needs of users and services.
- Make flexible and efficient use of various spectrum resources, including paired and unpaired spectrum, re-farmed spectrum and new spectrum, low-frequency and high-frequency bands, and licensed and unlicensed bands.
- Have stronger device-connection capabilities to deal with the access requirements of huge amounts of IoT devices.

In terms of operation and maintenance (O&M), 5G needs to:

- Improve network energy efficiency and the O&M cost-per-bit to cope with data traffic growth and the diverse needs of various services and applications.

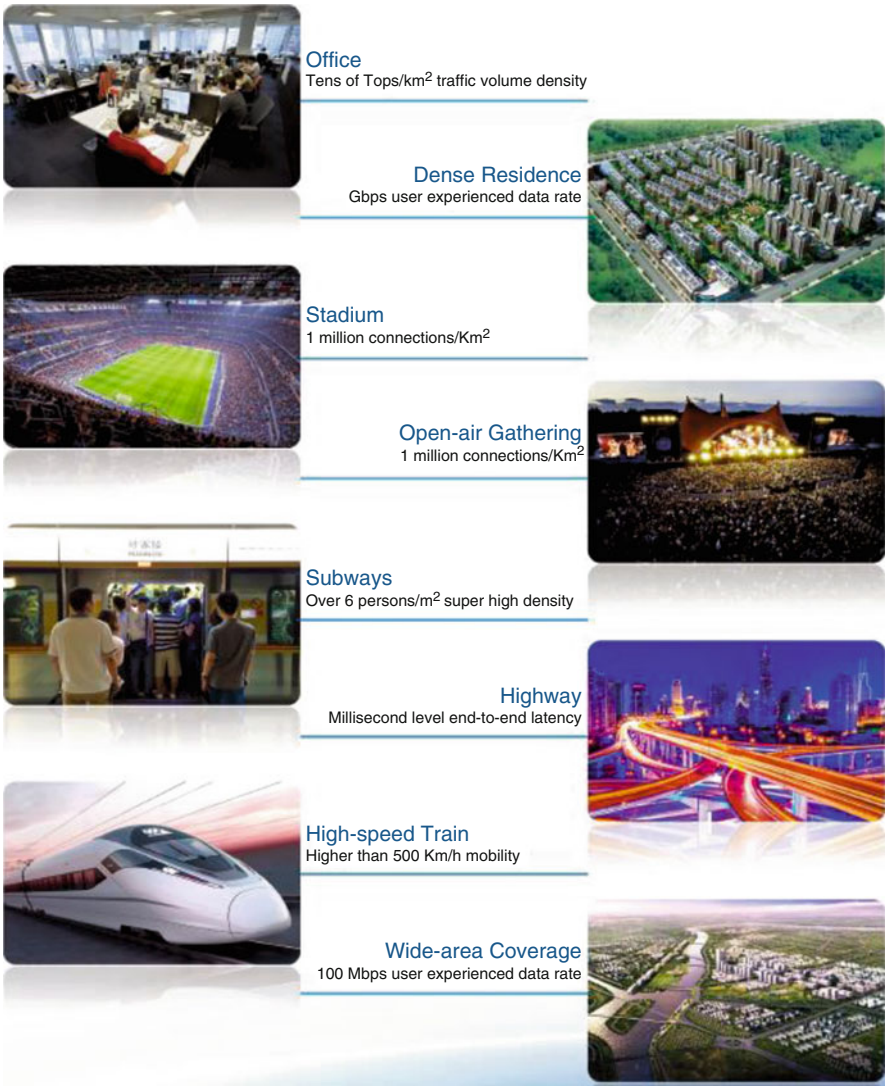


**Fig. 10** Challenging scenarios and performance indicators

- Reduce the complexity caused by the co-existence of multiple radio access technologies, network upgrades, and the introduction of new features and functions, to improve users' experience.
- Make intelligent optimization based on awareness of users behaviors and services contents
- Provide a variety of network security solutions to meet the needs of all types of devices and services of mobile internet and IoT.

**Table 2** 5G Key efficiency indicators

| Efficiency indicators | Definition |
|---|---|
| Spectrum efficiency (bps/Hz/cell or bps/Hz/km$^2$) | Data throughput per unit of spectrum resource per cell (or per unit area) |
| Energy efficiency (bit/J) | Number of bits that can be transmitted per joule of energy |
| Cost efficiency (bit/¥) | Number of bits that can be transmitted per unit cost |

Spectrum utilization, energy consumption and cost are the three key factors which must be addressed in sustainable mobile communication networks. In order to achieve sustainability, 5G needs to make significant improvements in the following aspects (Table 2):

- Spectrum efficiency: 3–5 times
- Energy efficiency: 100+ times
- Cost efficiency: 100+ times

5G systems must dramatically outperform previous generation systems. 5G should support

- User experienced data rate: 0.1–1 Gbps
- Connection density: one million connections per square kilometer
- End-to-end latency: millisecond level
- Traffic volume density: tens of Gbps per square kilometer
- Mobility: higher than 500 km per hour
- Peak data rate: tens of Gbps

Among these requirements, the user experienced data rate, connection density and end-to-end latency are the three most fundamental ones. Meanwhile, 5G needs to significantly improve the efficiency of network deployment and operations. Compared with 4G, 5G should have 3–5 times improvement on spectrum efficiency and more than 100 times improvement on energy and cost efficiency.

The performance requirements and efficiency requirements define the key capabilities of 5G, which can be illustrated as a "blooming flower" depicted in Fig. 11. The petals and leaves rely on each other. The petals represent the six key capabilities in terms of performance and can fulfill the diverse requirements of future services and scenarios. The leaves represent the three key capabilities in terms of efficiency, and can guarantee the sustainable development of 5G. The top of each petal means the maximum value of the corresponding capability.

The key capabilities of IMT-2020 defined by ITU-R are shown in Fig. 12, compared with those of IMT-Advanced. The values for each key capability are shown in Table 3.

All key capabilities may to some extent be important for most use cases, and the relevance of certain key capabilities may be significantly different, depending on the use cases/scenario. The importance of each key capability for the usage scenarios

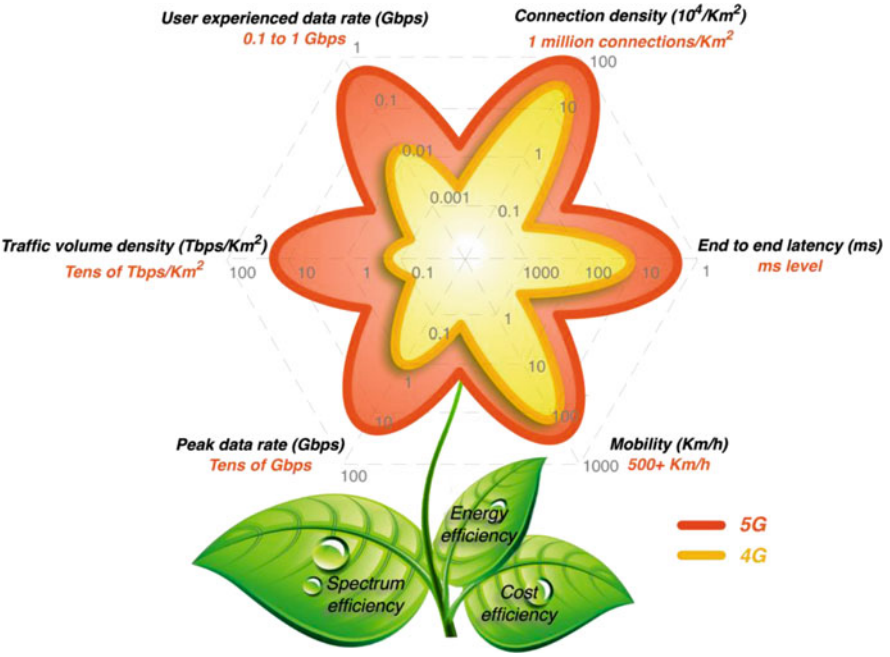


**Fig. 11** 5G key capabilities

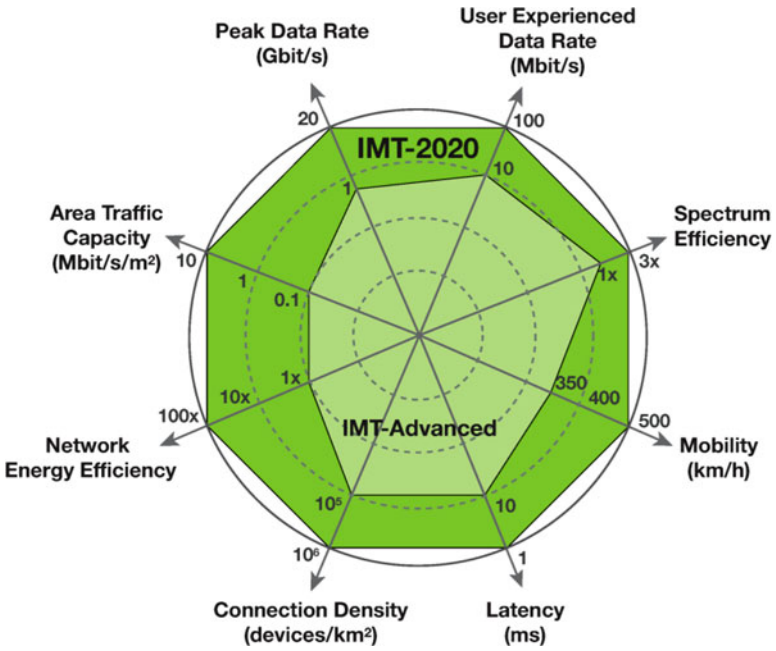


**Fig. 12** Enhancement of key capabilities from IMT-Advanced to IMT-2020

**Table 3** 5G Key capabilities and values from ITU-R

| Key capabilities | Values |
|---|---|
| Peak data rate | 20 Gbps |
| User experienced data rate | 0.1–1 Gbps |
| Latency | 1 ms over-the-air |
| Mobility | 500 km/h |
| Connection density | $10^6$/km$^2$ |
| Energy efficiency | 100 times compared with IMT-Advanced |
| Spectrum efficiency | 3–5 times compared with IMT-Advanced |
| Area traffic capacity | 10 Mbit/s/m$^2$ |

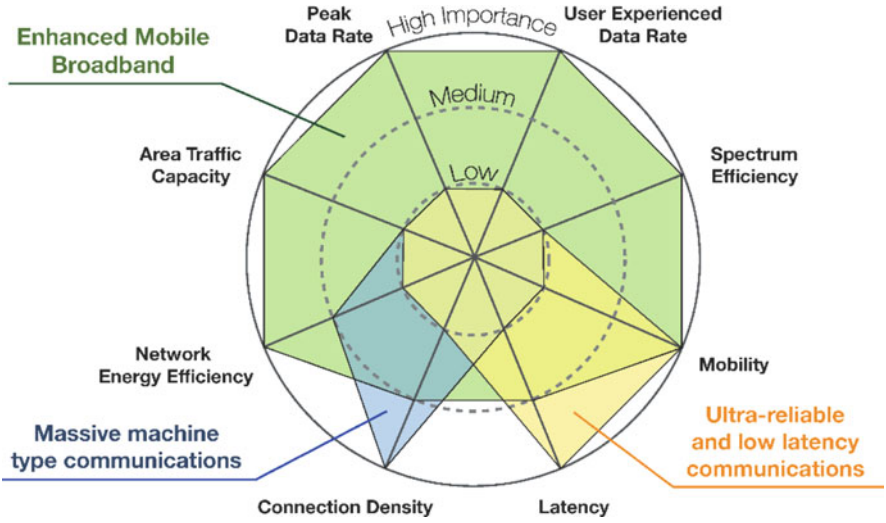


**Fig. 13** Importance of key capabilities in different usage scenarios

enhanced mobile broadband, ultra-reliable and low latency communication and massive machine-type communication is illustrated in Fig. 13. This is done using an indicative scaling in three steps as "high", "medium" and "low".

In the enhanced mobile broadband scenario, the user experienced data rate, area traffic capacity, peak data rate, mobility, energy efficiency and spectrum efficiency all have high importance, but mobility and the user experienced data rate would not have equal importance simultaneously in all use cases. For example, in hotspots, a higher user experienced data rate, but a lower mobility, would be required than in the wide area coverage case.

In the ultra-reliable and low latency communications scenarios, low latency is of highest importance, e.g., safety critical applications. Such capability would be required in some high mobility cases as well, e.g., transportation safety, while high data rates could be less important.

In the massive machine type communication scenario, high connection density is needed to support a tremendous number of devices in the network that may transmit only occasionally, at low bit rate and with zero/very low mobility. A low cost device with long operational lifetime is vital for this usage scenario.

## 4 Deployments Scenarios

Use cases will be delivered across a wide range of environments. To facilitate the study of 5G requirements and provide guidance to 5G technical design, several typical deployment scenarios need to be specified.

3GPP has started study on scenarios and requirements for next generation access technologies [6] in December 2015. After three months' study, there are totally 10 deployment scenarios specified for 5G in RAN plenary meeting in March 2016. 7 deployment scenarios including indoor hotspot, dense urban, rural, urban macro, high speed, extreme rural for the provision of minimal services over long distances and extreme rural with extreme long range are proposed mainly for eMBB, 1 deployment scenario, urban coverage for massive connection is proposed for mMTC, 2 deployment scenarios including highway scenario (for Internet of vehicles) and urban grid for connected car are proposed for URLLC. Some eMBB deployment scenarios may possibly be reused to evaluate mMTC and URLLC, or some specific evaluation tests (e.g., link-level simulation) can be developed to check whether the requirements can be achieved.

High-level descriptions on deployment scenarios including carrier frequency, aggregated system bandwidth, network layout including Inter-Site-Distance (ISD), UE distribution, UE moving speed and service profile are proposed in the following sections.

- Indoor hotspots

The indoor hotspot deployment scenario focuses on small coverage per cell and high user throughput or user density in buildings. The key characteristics of this deployment scenario are high capacity, high user density and consistent user experience indoor.

This scenario represents indoor offices with a total area of 120 m $\times$ 50 m. 12 small cells are deployed with an ISD of 20 m. The BS antenna height is 3 m. The carrier frequency options include 4 GHz, around 30 and 70 GHz. The bandwidth for 4 GHz is up to 200 MHz. The bandwidth for around both 30 and 70 GHz is up to 1 GHz. 10 users per cell are distributed uniformly and all users are indoors with 3 km per hour velocity. Full buffer and/or burst traffic model is assumed.

- Dense urban

The dense urban heterogeneous deployment scenario focuses on macro cells with micro cells and high user densities and traffic loads in city centres and dense urban areas. The key characteristics of this deployment scenario are high traffic loads,

outdoor and outdoor-to-indoor coverage. This scenario will be interference-limited, using macro TRPs with or without micro cells. A continuous cellular layout and the associated interference shall be assumed.

The ISD for the macro cells is 200 m. There are 3 micro cells per macro cell. The macro BS antenna height is 25 m and micro BS antenna height is 10 m. The carrier frequency for macro cell is 4 GHz. The carrier frequencies for micro cell include 4 GHz, around 30 GHz and around 70 GHz. The bandwidth for 4 GHz is up to 200 MHz and bandwidth for both around 30 GHz and around 70 GHz is up to 1 GHz. Full buffer and/or burst traffic model is assumed. 10 UEs are distributed per micro sector. 80 % users are indoor with a moving speed of 3 km per hour and 20 % is in cars with a velocity of 30 km per hour.

- Urban macro

The urban macro homogeneous deployment scenario focuses on large cells and continuous coverage. The key characteristics of this scenario are continuous and ubiquitous coverage in urban areas. This scenario will be interference-limited, using macro TRPs (i.e. radio access points above rooftop level).

The ISD in this scenario is 500 m. The BS antenna height is 35 m. The carrier frequency is 4 GHz, 2 GHz, and around 30 GHz. The bandwidth for 4 GHz is up to 200 MHz and bandwidth for 2 GHz is up to 100 MHz. Full buffer and/or burst traffic model is assumed. 10 UEs are distributed per cell. 80 % users are indoor with 3 km per hour velocity and 20 % is in cars with 30 km per hour velocity.

- Rural

The rural deployment scenario focuses on larger and continuous coverage. The key characteristics of this scenario are continuous wide area coverage supporting high speed vehicles. This scenario will be noise-limited and/or interference-limited, using macro TRPs.

The ISD in this scenario is 1732 m or 5000 m. The BS antenna height is 35 m. The carrier frequency is 700 MHz. The bandwidth is up to 20 MHz. Full buffer and/or burst traffic model is assumed. 10 UEs are distributed per cell. 50 % users are indoors with 3 km per hour velocity and 50 % is in cars with 120 km per hour velocity.

4 GHz and 2 GHz frequency are also considered in this scenario.

- High speed

The high speed deployment scenario focuses on continuous coverage along track in high speed trains. The key characteristics of this scenario are consistent user experience with very high mobility. In this deployment scenario, dedicated linear deployment along railway line is considered and UEs are located in train carriages. If the antenna of relay node for eNB-to-Relay is located at top of one carriage of the train, the antenna of relay node for Relay-to-UE could be distributed to all carriages.

- Extreme rural for the provision of minimal services over long distances

The extreme rural deployment scenario is defined to allow the provision of minimal services over long distances for low average revenue per user (ARPU) and low density areas including both humans and machines. The key characteristics of this scenario are macro cells with very large area coverage supporting basic data and voice services, with low to moderate user throughput and low user density.

- Extreme rural with extreme long range

The extreme rural long range deployment scenario is defined to allow for the provision of services for very large areas such as wilderness or areas with only highways. The key characteristics of this scenario are macro cells with very large area coverage supporting basic data speeds and voice services, with low to moderate user throughput and low user density.

- Urban coverage for massive connection

The urban coverage for massive connection scenario focuses on large cells and continuous coverage to provide mMTC. The key characteristics of this scenario are continuous and ubiquitous coverage in urban areas, with very high connection density of mMTC devices. This deployment scenario is for the evaluation of the KPI of connection density.

- Highway scenario

The highway deployment scenario focuses on scenario of vehicles on highways with high speeds. The main KPIs evaluated under this scenario include reliability/availability under high speeds/mobility (and thus frequent handover operations).

- Urban grid for connected car

The urban macro deployment scenario focuses on the scenario of highly dense deployed vehicles placed in urban areas. It can cover the scenario where freeways lead through an urban grid. The main KPI evaluated under this scenario are reliability/availability/latency in high network load and high UE density scenarios.

## 5 Detailed Technical Requirements

- Peak data rate

Peak data rate is the highest theoretical data rate which is the received data rate assuming error-free conditions assignable to a single mobile station, when all assignable radio resources for the corresponding link direction are utilized (i.e., excluding radio resources that are used for physical layer synchronization, reference signals or pilots, guard bands and guard times).

The target for peak data rates are 20 Gbps for downlink and 10 Gbps for uplink.

- Peak spectral efficiency

Peak spectral efficiency is the highest theoretical data rate (normalized by bandwidth), which is the received data rate assuming error-free conditions assignable to a single mobile station, when all assignable radio resources for the corresponding link direction are utilized (i.e., excluding radio resources that are used for physical layer synchronization, reference signals or pilots, guard bands and guard times).

The targets for peak spectral efficiency are 30 bps/Hz for downlink and 15 bps/Hz for uplink.

Higher frequency bands imply higher bandwidths but lower spectral efficiency, while lower frequency bands results in lower bandwidths but higher spectral efficiency. Thus, the peak data rate cannot be directly derived from the peak spectral efficiency and bandwidth multiplication.

- Bandwidth

Bandwidth means the maximal aggregated total system bandwidth. It may be supported by a single or multiple RF carriers.

The value for this KPI is for further study.

- Control plane latency

Control plane latency refers to the time to move from a battery efficient state (e.g., IDLE) to the start of continuous data transfer (e.g., ACTIVE).

The target for control plane latency should be 10 ms.

- User plane latency

The time taken to successfully deliver an application layer packet/message from the radio protocol layer 2/3 SDU ingress point to the radio protocol layer 2/3 SDU egress point via the radio interface in both the uplink and downlink directions, where neither device nor base station reception is restricted by DRX.

For URLLC the target for user plane latency should be 0.5 ms for UL and 0.5 ms for DL. Furthermore, if possible, the latency should also be low enough to support the use of next-generation access technologies as a wireless transport technology that can be used within the next-generation access architecture.

For other cases, the target for user plane latency should be 4 ms for UL and 4 ms for DL.

- Latency for infrequent small packets

For infrequent application layer small packet/message transfer, the time it takes to successfully deliver an application layer packet/message from the radio protocol layer 2/3 SDU ingress point at the mobile device to the radio protocol layer 2/3 SDU egress point in the RAN, when the mobile device starts from its most “battery efficient” state.

- Mobility interruption time

Mobility interruption time means the shortest time duration supported by the system during which a user terminal cannot exchange user plane packets with any base station during transition.

The target for mobility interruption time should be zero.

This KPI is for intra-system mobility.

- Inter-system mobility

Inter-system mobility refers to the ability to support mobility between the IMT-2020 system and at least one IMT system.

- Reliability

Reliability can be evaluated by the success probability of transmitting e.g., 20 bytes within 1 ms, which is the time it takes to deliver a small data packet from the radio protocol layer 2/3 SDU ingress point to the radio protocol layer 2/3 SDU egress point of the radio interface, at a certain channel quality (e.g., coverage-edge).

The target for reliability should be $1–10^{-5}$ within 1 ms.

- Coverage

"Maximum coupling loss" (MCL) on the uplink and downlink between device and Base Station site (antenna connector(s)) for a data rate of 160 bps, where the data rate is observed at the egress/ingress point of the radio protocol stack in uplink and downlink.

The target for coverage should be 164 dB.

- UE battery life

The UE battery life is determined by the battery life of the UE without recharge. For mMTC, the UE battery life in extreme coverage depends on the activity of mobile originated data transfer consisting of 200 bytes UL per day followed by 20 bytes DL from MCL of 164 dB, assuming a stored energy capacity of 5 Wh.

The target for UE battery life is beyond 10 years.

- UE energy efficiency

UE energy efficiency means the capability of a UE to sustain much better mobile broadband data rate while minimizing the UE modem energy consumption.

- Cell/Transmission Point/TRP spectral efficiency

TRP spectral efficiency is defined as the aggregate throughput of all users (the number of correctly received bits, i.e., the number of bits contained in the service data units (SDUs) delivered to Layer 3, over a certain period of time) divided by the channel bandwidth divided by the number of TRPs. A three-sector site consists of 3 TRPs. In case of multiple discontinuous "carriers" (one carrier refers to a continuous block of spectrum), this KPI should be calculated per carrier. In this case, the aggregate throughput, channel bandwidth, and the number of TRPs on the specific carrier are employed.

The target considered as a starting point for eMBB deployment scenarios is in the order of 3× IMT-Advanced requirements for full buffer.

- Area traffic capacity

Area traffic capacity means total traffic throughput served per geographic area (in Mbit/s/m$^2$). This metric can be evaluated by two different traffic models, i.e., full buffer model and non-full buffer model.

Full buffer model:

The total traffic throughput served per geographic area (in Mbit/s/m$^2$). The computation of this metric is based on full buffer traffic.

Non-full buffer model:

The total traffic throughput served per geographic area (in Mbit/s/m$^2$). Both the user experienced data rate and the area traffic capacity need to be evaluated at the same time using the same traffic model.

The area traffic capacity is a measure of how much traffic a network can carry per unit area. It depends on the site density, bandwidth and spectrum efficiency. In the special case of a single layer single band system, it may be expressed as:

$$\begin{aligned}\text{area capacity } \left(\text{bps}/\text{m}^2\right) = {} & \text{site density } \left(\text{site}/\text{m}^2\right) \times \text{bandwidth (Hz)} \\ & \times \text{spectrum efficiency (bps/Hz/site)}.\end{aligned}$$

Based on the above, it is proposed to use the spectrum efficiency results together with the assumptions on the available bandwidth and site density in order to derive a quantitative area traffic capacity KPI for information.

- User experienced data rate

User experienced data rate can be evaluated for non-full buffer traffic and for full buffer traffic. Non-full buffer simulations are preferred for the evaluation of this KPI.

For non-full buffer traffic, the user experienced data rate is 5 % of the user throughput. User throughput (during active time) is defined as the size of a burst divided by the time between the arrival of the first packet of the burst and the reception of the last packet of the burst.

The target values for the user experienced data rate are associated with non-full buffer evaluation. The non-full buffer user experienced data rate target is applicable at the non-full buffer area traffic capacity traffic level.

For full buffer traffic, the user experienced data rate is calculated as:

$$\text{user experienced data rate} = 5\% \text{ user spectrum efficiency} \times \text{bandwidth}$$

To improve user experienced data rates, 3GPP can develop standards with means for high 5 % user spectrum efficiency. To this end, 5 % user spectrum efficiency gains in the order of three times that IMT-Advanced are proposed. Furthermore, 3GPP can develop standards with means for large bandwidth support. Towards this end, it is proposed that at least 1 GHz aggregated bandwidth shall be supported.

The available bandwidth and site density, which both have a strong impact on the available user experienced data rates, are however beyond the control of 3GPP.

Based on this, the full buffer experienced user data rate is evaluated for information without numerical requirements.

- 5th percentile user spectrum efficiency

5th percentile user spectrum efficiency means the 5 % point of the cumulative distribution function (CDF) of the normalized user throughput. The (normalized) user throughput is defined as the average user throughput (the number of correctly received bits by users, i.e., the number of bits contained in the SDU delivered to Layer 3, over a certain period of time, divided by the channel bandwidth and is measured in bit/s/Hz. The channel bandwidth for this purpose is defined as the effective bandwidth multiplies the frequency reuse factor, where the effective bandwidth is the operating bandwidth normalized appropriately considering the uplink/downlink ratio. In the case of multiple discontinuous "carriers" (one carrier refers to a continuous block of spectrum), this KPI should be calculated per carrier. In this case, the user throughput and channel bandwidth on the specific carrier are employed.

The target considered as a starting point for eMBB deployment scenarios is in the order of 3× IMT-Advanced requirements for full buffer.

- Connection density

Connection density refers to the total number of devices fulfilling specific QoS per unit area (per $km^2$). QoS definition should take into account the amount of data or access request generated within a time X (to be studied in future) that can be sent or received within a given time, Y (to be studied in future), with Z% (to be studied in future) probability.

The target for connection density should be 1,000,000 device/$km^2$ in the urban environment.

- Mobility

Mobility means the maximum user speed (km/h) at which a defined QoS can be achieved.

The target for mobility target is 500 km/h.

- Network energy efficiency

The capability is to minimize the RAN energy consumption while providing much better area traffic capacity.

Qualitative KPIs as the baseline and quantitative KPI are for further study.

## 6 Requirements for RAN Architecture

The RAN design for the Next Generation Radio Access Technologies shall be designed to fulfill the following requirements:

- The RAN architecture shall support tight interworking between the new RAT and LTE.
- Considering high performing inter-RAT mobility and aggregation of data flows via at least dual connectivity between LTE and new RAT. This shall be supported for both collocated and non-collocated site deployments;
- The RAN architecture shall support connectivity through multiple transmission points, either collocated or non-collocated;
- The RAN architecture shall enable the separation of control plane signaling and user plane data from different sites;
- The RAN architecture shall support interfaces supporting effective inter-site scheduling coordination;
- Different options and flexibility for splitting the RAN architecture shall be allowed;
- The RAN architecture shall allow for deployment flexibility e.g., to host relevant RAN, CN and application functions close together at the edges of the network, when needed, e.g., to enable context aware service delivery, low latency services, etc;
- The RAN architecture shall allow for C-plane/U-plane separation;
- The RAN architecture shall allow deployments using Network Function Virtualization;
- The RAN architecture shall allow for the RAN and the CN to evolve independently;
- The RAN architecture shall allow for the operation of Network Slicing;
- The RAN architecture shall support sharing of the RAN between multiple operators;
- The design of the RAN architecture shall allow for rapid and efficient deployment of new services;
- The design of the RAN architecture shall allow the support of 3GPP defined service classes (e.g. interactive, background, streaming and conversational);
- The design of the RAN architecture shall enable lower CAPEX/OPEX with respect to current networks to achieve the same level of services;
- RAN-CN interfaces and RAN internal interfaces (both between new RAT logical nodes/functions and between new RAT and LTE logical nodes/functions) shall be open for multi-vendor interoperability;
- The RAN architecture shall support operator-controlled sidelink (device-to-device) operation, both in coverage and out of coverage.

## 7 Chapter Summary

Mobile Internet and IoT are the two main market drivers in the future development of mobile communications, and they will provide a broad range of prospects for 5G. There will be a massive number of use cases in 5G era, such as augmented reality, virtual reality, remote computing, eHealth services, automotive driving and so on. All these use cases can be grouped into three categories, i.e., eMBB, mMTC and URLLC. To facilitate the study of 5G requirements and provide guidance to 5G technical design, several typical deployment scenarios are specified including indoor hotspots, dense urban, urban macro, rural and high-speed scenarios.

This chapter also presents high-level key capabilities and detailed technical requirements for 5G. 5G will be able to sustainably satisfy the requirement of thousands of times mobile data traffic growth. 5G will provide users with fiber-like access data rate and "zero" latency user experience. It will be capable of connecting 100 billion devices. 5G will be able to deliver a consistent experience across a variety of scenarios including the cases of ultra-high traffic volume density, ultra-high connection density and ultra-high mobility. 5G will also be able to provide intelligent optimization based on services and user awareness, and will improve energy and cost efficiency by over a hundred times, enabling us all to realize the vision of 5G—"Information a finger away, everything in touch".

## References

1. Whitepaper, 5G vision and requirements, IMT-2020 (5G) Promotion Group
2. Recommendation M.2083, Framework and overall objectives of the future development of IMT for 2020 and beyond, ITU-R
3. Workplan, timeline, process and deliverables for the future development of IMT, ITU-R WP5D
4. New WID on New Services and Markets Technology Enablers – SmarTer, Vodafone, February 2015
5. 5G white paper, NGMN, March 2015
6. RP-160689, Study on Scenarios and Requirements for Next Generation Access Technologies, 3GPP TR 38.913 V0.3.0, March 2016
7. Further Elaboration on NGMN Requirement Metrics and Deployment Scenarios for IMT-2020 in Response to Liaison Statement from ITU-R WP5D from 4th February 2015, NGMN Alliance, February 2016
8. Further elaboration on NGMN KPIs and requirements for 5G, NGMN Alliance, March 2016
9. Deliverable D1.1, Scenarios, requirements and KPIs for 5G mobile and wireless system, METIS
10. Whitepaper, 5G NEW WAVE TOWARDS FUTURE SOCIETIES IN THE 2020S, 5G forum
11. Whitepaper, 4G Americas Recommendations on 5G Requirements and Solutions, 4G Americas
12. Report, IMT Traffic estimates for the years 2020 to 2030, ITU-R

# Spectrum Analysis and Regulations for 5G

**Tan Wang, Gen Li, Biao Huang, Qingyu Miao, Jian Fang, Pengpeng Li, Haifeng Tan, Wei Li, Jiaxin Ding, Jingchun Li, and Ying Wang**

**Abstract** In 5G vision, the spectrum issue is one of the most important parts. Governments, agencies, standardization organizations and research institutions from many countries pay high attention to the 5G spectrum strategies. The appeals for international harmonized spectrum and full band spectrum access are intense, the range of which are possibly from 0-100GHz. This chapter addresses the current spectrum for mobile communications, the future spectrum demand, possible candidate frequency bands and spectrum management considerations. Information from international and regional telecommunications such as ITU, CEPT, APT, and also from different countries are collected and analyzed. Some academic views are also provided for future work.

## 1 Introduction

Mobile communication has gone deep into people's lives. For the year 2020 and beyond, all kinds of new services will continue to emerge, mobile data traffic will show explosive growth, and connections from terminals will increase exponentially. With the rapid development of Mobile Internet and Internet of Things, the 5G will be integrated into each field of the society in the future, and construct a full range of information ecosystem, providing a broad prospect.

Spectrum is the one of the most valuable resource for mobile communications. Among the 5G studies and pre-standardization works, the spectrum issue is one of the most important parts. In this context, the global 5G research institutions pay high attention to the research on spectrum, such as EU FP7 METIS project [1] and China's IMT-2020 Promotion Group [http://www.imt-2020.org.cn/zh]. In addition, some countries put the spectrum for next generation of mobile communication in

T. Wang (✉) • B. Huang • J. Fang • P. Li • H. Tan • W. Li • J. Ding • J. Li
State Radio Monitoring Center/State Radio Spectrum Management Center, MIIT, Beijing, China
e-mail: wangtan@srrc.org.cn

G. Li • Q. Miao
Ericsson, China Radio Research Lab, Beijing, China

Y. Wang
Beijing University of Posts and Telecommunications (BUPT), Beijing, China



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_2


**Fig. 1** General consideration of spectrum for 5G

a prominent position [2, 3]. Furthermore, in the framework of ITU, the ITU-R study groups (e.g., WP 5D) and related regional telecommunication organizations (including ASMG, APT, CEPT, CITEL, RCC, ATU and so on) are actively conducting studies for 5G spectrum.

According to current research, generally, spectrum below 6 GHz is the best resource in the near future [4]. These bands mainly include re-farmed 2G/3G spectrum, identified frequency band for IMT in Radio Regulations and also WRC-15 candidate bands. However, due to its scarcity and increasingly difficulty to realize international harmonization after WRC-15, it is the time to seek spectrum above 6 GHz, as shown in Fig. 1. The 5G spectrum solution will be a comprehensive outcome, indicating combinations of different frequency ranges for different scenarios. It is expected that 5G requires more in the total amount, wider with respect to individual bandwidths, greater in the range and more flexible in the usage and management pattern.

In this chapter, 5G spectrum related issues are investigated, discussed and summarized, taking typical regions and countries as examples around the world including Asia-Pacific, Europe and Americas. In Sect. 2, the current used spectrum for mobile communication is described. Section 3 focuses on the 5G spectrum demand, followed by the 5G potential candidate frequency bands in Sect. 4. Some new ideas in 5G spectrum management are investigated in Sect. 5 to give some food for thought. We conclude this chapter in Sect. 6.

## 2 Current Spectrum for Mobile Communications

The sharing of information on frequency allocation and spectrum usages will be beneficial to reach harmonization on spectrum usages among all countries around the world, especially for mobile communications. Moreover, as described in Sect. 4, the current allocated spectrum for International Mobile Telecommunication (IMT) is an important part of 5G candidate spectrum. In this regard, this chapter starts with the usage status of mobile communication spectrum. In different countries, due to the different development stages of mobile communication industry, the total amount of spectrum, different frequency ranges for 2G/3G/4G vary widely, but also reflects a certain degree of consistency.

Take the situation of Asia as the first step. In Asia, it was agreed to develop an APT Report on information of mobile operators' frequencies, technologies and license durations in Asia Pacific countries [5]. The APT report is a compilation

**Table 1** Operators, frequency bands, used technology and license duration of Japan

| Japan | | | | |
|---|---|---|---|---|
| | Frequency band (MHz) | | | |
| Operator | Uplink | Downlink | Technology | License duration |
| NTT DoCoMo, Inc. | 728–738 | 783–793 | LTE | Period: 5 years (Note) |
| | 830–845 | 875–890 | WCDMA/LTE | |
| | 1447.9–1462.9 | 1495.9–1510.9 | LTE | |
| | 1764.9–1784.9 | 1859.9–1879.9 | WCDMA/LTE | |
| | 1940–1960 | 2130–2150 | WCDMA/LTE | |
| | 3480–3520 (TDD) | | LTE | |
| KDDI Corporation | 718–728 | 773–783 | LTE | |
| | 815–830 | 860–875 | CDMA2000/LTE | |
| | 1437.9–1447.9 | 1485.9–1495.9 | LTE | |
| | 1920–1940 | 2110–2130 | CDMA2000/LTE | |
| | 3520–3560 (TDD) | | LTE | |
| Softbank Mobile Corp. | 900–915 | 945–960 | WCDMA/LTE | |
| | 1427.9–1437.9 | 1475.9–1485.9 | WCDMA | |
| | 1960–1980 | 2150–2170 | WCDMA/LTE | |
| | 3560–3600 (TDD) | | LTE | |
| Ymobile Corp. | 738–748 | 793–803 | LTE | |
| | 1749.9–1764.9 | 1844.9–1859.9 | WCDMA/LTE | |
| Ymobile Corp. | 1884.5–1915.7 (TDD) | | PHS | |
| Wireless City Planning Inc. | 2545–2575 (TDD) | | AXGP (Advanced eXtended Global Platform) | |
| UQ Communications Inc. | 2595–2645 (TDD) | | WiMAX/WiMAX2+ | |

*Note*: License is granted to Base Station each. Expiration date of license for each station is different

of responses to the Questionnaire and up to now includes information from 26 members such as Australia, China, India, Iran, Japan, Korea, and Thailand, etc. One example with relatively new information is shown in Table 1.

For US, the public mobile communications spectrum is mainly reflected in Title 47 Telecommunication [http://reboot.fcc.gov/reform/systems/spectrum-dashboard]. As for Europe, the ECC is in response of developing European common positions and proposals for efficient use of radio spectrum in the framework of international and regional bodies. A European Table of Frequency Allocations and Applications (ECA Table) [6] is provided by CEPT.

In general, the frequency table for mobile communication in countries can be found in national official publications or output documents in regional organizations. It is shown that almost all the frequency ranges are focus on the spectrum below 3 GHz. Another observation from these data is that generally the amount of spectrum currently used for mobile communications is several hundred megahertz.

# 3 5G Spectrum Demand

Every few years, ITU-R sets up agenda items in advance to study the future spectrum demand for IMT, and to support the consideration of additional spectrum allocations. Currently, ITU-R has already finished the study on IMT spectrum demand towards the year 2020. Some countries have also started the study on spectrum demand beyond 2020.

## *3.1 Demand in the Year 2020*

How to calculate IMT spectrum demands? Generally, a methodology starts with an analysis of future market and traffic volume, moves on to calculate and distribute the traffic on different RATs, and then calculates the required capacity, before concluding the estimation. The actual calculation process can be very complicated, when there are a variety of traffic types, different environments and multiple cell types of different RATs. For example, imagine how to estimate the data rates of a high quality video streaming user located in indoor offices, connecting with future 5G small cells, in the year 2025.

Many countries have made contributions to the calculation methodology and output results. Table 2 summarizes national spectrum requirements as provided by some countries and organizations during the study of ITU-R, in the form of total amount for all the operators in one country [7]. It should be noted that these national spectrum requirements have differences in the methodology used and assumptions made (e.g., differences in traffic/radio-aspects related parameters, differences in estimation year, differences in estimates based on whether the spectrum requirements are total or additional, etc.).

Some of them, such as GSMA (GSM Association) and UK, focused on the improvements of existing ITU-R method specified in Recommendation M.1768-1.

The methodology of Recommendation ITU-R M.1768-1 is developed by ITU-R Study Group, which is used in WRC-07 to calculate the spectrum demand in the future. The methodology provides the spectrum requirements of IMT as a whole, and divided between two radio access technique groups (RATGs):

- RATG 1: Pre-IMT systems, IMT-2000 and its enhancements;
- RATG 2: IMT-Advanced.

The methodology reflects certain recent advances in IMT technologies and the deployment of IMT networks such as the introduction of spectrum sharing between the macro and micro cell layers in IMT Advanced, and the introduction of a new spectrum granularity parameter for IMT systems.

Figure 2 summarizes the steps of the calculation algorithm employed in Recommendation ITU-R M.1768-1 in conjunction with relevant input parameters (detailed descriptions of these parameters are provided in Sect. 4) [7]. The methodology starts

**Table 2** Summary of national spectrum requirements in some countries [7]

| Source | US | Australia | Russia | China | GSMA | India | UK |
|---|---|---|---|---|---|---|---|
| Estimation year | Until 2014 | Until 2020 | 2020 | 2015, 2020 | 2020 | 2017, 2020 | 2020 |
| Spectrum requirements | Additional 275 MHz by 2014 | 1081 MHz in total | 1065 MHz in total | 570–690 MHz in total by 20151490–1810 MHz in total by 2020 | 1600–1800 MHz in total for some countries | Additional 300 MHz by 2017Additional another 200 MHz by 2020 | 775–1080 MHz in total for low2230–2770 in total MHz for high |
| Methodology | Original | Original | Original | Rec. ITU-R M.1768-1 | Complementary to Rec. ITU-R M.1768-1 | Original | Rec. ITU-R M.1768-1 |

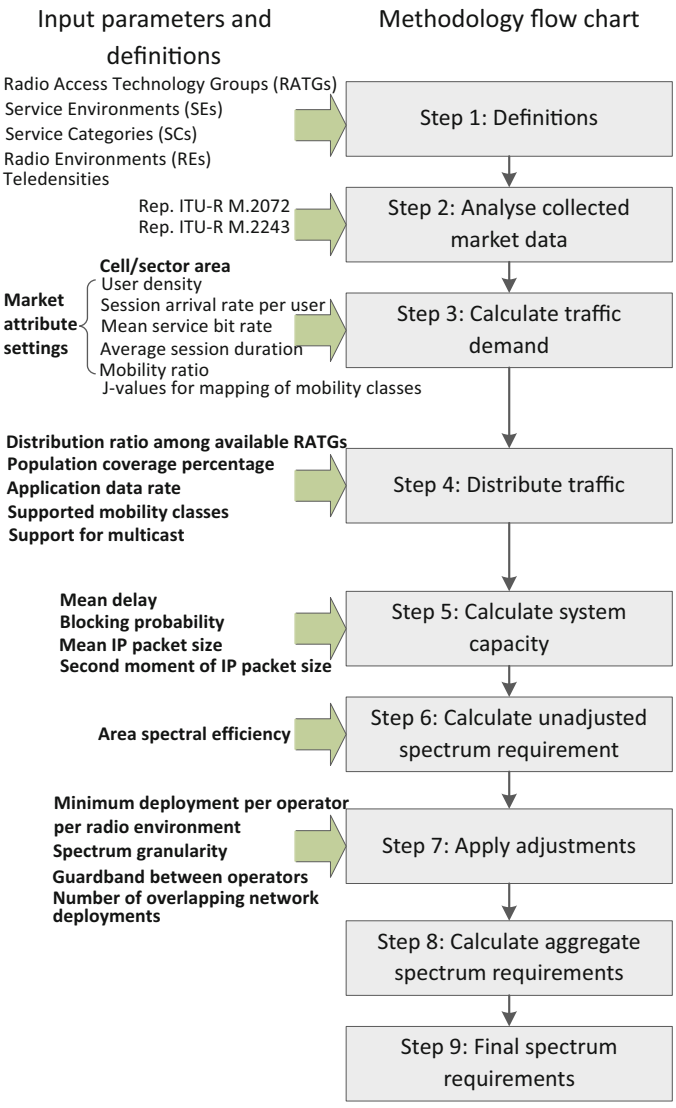


**Fig. 2** Steps of calculation algorithm and relevant input parameters [3]

from market studies that characterize all of the traffic carried by IMT and other mobile systems, corresponding to Steps 2–3. In Step 4, the total traffic obtained from the market studies is distributed among different radio environments (cell layers) and RATGs according to factors such as traffic characteristics (required data rates and user mobility), RATG capabilities (supported data rates, available cell types and their coverages), etc. The system capacity required to carry the offered traffic is then calculated in Step 5 using separate capacity calculation algorithms

**Table 3** Total spectrum requirements for both RATG 1 and RATG 2 in the year 2020 [7]

| Total spectrum requirements | RATG 1 | RATG 2 | RATGs 1 and 2 |
|---|---|---|---|
| Lower user density settings | 440 MHz | 900 MHz | 1340 MHz |
| Higher user density settings | 540 MHz | 1420 MHz | 1960 MHz |

for reservation-based traffic and packet-based traffic respectively. Initial spectrum estimates are obtained from the capacity requirements by dividing the latter by the spectral efficiencies (Step 6). Then, adjustments are made to take into account network deployments with the spectrum requirements being aggregated over the relevant deployments (Steps 7–8). Finally, the methodology outputs the overall spectrum requirements of RATG 1 and RATG 2, which collectively denote IMT systems (Step 9).

There are differences in the markets and deployments and timings of the mobile data growth in different countries. Therefore, two settings are developed to characterize lower and higher user density settings. These two sets of market study input parameter values are considered in the calculations to characterize differences in the user densities in different countries. Table 3 shows the calculated spectrum requirements for both RATGs 1 and 2.

In some countries, national spectrum requirement can be lower than the estimate derived by lower user density settings and in some other countries, national spectrum requirement can be higher than the estimate derived by higher user density settings.

Besides using existing methodologies, there are also original methodologies proposed by US [8], Australia [9], Russia [10] and so on. These methodologies are reflected as case studies based on national considerations. In the following, a brief example is shown from China, which is conducted from the IMT-2020 Promotion Group.

The main point of the new method is to estimate the upper bound of demand. In fact, it is proved by the operational data that the area which has the largest spectrum demand is always in metropolises. When requirements of such a scenario are satisfied, the solution might be applicable for others as well. In the calculation, a typical hotspot zone in Beijing is selected as the research area. To gather the historical data within, the government issued an investigation letter to all network operators. The data survey is based on the operator's network management system in China, which is a major difference between the proposed method and other international methods. Supported by these first-hand data, analyses are made to estimate the traffic increase, traffic distribution and Base Station (BS) deployments in the future. The general flow chart of the methodology is shown in Fig. 3. According to the calculation, the total IMT spectrum demand in China is 1350–1810 MHz in 2020 [11], which are a little different from but similar with the result from the calculation by M.1768-1 (1490–1810 MHz).

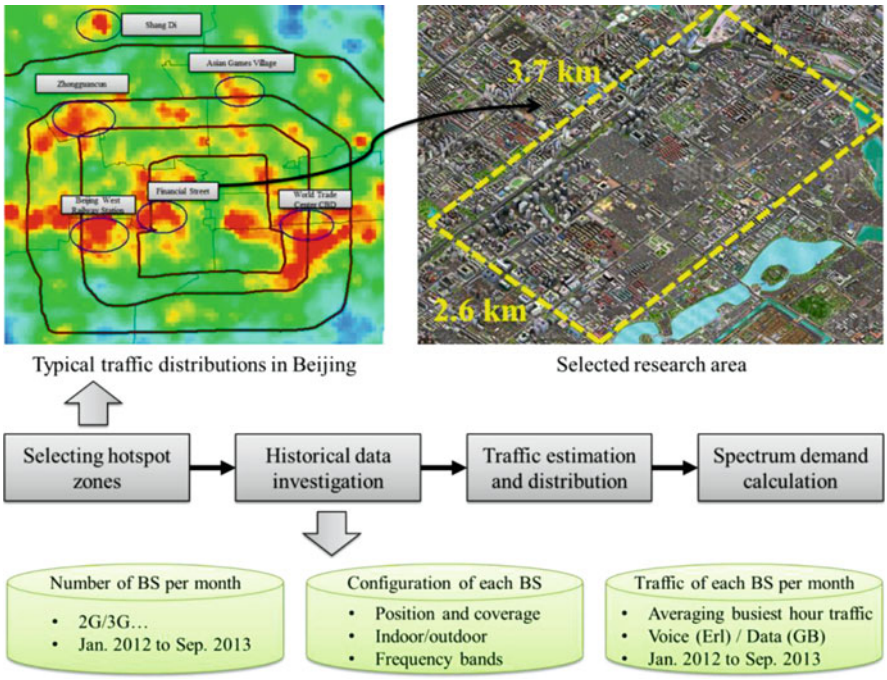


**Fig. 3** Flow chart of the spectrum demand methodology proposed by China, where a 3.7 km × 2.6 km square area around Beijing Financial Street is selected as the typical dense traffic area [4]

The obtained estimation results regardless from different methods are huge. Compared with current identified spectrum for IMT in different countries, there are still several hundred to even 1000 MHz deficit. Moreover, in the period of 5G commercialization after 2020, the total spectrum demand may continue to increase.

## *3.2 Demand Beyond the Year 2020*

The above result is, to some extent, a total amount of spectrum and the range depends on the assumptions made in the estimation process. However, in view of 5G beyond 2020, it is probably difficult to measure the new demand only in a total number. For example, in 5G indoor high traffic scenarios, in order to achieve high peak data rates, the demand for frequency bandwidth may be up to several GHz. This can be solved by using higher frequencies and denser deployments [12]. But these solutions may not be the best choice for outdoor wide area scenarios. Therefore, it will be useful to separately estimate how much spectrum is required for coverage, capacity, performance and connections for each 5G scenario as Fig. 4 shows, to

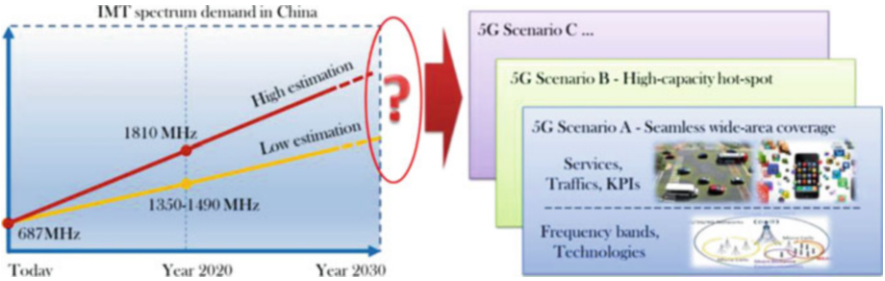


**Fig. 4** A scenario-based solution for spectrum demand in 2030 [http://www.imt-2020.org.cn/zh]

perform a mapping onto different frequency bands, such as bands below 1 GHz, between 1 and 6 GHz and above 6 GHz. It may be believed that the results from different scenarios will vary a lot.

Generally, the 5G spectrum demand estimation will be a comprehensive outcome, indicating spectrum solutions for different scenarios. It is expected that 5G requires more in the total amount, wider with respect to individual bandwidths, greater in the range and more flexible in the usage pattern.

## 4 5G Candidate Frequency Bands

Is the world ready to provide sufficient frequency bands for 5G? From the study point of view, the potential bands can be divided into two parts: bands below 6 GHz and above [4].

### *4.1 Spectrum Below 6 GHz*

For wireless communications, lower frequencies provide better coverage. Currently, almost all countries are using spectrum below 6 GHz for IMT systems. Besides achieving high data rates, it is also necessary to guarantee wide-area coverage and outdoor-to-indoor coverage in 5G. Therefore, spectrum below 6 GHz forms a very important part for the 5G spectrum solution. Potential 5G spectrum below 6 GHz includes the following aspects:

- Spectrum re-farming

In order to realize the benefits of new generation of mobile communication systems, regulators need to deliver efforts to enable re-farming of spectrum in the frequency band occupied by the old ones.

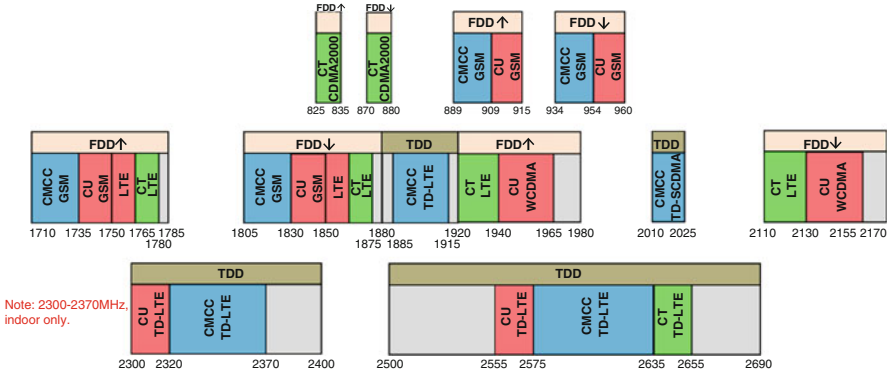


**Fig. 5** Frequency bands used for mobile technology in China (CT: China Telecom, CM: China Mobile, CU: China Unicom)

Taking China as an example, up to Feb. 2016, 517 MHz been allotted to three operators for eight networks, as shown in Fig. 5. For the two operators authorized for LTE hybrid network, China Telecom has $2 \times 15$ MHz spectrum for LTE FDD, while China Unicom only has $2 \times 10$ MHz in the beginning. Market competition has prompted Chinese Unicom to accelerate the pace of 2G spectrum re-farming for 3G and 4G networks. On the other hand, owing the world largest TD-LTE network, China Mobile has also expressed willingness to re-farm 2G spectrum for LTE FDD.

When 5G is put into use, it is expected that some spectrum from the old generation could be re-farmed for it as well. However, re-farming does not increase the total spectrum amount.

- Identified spectrum

In the past ITU-R studies, sometimes it is difficult to reach global or even regional agreement for the usage of some frequency bands. Therefore some Resolutions from WRC will not modify the Table of Frequency Allocations uniformly, but usually take effect in the form of adding footnotes for some countries. There are many cases for IMT frequency bands planning in this way.

For example, footnote **5.432A** specifies that "In Korea (Rep. of), Japan and Pakistan, the band 3400–3500 MHz is identified for International Mobile Telecommunications (IMT). This identification does not preclude the use of this band by any application of the services to which it is allocated and does not establish priority in the Radio Regulations . . . "

However, some identified IMT spectrum is not for immediate use. For example, in Regulations of Radio Frequency Division of China, there are several frequency bands identified for IMT specified in footnote CHN28 such as 2300–2400 MHz and 3400–3600 MHz [13], but only when the compatibility studies are finished. Currently in China, LTE-Hi (LTE Hotspot/indoor) is proposed to meet the coverage requirement of hotspot and indoor scenarios. Compared with current IMT, LTE-Hi is targeting higher frequency, including 3400–3600 MHz.

For 2300–2400 MHz, after careful studies, IMT systems have proven the capability of coexisting with radio location services, but only limited to indoor use. This band has been assigned in Dec. 2013, for deploying TD-LTE systems.

Nevertheless, other frequency bands are still to be studied with respect to compatibility before official use. For example, 3400–3600 MHz is already used as extended C band for satellite services in China, since it can provide better propagation characteristics against the rain attenuation than higher frequency bands. Therefore, up to now, allowing this band for IMT systems needs further demonstration and coordination.

- New frequency bands after WRC-15

In accordance with resolves to invite ITU-R 2 of Resolution 233 (WRC-12), there are 19 candidate frequency bands proposed for IMT from different administrations, in the framework of WRC-15 Agenda Item (AI) 1.1 [14], as shown in Fig. 6.

These frequency bands are considered as potential candidate frequency bands from among the suitable frequency ranges provided by the ITU-R. These frequency bands were proposed by one or more administrations and were studied by the ITU-R. For a frequency band to be included, the ITU-R JTG 4-5-6-7 developed the following criteria: it must have been proposed by an administration and have been studied.

In studying AI 1.1, inputs were received reflecting views on certain frequency ranges/bands. Some of the views are listed in Table 4. Studies related to various frequency bands (studies initiated, carried out, not completed, completed), as well as an overview of which services were and were not studied for each band, are addressed in [14].

Different administrations or regions may have different considerations in supporting these frequency bands. The final results could be found in Provisional Final Acts of WRC-15 [15]. It shows that, to realize global harmonization, it is necessary to achieve further regional convergence.

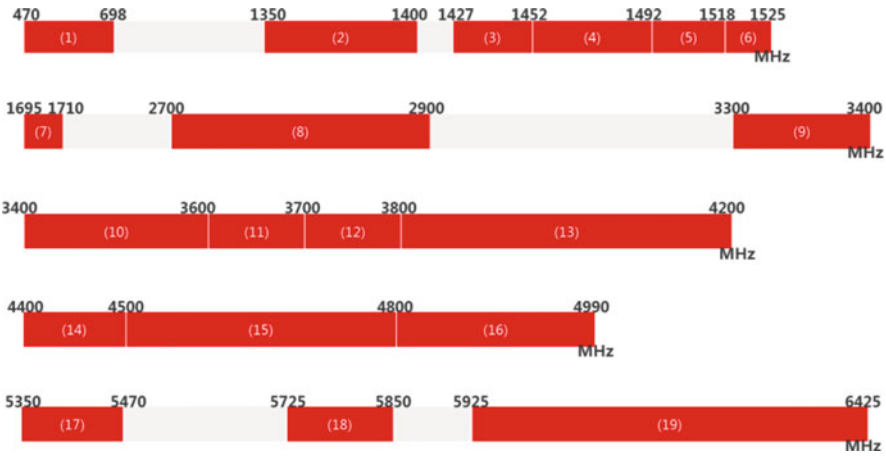


**Fig. 6** Possible candidate bands in WRC-15 Agenda Item 1.1 before the conference

**Table 4** Views on suitable frequency ranges/bands

| | | Views[a] | | | | | |
|---|---|---|---|---|---|---|---|
| Frequency ranges (MHz) | | EU | US | Russia | Japan | Korea | China |
| 1 | 470–694/698 | s | √ | × | × | √ | |
| 2 | 1350–1400 | × | | × | | | |
| 3 | 1427–1452 | √ | × | × | √ | | |
| 4 | 1452–1492 | √ | × | × | √ | √ | |
| 5 | 1492–1518 | √ | × | × | √ | | |
| 6 | 1518–1525 | × | × | × | √ | | |
| 7 | 1695–1710 | × | √ | | | | × |
| 8 | 2700–2900 | × | s | | | | × |
| 9 | 3300–3400 | × | | | | | √ |
| 10 | 3400–3600 | √ | s | × | √ | √ | × |
| 11 | 3600–3700 | √ | s | × | √ | √ | × |
| 12 | 3700–3800 | √ | s | × | √ | √ | × |
| 13 | 3800–4200 | × | s | × | √ | √ | × |
| 14 | 4400–4500 | × | | | √ | | √ |
| 15 | 4500–4800 | × | | | √ | | × |
| 16 | 4800–4990 | × | | | √ | √ | √ |
| 17 | 5350–5470 | × | √ | | | | |
| 18 | 5725–5850 | s | | | | | |
| 19 | 5925–6425 | s | | √ | | | × |

[a] √: support, ×: oppose, s: study

Spectrum below 6 GHz are the best resources for IMT in the near future. However, due to its scarcity and increasingly difficulty to realize international harmonization, it is the time to seek spectrum above 6 GHz.

## *4.2 New Spectrum After WRC-15, Especially the Bands Above 6 GHz*

Traditionally, spectrum usage above 6 GHz such as millimeter wave (mmWave) often applies for point-to-point communications in large power systems such as satellite systems and microwave systems.

From the perspective of public mobile communication, mmWave communication technology also attracts peoples' attention. Samsung from Korea conducted studies and tests for the above 6 GHz, especially at 13.4–14 GHz, 18.1–18.6 GHz, 27.0–29.5 GHz and 38.0–39.5 GHz. At 28 GHz, deploying 64 antennas, Samsung has done the experiments of the adaptive beam-forming technology to implement 1 Gbps peak rate within 200 m, and its error rate is less than 0.01 % [16]. Moreover, at 11 GHz, NTT DoCoMo implemented prototype for high-speed mobile vehicle communications. In addition, manufacturers such as Ericsson, Huawei, and

NSN show their vision for mmWave communications on various bands such as 14 GHz, E-band (71–76 GHz/81–86 GHz) and so on. It is obvious that mmWave communication is drawing attentions all over the world as one of the potentially essential technologies of the next generation of mobile communication systems.

The 60 GHz International Broadband Wireless Access Systems (BWAS) are drawing peoples' attention. It mainly focuses on short-distance, high-rate communications. The transmission distance is usually within 10 m, and TDD mode is applied. The specific operating frequency is 56–66 GHz (59–64 GHz in China). The three related international standards are ECMA 387, IEEE 802.15.3c and IEEE 802.11ad. In addition, IEEE establishes IEEE 802.11aj to formulate the specific criteria of 60 and 45 GHz usage in China.

In 2001, FCC allotted a continuous section of spectrum (57–64 GHz) for wireless communication in unlicensed mode [17]. In Europe, ECC also allotted 57–66 GHz unlicensed spectrum. Specifically, 62–63 GHz and 65–66 GHz were for Mobile Broadband System, and 59–62 GHz were for Wireless Local Area Network (WLAN) [18]. In 2006, 59–64 GHz was planned in China for wireless short-distance technologies. Similarly, Japan and Australia conducted the related spectrum planning in this frequency band. The detail procedure is shown in Fig. 7. In 2013, China allotted 40.5–42.3 GHz band and 48.4–50.2 GHz band for fixed point-to-point wireless access system (light license management), 42.3–47 GHz band and 47.2–48.4 GHz band for mobile point-to-point wireless access system (unlicensed management).

However, due to the characteristics of high frequency, the key technologies applied for low frequency are hard to directly put into use. How to utilize the advantage and overcome the disadvantage are both opportunities and challenges [19].

- Principles for candidate bands selection [20]

In the Table of Frequency Allocations, spectrum above 6 GHz are mainly allocated to fixed service, mobile services, radio location service, radio navigation service, fixed satellite service, inter-satellite services and broadcasting satellite service, etc. In practical use, for example, there are digital microwave relay systems, navigation

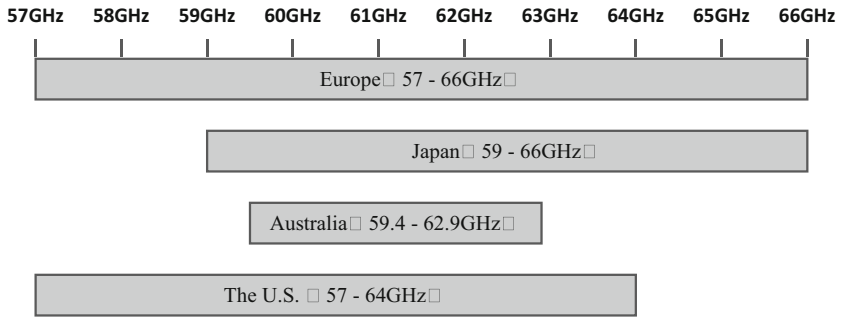


**Fig. 7** International spectrum allocation of 60 GHz

and air traffic control radar system and satellite communication systems. High frequency radio communication has different characteristics in radio propagation antenna and RF, compared with lower frequency communication. For candidate bands selection for mobile communications, the following principle could be considered.

**Compliance** Bands allocated to Mobile Services (or in footnotes identified) are preferred.

**Consistency** To select potential candidates for international harmonization, to take into account the future promotion in WRC.

**Security** Fully consideration on the electromagnetic compatibility between systems, to ensure the protection of mobile systems as well as other systems.

**Continuous** To take the advantage of broadband (e.g., >500 MHz) from high frequency to guarantee that the system can obtain a higher efficiency.

**Effective** Taking into account the propagation characteristics of high frequency and industrial hardware manufacturing capability, to select the appropriate frequency to ensure the effective design and realization of the system, terminal, instrument, etc.

- Some potential frequency ranges for 5G above 6–100 GHz

Preliminary studies from different administrations or organizations show that a lot of potential suitable frequency ranges could be found from 6 to 100 GHz. This sub-section summaries some considerations from research institutes, countries and regional organizations.

### 4.2.1 METIS

The METIS project has delivered study results on potential suitable frequency ranges [21] in terms of a prioritization of the bands as shown in Figs. 8 and 9,

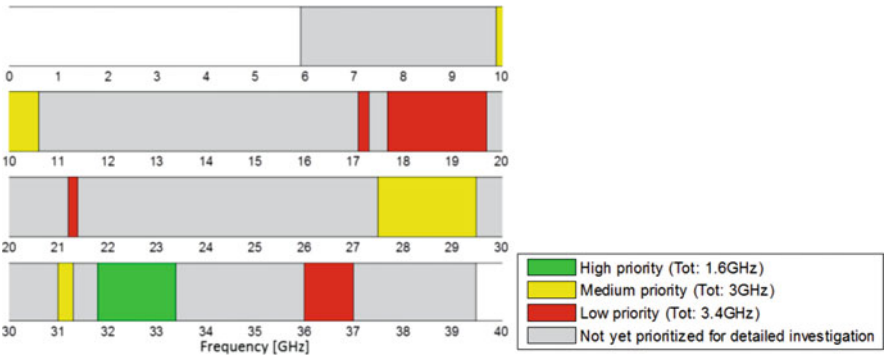


**Fig. 8** Opportunities of potential sharing with current incumbent within 5.925–40.5 GHz [21]

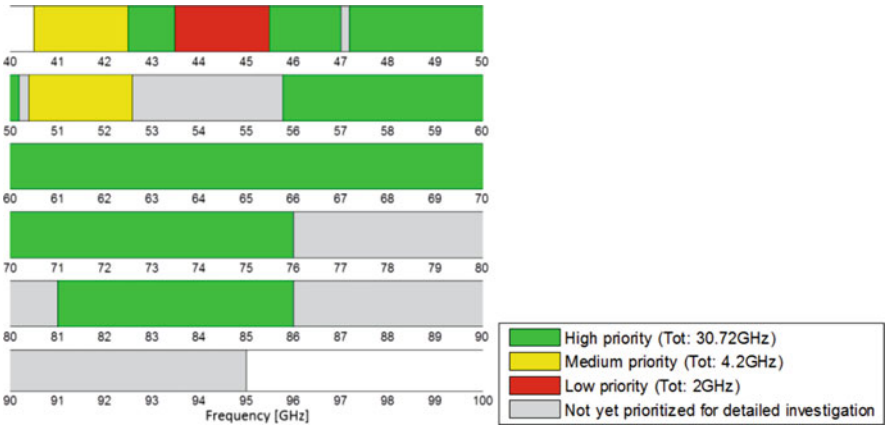


**Fig. 9** Opportunities of potential sharing with current incumbent within 40.5–100 GHz [21]

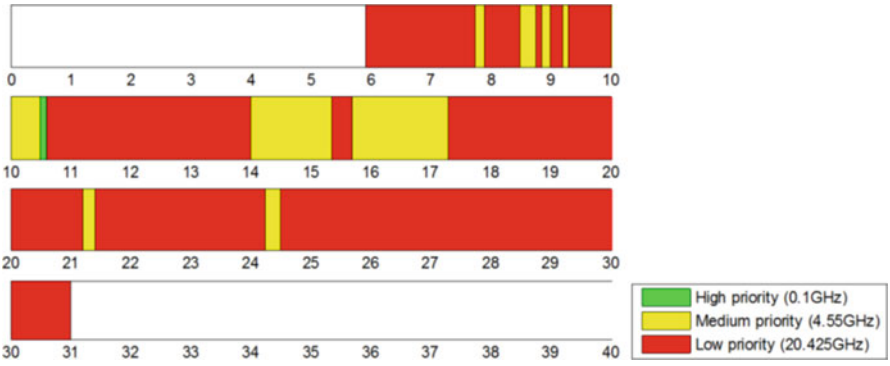


**Fig. 10** Opportunities of potential sharing with current incumbent for outdoor scenario [21]

indicating the most relevant bands for initial consideration in regulation when starting the procedure to find additional bands for 5G systems considering only current allocations.

To further shed light on the opportunities in the 5.925–31 GHz range, an additional band assessment with modified assessment criteria has been performed, as shown in Fig. 10. In this assessment, a reduced target minimal contiguous bandwidth of 60 MHz was used, and a larger focus was given to outdoor deployments.

### 4.2.2 Ofcom [22, 23]

For spectrum above 6 GHz, the Ofcom started with a national Call For Input (CFI) for the potential candidate bands and summarized the response as shown in Table 5.

**Table 5** Ofcom 5G candidate frequency bands above 6 GHz

| Frequency ranges | 6–20 GHz | 20–40 GHz | 40–60 GHz | | 60–100 GHz |
|---|---|---|---|---|---|
| Candidate bands | 10.125–10.22510.475–10.575 | 31.8–33.4 | 40.5–43.5 | 45.5–48.9 | 66–71 |

**Table 6** FCC 5G candidate frequency bands above 6 GHz

| Frequency ranges | Candidate bands (GHz) |
|---|---|
| 24 GHz | 24.25–24.45, and 25.05–25.25 |
| LMDS | 27.5–28.35, 29.1–29.25 and 31–31.3 |
| 39 GHz | 38.6–40 |
| 37/42 GHz | 37–38.6 and 42–42.5 |
| 60 GHz | 57–64 and 64–71 |
| 70/80 GHz | 71–76 and 81–86 |

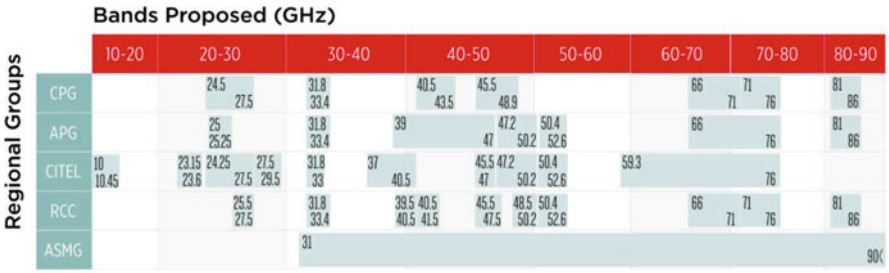


**Fig. 11** Views from different regional organizations on identification of frequency bands for IMT in portion(s) of the frequency range between 6 and 100 GHz

### 4.2.3 FCC [24]

Similarly, FCC started with a Notice Of Inquiry (NOI) the potential candidate bands and only consider bands above 24 GHz as shown in Table 6.

### 4.2.4 Regional Organizations

Up to Oct. 2015, several regional organizations have made preliminary views for WRC-15 Agenda Item 10. Almost all the organizations have agreed to establish a new agenda item for WRC-19, to consider identification of frequency bands for IMT in portion(s) of the frequency range between 6 and 100 GHz. These views are shown in Fig. 11.

From the regulatory perspective, it is vital to perform solid research of these bands, including channel measurements, system modeling and detailed compatibility studies with currently used services [25].

### 4.2.5 Provisional Final Acts of WRC-15

After WRC-15, a new Agenda Item 1.13 is established in WRC-19, to consider identification of frequency bands for the future development of IMT, including possible additional allocations to the mobile service on a primary basis [15]. The related frequency bands are:

– 24.25–27.5 GHz2, 37–40.5 GHz, 42.5–43.5 GHz, 45.5–47 GHz, 47.2–50.2 GHz, 50.4–52.6 GHz, 66–76 GHz and 81–86 GHz, which have allocations to the mobile service on a primary basis; and
– 31.8–33.4 GHz, 40.5–42.5 GHz and 47–47.2 GHz, which may require additional allocations to the mobile service on a primary basis.

## *4.3 Spectrum for IoT*

Compared with previous generations of mobile communications, 5G needs to meet extremely high performance requirements in more diverse scenarios. Besides scenarios such as seamless wide-area coverage and high-capacity hot-spot, there is also a need of low-power massive-connections including machine-to-machine and man-to-machine, driven by internet of things (IoT).

Low-power massive-connection scenario mainly targets sensing and data collecting use cases, such as environmental monitoring and intelligent agriculture. This scenario is characterized by small data packets, low power consumption, low cost, and massive connections. Specifically, probably at least one million connections per squared kilometer need to be supported [26].

The suitable frequency ranges of such 5G usage scenarios will be wide and rather different with those for communication between people. An example is shown in Table 7, which is the list the frequency bands used for smart metering application in Power Grid Management Systems in some parts of the world [27].

When considering certain spectrum for Short Range Devices (SRDs) as one part of 5G, e.g., RFID, the 5G potential frequency ranges will be much wider.

# 5 Spectrum Management Considerations

New scenarios, service types, spectrum demand and usage methods of 5G will definitely bring new challenges to current spectrum management. In this section, some new policies and policy related studies on spectrum management for mobile communications from a number of countries are introduced, reflecting the future 5G spectrum management development trends.

**Table 7** Example of frequency bands in wireless metering in Power Grid Management Systems [27]

| Frequency (MHz) | Area/region |
|---|---|
| 40–230 (part of), 470–694/698 | North America, UK, Europe, Africa, and Japan |
| 169.4–169.8125 | Europe |
| 220–222 | Some parts of ITU Region 2 |
| 223–235 | China |
| 410–430 | Parts of Europe |
| 450–470 | North America, parts of Europe |
| 470–510 | China |
| 470–698 | North America and Europe |
| 779–787 | China |
| 868–870 | Europe |
| 870–876 | Parts of Europe |
| 896–901 | North America |
| 901–902 | North America |
| 902–928 | North America, South America, Australia |
| 915–921 | Parts of Europe |
| 917–923.5 | Korea |
| 920–928 | Japan |
| 928–960 | North America |
| 950–958 | Japan |

To be clear, some new ideas mentioned in this chapter do not represent the current policy and are considered as research work aiming to give some food for thought. But from various aspects of the related work, they could be a possible trend.

## *5.1 US*

Foreseeing the rapid development of mobile communications, US pay high attention to the related spectrum resources management, with carefully planned steps to carry out a series of related work.

In June 2010, the Presidential Memorandum entitled of Unleashing the Wireless Broadband Revolution was released. Accordingly, the National Telecommunications and Information Administration (NTIA) collaborating with the Federal Communications Commission (FCC) are planned to make 500 MHz of Federal and nonfederal spectrum available for wireless broadband use within 10 years.

Furthermore, in July 2012, the report from PCAST (President of the Science and Technology Advisory Committee) to the US President propose to released 1000 MHz band to “create of spectrum super-highway” [2].

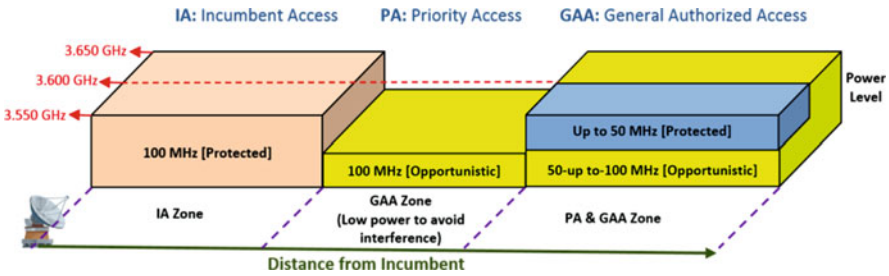


**Fig. 12** Proposed Licensing Model for 3.5 GHz (source: http://www.ischool.pitt.edu/)

In June 2013, based on the above report, the Presidential Memorandum entitled of Expanding America's Leadership in Wireless Innovation was released. A Spectrum Policy Team was set up to seek how to bring spectrum sharing measures into current spectrum management of NTIA and FCC for a study period of 1 year.

July 11, 2014, NTIA and the FCC issued a joint notice for comments on Model City for demonstrating and evaluating advanced spectrum sharing technologies recommended by PCAST. More than a dozen replies are received from AT&T, Wireless Innovation Forum, Dynamic Spectrum Alliance, Telecommunications Industry Association, etc.

When seeking the candidate bands to release 1000 MHz, the 3550–3700 MHz band are selected as a possible one. On April 2014, FCC proposed Rules to make up this 150 MHz of spectrum available for new citizens broadband radio service. This innovative rule significantly promotes the intensive shared use of spectrum in the 3.5 GHz Band.

On April 2015, the FCC finally adopted rules for the Citizens Broadband Radio Service, creating a new spectrum band and taking a major step forward in spectrum policy by authorizing advanced spectrum sharing among commercial and federal operators, as shown in Fig. 12.

## *5.2 EU*

Recently, ECC states that License Shared Access (LSA) is a complementary spectrum management tool that fits under an "individual licensing regime" [28] and the national legislation framework is shown in Fig. 13. National Tables of Frequency Allocations (NTFAs) primarily specify the radio services authorized by an individual administration in frequency bands and the entities which have access to them. National frequency assignments, as derived from the ITU concept, allows the fine management of frequency bands in accordance with the rules set in NTFAs, particularly in bands shared by different type of users and also in respect of coexistence issues in adjacent bands. They may contain sensible data and their management requires confidentiality procedures. Under this, two different ways

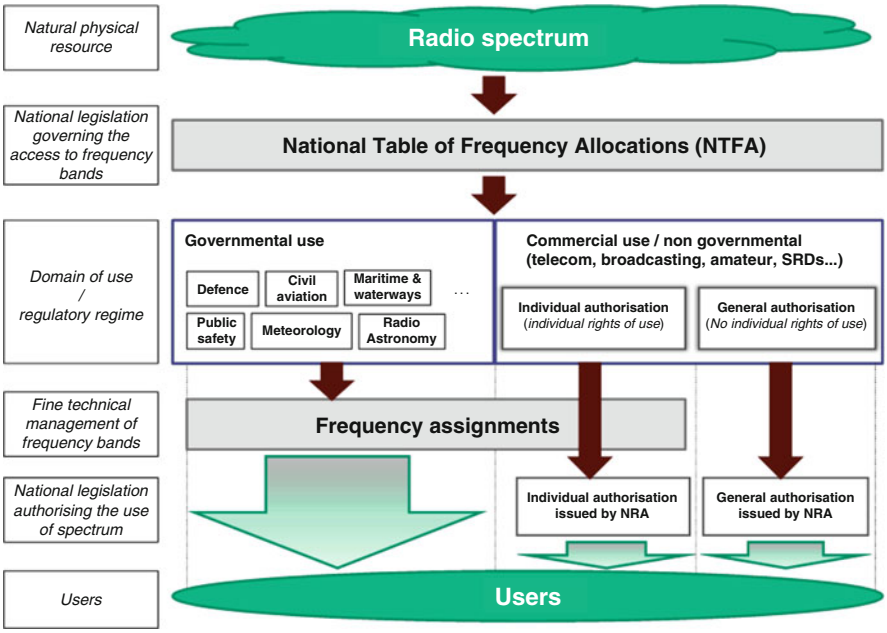


**Fig. 13** National legislation from the radio spectrum to users [28]

of using spectrum are introduced: Government use covering various domains (e.g. defense, civil aviation, maritime & waterways, public safety, meteorology, science), and commercial use understood as the public legal act issued by NRAs for the purpose of delivering spectrum usage rights to private entities or citizens. The following two terminologies should also be distinguished for commercial use:

- Individual authorization (Individual rights of use);
- General authorization (No individual rights of use).

Individual rights of use are given for limited duration and do not constitute property act of the frequencies by the operator as frequencies are part of the national domain.

## *5.3 China*

At present, the spectrum license pattern in China mainly includes both licensed and unlicensed use. In addition, there is a light license mode, in which the user only needs to report but not apply for setting up a new radio station to the government for recording.

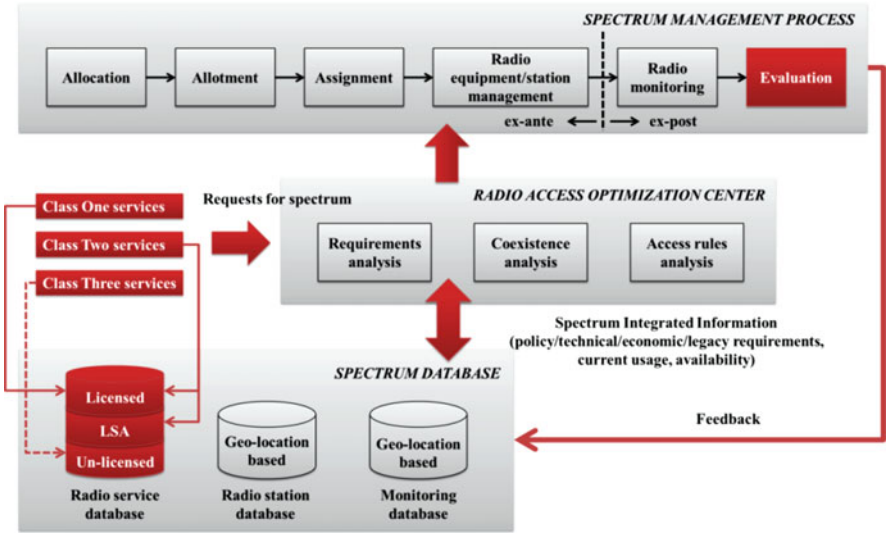


**Fig. 14** New elements studied in current spectrum management framework in China [4]

In such a management framework, spectrum sharing is not well supported. As 5G brings much more challenges in spectrum demand and also new spectrum usage methods, the spectrum management has to keep pace with such innovation. Currently based on the current framework of spectrum management in China shown in Fig. 14, two new elements are being studied, namely services classification and spectrum assessment.

The *basic spectrum management process* is shown at the top of Fig. 14. The spectrum allocation (to radio services), allotment (to users or systems) and assignment (to specific radio stations) as well as the radio equipment and station management belong to ex-ante management. The radio monitoring belongs to ex-post management, including radio occupancy measurements, signal parameter and transmitter inspection, illegal transmitter detection and finding and so on [http://www.srrc.org.cn/english/]. The *radio access optimization center*, composed of relevant organizations and institutions, is in charge of spectrum demand calculation, coexistence and access analysis. The *spectrum database*, established in State Radio Monitoring Center and State Radio Spectrum Management Center (SRMC/SRSMC), stores all the information of radio services, stations and monitoring reports, which indicate the current spectrum usage and availability. When a new service access request arrives, the *center* checks the *database*, performs integrated analysis and outputs the solution to support the *process*, which conversely inputs the information to the *database* for further use, establishing a closed-loop management.

- Services classification

Intuitively, not every radio service can be shared with IMT systems. Therefore, it is vital to make clear which types of service can be put forward for sharing. Here, it is proposed to divide them into three classes.

Class One services are those involving government affairs, national security and people's safety. They usually occupy specified licensed spectrum for free and need strict protection. Their spectrum cannot be shared with others.

Class Two services include commercial radio services, and also general public and dedicated services. The spectrum for public mobile communication in 5G is included. Licensed or LSA mode could be used here. The government would flexibly charge the Class Two services.

Class Three services are dedicated for public free use. The access process might also need spectrum sensing, such as WiFi.

- Spectrum assessment

Spectrum assessment belongs to ex-post management. It is not only the basis for whether a frequency band can be allotted to a new service, but also a way of supervising its utilization efficiency. In [4], the assessment of the current spectrum usage is proposed in China for the first time. Based on assessment results, the government can adjust or even retrieve the spectrum allotment.

To make fair assessment, a scientific and effective Key Performance Indicator (KPI) system is required to be established. The KPI currently under studied mainly includes three aspects.

– Radio monitoring related information, such as noise and radio occupancy measurement results in different scenarios and locations.
– Radio stations related information, such as location, RF parameters and also the related statistics.
– Service related information, e.g., different Quality of Service (QoS) and protection requirements.

The KPI system is necessary for making reasonable conclusions in spectrum management. It is important to study the feasibility of providing such information in a technology- and service-neutral way, and whether it allows for a practical and efficient spectrum evaluation process.

## 6 Summary

Different aspects of 5G spectrum issues are discussed in this chapter. Generally, it is expected that 5G requires more in the total amount, wider with respect to individual bandwidths, greater in the range and more flexible in the usage and management pattern.

The suitable frequency ranges of 5G will include those bands below 6 GHz such as re-farmed 2G/3G spectrum, identified frequency bands for IMT and also WRC-15 candidate bands. However, due to the scarcity of spectrum below 6 GHz, which

has almost been fully utilized and it will be even more difficult to find internationally harmonized spectrum after WRC-15, it is necessary to seek potential frequency ranges above 6 GHz.

The controlled spectrum sharing is an important way of re-using spectrum to complement current licensed dedicated spectrum, which is still the basis for operation of 5G systems.

**Acknowledgments** This chapter is supported by Chinese National Key Project under Grant No. 2014ZX03001027 and No. 2015ZX03002008, National High Technology Research and Development Program ("863" Program) of China under Grant No. 2014AA01A707, and Beijing Natural Science Foundation(4132050).

## References

1. ICT-317669 METIS project, Intermediate description of the spectrum needs and usage principles. Deliverable D5.1, April 2013, https://www.metis2020.com/documents/deliverables
2. G. Documents, *Realizing the Full Potential of Government-Held Spectrum to Spur Economic Growth - PCAST - July 2012* (IEEE USA Books & eBooks, 2012)
3. Ofcom, Spectrum management strategy - Ofcom's strategic direction and priorities for managing spectrum over the next 10 years, April, 2014
4. T. Wang, G. Li, J. Ding, Q. Miao, J. Li, Y. Wang, 5g spectrum: is China ready? Commun. Mag. IEEE **53**, 58–65 (2015)
5. AWG-18/TMP-18, A Draft Revision of APT Report on Information of Mobile Operators' Frequencies, Technologies and License Durations in Asia Pacific Countries, Kyoto, Japan, March 2015
6. Electronic Communications Committee (ECC), The European Table of Frequency Allocations and Applications in the Frequency Range 8.3 kHz to 3000 GHz (ECA Table), May 2014
7. Report ITU-R M.2290-0, Future spectrum requirements estimate for terrestrial IMT, Dec. 2013
8. Federal Communications Commission, Mobile Broadband: The Benefits of Additional Spectrum, Oct. 2010
9. ITU-R Document 5D/66, Relevant Work Undertaken by Australia Regarding the Potential Identification of Additional Spectrum for IMT, July 2012
10. ITU-R Document 5D/118, Future IMT Spectrum Requirements Assessment for the Russian Federation, Sep. 2012
11. H. Biao, W. Tan, Spectrum requirements calculation based on Urban Hotspot in China for mobile communication in the year 2020. ZTE Technol. J. **20**(2), 5–10 (2014)
12. S. Chen, J. Zhao, The requirements, challenges and technologies for 5G of terrestrial mobile telecommunication. IEEE Commun. Mag. **52**(5), 36–43 (2014)
13. MIIT of China, Regulations of Radio Frequency Division of People's Republic of China, Feb. 2014
14. ITU-R, Report of the CPM on operational and regulatory/procedural matters to the World Radio Communication Conference, 2015
15. Provisional Final Acts WRC-15, Geneva, http://www.itu.int/pub/R-ACT-WRC.11-2015
16. W. Roh et al., Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results. IEEE Commun. Mag. (2014)
17. FCC, Code of Federal Regulation, Title 47 Telecommunication, Chapter 1, Part 15.255
18. ETSI DTR/ERM-RM-049, Electromagnetic compatibility and radio spectrum matters (ERM); system reference document; technical characteristics of multiple gigabit wireless systems in the 60 GHz range. Mar. 2006

19. S. Chen, J. Zhao, M. Ai, D. Liu, Y. Peng, Virtual RATs and a flexible and tailored radio access network evolving to 5G. IEEE Commun. Mag. **53**(6), 52–58 (2015)
20. APG, APG15-5/OUT-40: Preliminary APT Common Proposals on WRC-15 Agenda Item 10, July 31st 2015
21. METIS deliverable D5.3, Description of the spectrum needs and usage principles, Aug. 2014
22. Ofcom, Spectrum above 6 GHz for future mobile communications: Call for Input, Jan. 2015
23. Ofcom, Laying the foundations for next generation mobile services: Update on bands above 6 GHz, April 2015
24. FCC, Notice of Inquiry, Oct. 2014
25. Y. Wang, J. Xu, L. Jiang, Challenges of system-level simulations and performance evaluation for 5G wireless networks. IEEE Access **2**, 1553–1561 (12/2014)
26. IMT-2020 (5G) Promotion Group, 5G Wireless Technology Architecture White Paper, May 2015
27. ITU-R Working Party 1A, REPORT ITU-R SM.2351-0 Smart grid utility management systems, June 2015
28. ECC Report 205, licensed shared access, Feb. 2014

# Spectrum Sharing for 5G

**Gen Li, Tan Wang, Qingyu Miao, Ying Wang, and Biao Huang**

**Abstract** Spectrum sharing for 5G is motivated by the fact that significantly more spectrum and much wider bandwidths than what is available today will be needed in order to realize the performance targets of 5G. First, spectrum sharing scenarios, i.e. vertical sharing and horizontal sharing, are summarized for different type of spectrum. Second, a thorough review of current spectrum sharing techniques are provided including coordination protocol, GLDB, DSA and MAC-based collision avoidance. Then spectrum sharing is analyzed specific to 5G design and one general architecture to enable 5G spectrum sharing is proposed. Finally, it is concluded that spectrum sharing becomes more and more important for 5G systems as a complementary way of using spectrum.

## 1 Introduction

Currently, traditional 2G, 3G and 4G cellular networks have so far been designed for dedicated, licensed spectrum. With that, cellular technology can only operate in a subset of the spectrum that is in principle available. More recently, other forms of shared spectrum including unlicensed spectrum is close to application in near future for 4G systems, i.e. licensed assisted access (LAA) LTE which is ongoing work in 3GPP. Significantly more spectrum and much wider bandwidths than what is available today will be needed in order to realize the performance targets of 5G, i.e. more system capacity and better user experience. One way is to try the best to fill the spectrum gap by finding new spectrum, especially going to higher frequency band which is less crowded than what is being used today. The other way is to improve

G. Li (✉) • Q. Miao
Ericsson, China Radio Research Lab, Beijing, China
e-mail: gen.li@ericsson.com

T. Wang • B. Huang
State Radio Monitoring Center, MIIT, Beijing, China

Y. Wang
Beijing University of Posts and Telecommunications (BUPT), Beijing, China



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_3

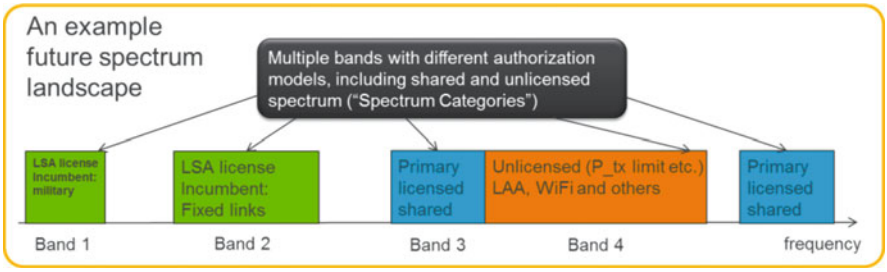


**Fig. 1** Example future spectrum landscape [1]

total spectrum efficiency by spectrum sharing. Generally speaking, it is expected that the spectrum for 5G will, at least partly, be available under spectrum sharing regimes.

As an example future spectrum landscape one can envision the situation depicted in Fig. 1. Therein the 5G system is able to access primary and licensed bands as well as other bands in an LSA fashion or unlicensed bands. The typical choice would be to go with the primary band for reliability, but local interference situations could make it more beneficial for some nodes to use other bands with other regulatory frameworks.

This chapter first introduces various spectrum sharing scenarios corresponding to different type of spectrum in Sect. 2. Then Sect. 3 presents a thorough review of current spectrum sharing techniques, including coordination protocol, GLDB, DSA and MAC-based collision avoidance. Next in Sect. 4, it also analyzes the limitations of the aforementioned techniques and proposed general architecture and the corresponding research directions in 5G. Finally Sect. 5 concludes this chapter.

## 2 Spectrum Sharing Scenario

It is well known that current 2G, 3G and 4G systems only use licensed dedicated spectrum. However, according to the analysis in this chapter, 5G systems will need significantly more spectrum than today and it is hard to find enough bands. Thus using only licensed mode may not meet 5G requirement very well. This section provides a complete summary of spectrum sharing scenarios as the following categories [1, 2]:

- *Vertical sharing* refers to spectrum sharing between users of different priority (e.g. primary and secondary), i.e., unequal rights of spectrum access.
- *Horizontal sharing* is sharing between systems that have the same priorities in the spectrum, i.e. different users have fair access rights to the spectrum. If the sharing users in the spectrum adopt the same technology, it is called *homogenous horizontal sharing*; otherwise *heterogeneous horizontal sharing*.

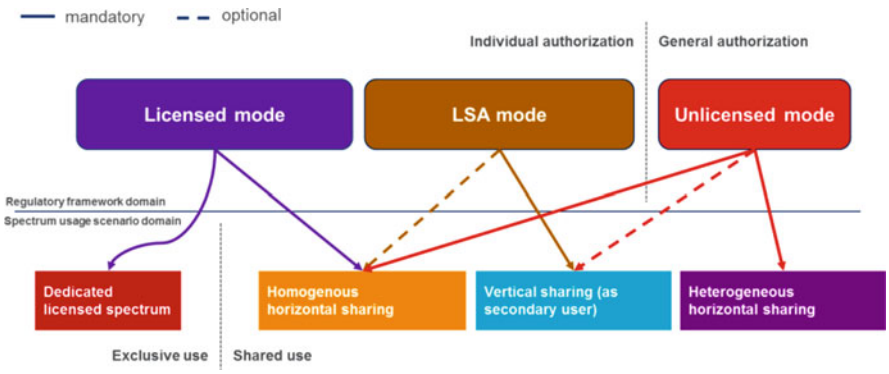


**Fig. 2** Illustration of spectrum type and related usage scenario [2]

First, as shown in Fig. 2, the above-mentioned spectrum sharing scenarios are mapped into different types of spectrum, i.e. licensed, licensed shared access (LSA) and unlicensed as follows:

## *2.1 Licensed Mode*

In licensed mode, spectrum is allotted to wireless systems as primary use. The only relevant spectrum sharing scenario is homogenous horizontal sharing, i.e. sharing spectrum with other operators using the same RAT, which is also called co-primary or inter-operator spectrum sharing. There are two different ways to achieve horizontal sharing, i.e. mutual renting and limited spectrum tool.

- In the *mutual renting* sub-scenario the spectrum resources in a band are subdivided into several blocks and each block is licensed to one operator. Operators mutually allow other operators to "rent" parts of their licensed resources. An operator can rent resources from multiple other operators simultaneously. The actual "owner" of a resource has always strict priority in accessing its licensed part of the spectrum, including the possibility of preemption at any time. This approach could be applied to bands that have initially been used in a dedicated licensed way where there is a need to increase peak data rate beyond what is possible within one licensed block.
- The *limited spectrum pool* scenario allows an operator to obtain an authorization, usually a license, to use up to the whole band on a shared basis with a limited number of other known authorized users. This setup does not provide guarantee for instantaneous access to a minimum amount of spectrum, but it is envisioned that mutual agreements between licensees are such that the long term share of an individual operator has a predictable minimum value. This is similar to

the sharing situation in unlicensed bands but it is a priori known how many authorizations a regulator will give out and what the rules for the band will be, hence providing the necessary certainty for investments into large networks.

## 2.2 LSA Mode

In LSA mode, a licensee has the authorization to access spectrum that is unused by an incumbent user at certain locations and/or times. This *vertical sharing* is based on well-defined conditions which are parts of a sharing license. In the initially developed LSA concept, the licenses are expected to be long term and exclusive for one system or operator only. As it evolves, the LSA concept may in the longer term be combined with *homogenous horizontal sharing* so that the issued licenses could be non-exclusive and require several licensees to coexist.

LSA is currently being developed through the European Commission's Radio Spectrum Policy Group (RSPG), which issued a report on LSA in November 2013 with recommendations to member countries. Now ECC FM 52 has essentially concludes this LSA regulation work and only the final round of ECC approval remains. CEPT ECC has started a study to use the 3.6–3.8 GHz band for LSA in order to enable sharing between FSS and mobile systems. Standardization activities on LSA technical conditions and geo-location databases are ongoing at ETSI (the European Telecommunications Standards Institute), where they are being carried out by the Technical Committee on Reconfigurable Radio Systems (RRS).

## 2.3 Unlicensed Mode

In unlicensed mode, a wireless system has to share spectrum with other unlicensed systems. For *heterogeneous horizontal sharing* in an unlicensed band a system must be prepared for coexistence with any other technology that may be present in the band [3]. Meanwhile, it is unavoidable to have *homogeneous horizontal sharing* between different systems with the same RAT. *Vertical sharing* functionality may also be required for a system operating in unlicensed mode if a primary user exists in the band. Therefore the complexity to implement spectrum sharing is more complicated and seems non-reliable compared to others since it almost involves all kinds of sharing scenario.

Furthermore, due to the usual restrictions on maximum transmit power in unlicensed bands; building standalone contiguous coverage areas for mobility in unlicensed spectrum requires a very large number of sites. Unrestricted access to the bands also creates unpredictable interference conditions that typically lead to a decrease in accessible capacity when traffic levels are high. The result is high costs and intense operational complexity. License-exempt spectrum creates a low barrier to market entry. This has market advantages but also causes issues with controlling

the numbers of users and operators, and in turn with controlling the quality of access and service. It is also difficult to repurpose license-exempt spectrum to adapt to changing regulatory needs, causing trouble in terms of long term spectrum management.

## 3 Spectrum Sharing Techniques

There are several different ways to technically realize the above spectrum sharing and to implement corresponding regulatory requirements. As illustrated in Fig. 3, four different kinds of spectrum sharing techniques are mapping to the most possible application scenarios. In this section we outline a few of the most common or promising techniques.

### *3.1 Coordination Protocol*

Coordination protocol means explicit information exchange between the sharing systems via some well-defined interface to avoid interference in sharing spectrum. The protocol defines the behavior of the nodes when receiving certain messages or when certain events take place. This includes the possibility for defining behavior depending on what relation nodes have with each other, e.g. interference relation. Both centralized and distributed structure can be applied to coordination as illustrated in Fig. 4. Here the connections between networks or between network and central manager can be backbone wired backhaul or over the air. The advantage of this approach is that it can achieve more efficient spectrum sharing with enough information on the interference between different networks by employing smart resource allocation. One obvious drawback is the mandatory need of connection between different networks or to central manager for multiple networks, which may not be easy to have in most scenarios.

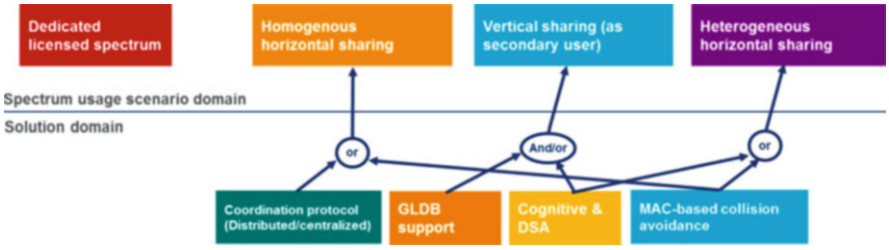


**Fig. 3** Summary of mapping from spectrum sharing techniques to scenarios

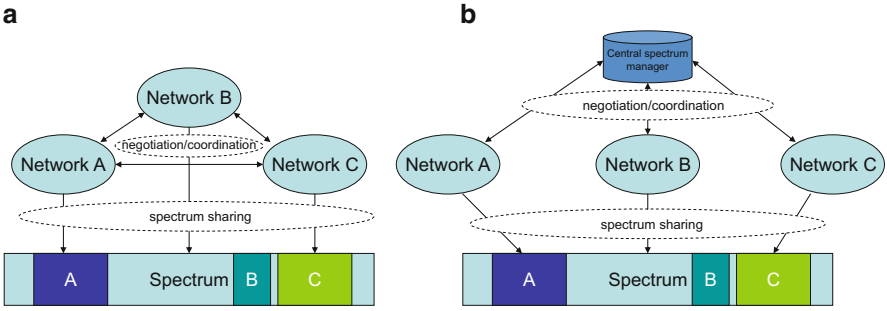


**Fig. 4** Coordination structure: (**a**) distributed coordination; (**b**) centralized coordination

An example for such an approach is the Coordinated Coexistence Protocol of IEEE 802.16h [4]. 802.16h employs distributed structure with over-the-air connection. When a base station (BS) needs to send coordination messages to its neighbor BS, it may select one or more user equipments (UE) as forwarders and send the message over the air. The detailed procedures are listed as follows:

- System selects the forwarding UE according to the Radio Signal Strength Indicator (RSSI) of neighbor BS detected by the interference victim SSs, i.e. the SSs with higher neighbor BS RSSI is better;
- After identifying the forwarding UE, the serving BS requests access to the neighbor BS through the forwarding UE;
- Upon receiving the above access neighbor BS request from serving BS, the forwarding UE will try to access the neighbor BS following a procedure similar to the normal network entry process;
- If the forwarding UE accesses the neighbor BS successfully, it shall respond to its serving BS with a message that includes the parameter "UE has successfully accessed the requested neighbor BS." The serving BS and its neighbor BS can communicate with each other via the forwarding UE.

After connection establishment, resource is coordinated between different systems. IEEE 802.16h specifically defines a coexistence frame structure used to implement multiple systems coexist as shown in Fig. 5. A coexistence frame consists of four IEEE 802.16 MAC frame (MAC frame is called subframe here), and is divided into the following types of subframes: master subframe, slave subframe and shared subframe. Each of the coexistence system is required to obtain a subframe as its master frame. Other coexisting systems claim slave frame on this subframe. It is worth mentioning that each subframe can only be the master frame by only one coexistence system. In other networks, if a subframe is announced as master frame for one system, this subframe can only be slave subframes of other systems. Systems can use the maximum power for transmission in the period of master frame. Within slave subframe, systems cannot transmit at all or only be allowed to transmit following restrictions by master systems.

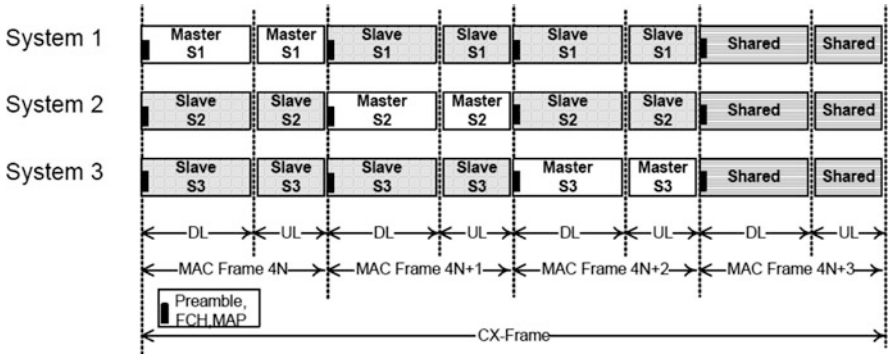


**Fig. 5** Coexistence frame structure in 802.16h network [4]

## *3.2 Geolocation Database (GLDB) Support*

In this solution the system queries a data base to acquire information on what resources are available (i.e., unused by other systems of typically higher priority) for operation in its current location. The geo-location data base provides rules for operating in the geographical region of the system. A rudimentary geo-location data base would provide a list of available channels and possibly associated transmits power (EIRP) limitations to a querying system for a specified geographical region.

The GLDB approach has been considered as an appropriate solution to manage unlicensed access to TV white spaces while ensuring that the bands may, in the future, be vacated of unlicensed transmissions. In this case this solution does not provide exclusive use of the spectrum to any system. The presence of a centrally controlled resource manager opens up for the possibility of future repurposing of the bands, since the behavior of the radios the database controls can be influenced, including the possibility to prevent transmissions completely. In this way a band could be opened up for usage not compatible with the database-controlled devices. For example, 802.22 use such approach to enable coexistence with TV systems. It is assumed that all devices must be equipped with satellite technology (e.g. GPS) which provide positioning and also gives time reference for inter-BS synchronization. Before communication, UE needs to inform BS the location and possibly antenna pattern, height and etc. Then BS queries database for allowed channels and powers. Finally BS will communicate with corresponding UE with received configuration limit.

GLDB is also seemed as a good solution to achieve vertical sharing in LSA regime [5]. Typical examples are government and military spectrum subject to sparse use, either in terms of geographic coverage or temporal characteristics. One simple LSA GLDB structure is illustrated in Fig. 6. The LSA controller in LSA licensee retrieves the information about spectrum made available under the LSA scheme from the LSA GLDB of the incumbent spectrum user and therefore has information on the available spectrum in any given location where its base stations are sited. Under this scheme, the base stations are allowed to transmit in the LSA

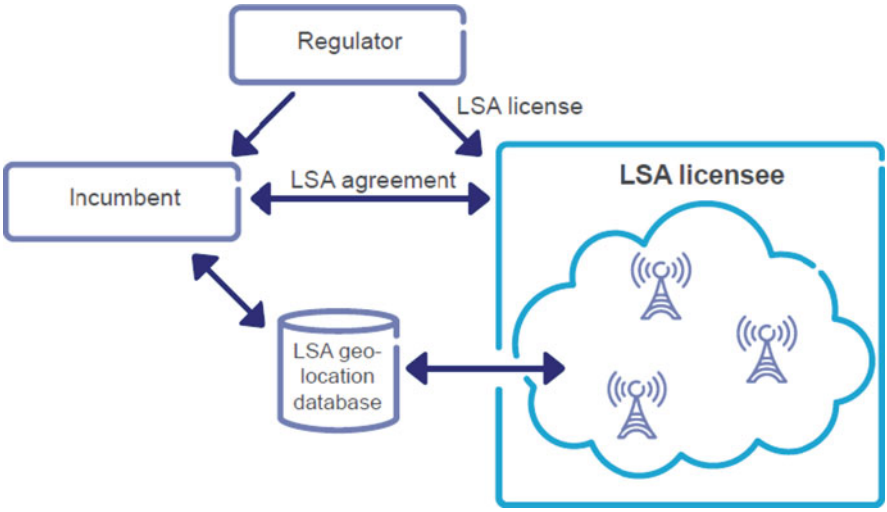


**Fig. 6** GLDB architecture for LSA [5]

band only where the LSA controller indicates that the spectrum is available and no harmful interference is caused to the incumbent. This management of access to LSA spectrum is completely transparent to the mobile device which simply roams between spectrum bands according to the Radio Resource Management (RRM) messages it receives from the network.

## *3.3 Cognitive/Dynamic Spectrum Access (DSA)*

Cognitive radio is an automatic machine that uses software defined radio to change frequency band to adapt and sense the surrounding environment. They recognize radio spectrum by sensing when it is unused by the incumbent radio system and use this spectrum in an intelligent way [6].

The idea of cognitive radio (CR) was first presented officially in an article by Joseph Mitola III and Gerald Q. Maguire, Jr in 1999. It was a new approach in wireless communications that Mitola described as: "The point in which wireless personal digital assistants (PDAs) and the related networks are sufficiently computationally intelligent about radio resources and related computer-to-computer communications to detect user communications needs as a function of use context, and to provide radio resources and wireless services most appropriate to those needs" [7]. Later, other researchers or organizations have evolving definitions. For example, the U.S. Federal Communications Commission (FCC) defines cognitive radio as an aware adaptive radio that senses the radio environment and adapts to it to enable more

efficient use of the available spectrum [8]. In general, all definitions of cognitive radio have the common features that the cognitive radio should, to some extent, be aware of its radio frequency environment and be able to adapt to optimize its operation.

There are two different types of systems present in situations in which cognitive radios are used to dynamically access the spectrum: (1) A *primary user* on a portion of the spectrum is a user who owns a license that legally permits the user to use the spectrum for a purpose specified by the license; (2) A *secondary user* is a user that access spectrum that is licensed by a primary user. This access has typically to be regulated in such a way that the secondary user does not cause harmful interference to the license holder.

There are two important functions to achieve good spectrum sharing between primary user and secondary user:

- *Spectrum sensing* function available in secondary systems to sense status of primary systems to find spectrum opportunity;
- *Dynamic spectrum access (DSA)* function available in secondary system to control spectrum access dynamically according to varying environment of primary systems.

### 3.3.1 Spectrum Sensing

The main challenge of spectrum sensing is the strict requirement of sensitivity. In order to incorporate the noise uncertainty and multi-path fading/shadowing uncertainty into the specification, the sensing node requires identifying the presence of primary signal in the low signal to noise ratio (SNR) region. In Table 1, several typical spectrum sensing algorithms are summarized including basic energy detection and other feature detection methods that improve the sensitivity by exploiting a priori knowledge embedded in the primary signal.

In general, the above schemes are oriented for local spectrum sensing in one node. However, if the primary signal is in the deep fading or obstructed by a large obstacle, the received power at secondary node is too weak to detect. Thus hidden node problem occurs in this situation. This uncertainty can be effectively mitigated by the spatial diversity in the cooperative sensing. Additionally, the cooperative sensing can relax the sensitivity requirement of local sensing and increase the agility of making sensing decision. The procedure of cooperative sensing is as follows: all cognitive radios that have agreed to participate in the cooperative sensing process sense the spectrum and communicate their results to the other cognitive radios. Based on the information type sent by the individual CR, the cooperative sensing can be categorized by soft decision fusion and hard decision fusion.

- *Hard decision fusion* is the merge of the information to produce a common final decision for the whole system if the information that it has received consists of individual preliminary decisions;

**Table 1** Summary of spectrum sensing technique

| Technique | Description | Requirement | Remark |
|---|---|---|---|
| Energy detection [9, 10] | Measure the radio frequency energy or the received signal strength indicator over certain observation time and compare the accumulated metric to a predefined threshold | No | (+) Easily implemented without a priori knowledge<br>(−) Sensitive to the noise and interference uncertainty<br>(−) Can't distinguish primary signal with other interference |
| Coherent detection [11, 12] | Detect known pilot signal of primary systems with matched filter and compare obtained metric to a predefined threshold | Requires the knowledge of primary signal at both PHY layer (e.g. pulse shaping, modulation) and MAC layer | (+) can distinguish primary signal from interference and noise<br>(+) Work well in low SNR<br>(+) Shorter observation time<br>(−) More implementation complexity by performing timing, carrier sync and equalization |
| Covariance-based detection [13] | Estimate sample covariance matrix of the received signal and make spectrum whiteness tests from eigenvalue-based or covariance statistic-based statistics | Primary signal correlated due to dispersive channel, correlation of multiple receiver antennas or over-sampling effect | (+) Exploit more general feature of primary signal without signal-specific knowledge<br>(−) Performance highly depends on correlation of primary signal |
| Cyclostationary detection [14, 15] | Analyze the cyclic autocorrelation function or its equivalent Fourier transformation spectrum correlation to seek cyclostationary feature | Cyclostationary feature in primary signal caused by modulation and coding, hopping sequence, cyclic prefix | (+) Capability of differentiating the primary signal from the interference and noise<br>(−) More implementation complexity with wide scan of the cyclic frequencies |

- *Soft decision fusion* means what it receives is channel statistics or raw data the procedure of creating the final decision.

The soft decision fusion outperforms the hard decision fusion at the expense of large signaling overhead to forward the sensing data. As an alternative, the quantized decision fusion is desirable to achieve the tradeoff between these two issues. It is shown that even a simple cooperative scheme, such as the OR-rule,

may improve the performance of the cognitive secondary system significantly. The probability of false detection can be made arbitrarily small while maintaining the level of interference protection of the primary system. This is achieved by adding more cooperating cognitive users that are experiencing uncorrelated fading or shadowing. Considering only few cooperating uncorrelated users give considerable improvement in the sensitivity, compared to the case without cooperation.

### 3.3.2 Dynamic Spectrum Access

After sensing the status of primary systems, secondary systems need to determine how to access the spectrum according to dynamic sensing results. There are three modes of operation dealing with the spectrum access of the cognitive radio (or secondary user) in fundamentally different ways and it is thus reasonable to use them to classify the cognitive radio behavior.

#### Interweave Operation

The basic idea of interweave operation is that the secondary systems scan the radio environment to find unused parts of the spectrum, so called spectrum holes. Once such holes are identified the secondary users start to communicate using these unused frequencies. Since the radio frequency environment is constantly changing the available spectrum holes change with time and the secondary users must constantly scan the spectrum to keep track of the spectrum holes and possibly predict where they may appear. To enable communication using spectrum to be possible, the secondary transmitter needs to communicate to the intended receiver with the chosen transmission parameters. This might be performed over a dedicated pilot channel, possibly located in an unlicensed band. An alternative way to initialize communication in a secondary system is to send a trigger on the frequency intended to be used. The trigger may be followed by a standardized pilot signal containing the necessary transmission parameters and the intended receiver identity. This possibility arises since the secondary receiver constantly scans the entire frequency range and hence, may be able to detect the trigger signal and start to receive using the parameters specified by the pilot following the trigger signal.

#### Underlay Operation

Underlay access means operating below the noise floor of the primary systems, involving an undercurrent of secondary system communications without primary systems being aware of. In details, a secondary system that operates in the underlay mode transmits at the same time as the primary systems of the spectrum under consideration. However, the secondary system uses a spectrum mask to make sure that the interference caused to the primary users is kept below the noise floor, or

below the interference limit. In order for the secondary systems to obtain a sufficient signal-to-noise ratio (SNR), a wide bandwidth might be used for the secondary signal.

There are obvious difficulties of ensuring a sufficiently low level of interference at a primary receiver that is most likely hidden from the secondary transmitter. For this reason underlay operation is most likely only applicable for short and low power communication between secondary users. One possible implementation of underlay spectrum access is via ultra-wide band (UWB) techniques at low power.

Overlay Operation

For overlay operation, the level of interference at a primary receiver is kept below a certain tolerable level. The idea in this channel model is that the secondary transmitter (cognitive radio) obtains the message to be transmitted in the primary system in a non-causal way and uses dirty paper coding technique, or possibly linear interference cancellation techniques, to eliminate the interference at the secondary receiver due to the message sent in the primary system. This approach does not try to limit the interference to the primary system and hence operation according to this scheme violates the idea of a cognitive radio acting as a secondary system not causing harmful interference to the primary system. For this reason the cognitive radio channel is sometimes denoted a selfish approach to overlay operation.

### 3.3.3 Application System Examples

The most simple application example is Dynamic Frequency Selection (DFS) in 5 GHz spectrum for Wi-Fi coexistence with Radar systems. Radar systems are given priority over all other devices operating in the DFS bands, i.e. 5250–5350 MHz and 5470–5725 MHz. As a result, all devices operating at these frequencies must stop transmitting when radar is detected on the current channel. So this is a typical vertical sharing scenario. Secondary system needs to detect if there is Radar systems all the time. It adopts interweave operation mode, i.e. secondary systems must not transmit on that channel for 30 min (non-occupancy period) from the time of detection.

Another example is IEEE 802.22 protocol operating on TV band. It is a wireless communication standard for operating as a non-interfering secondary user in TV whitespace with fixed communication nodes, aimed at bringing broadband internet access to rural environments worldwide. Along with GLDB support, spectrum sensing is also adopted to mitigate the interference to TV and wireless microphones. Spectrum sensing is mandatory in both BSs and user terminals. Three types of incumbents are needed to sense, i.e. analog TV, digital TV and low power licensed device (e.g. wireless microphones). There are several sensing requirements specified as sensing receiver sensitivity (0 dBi in all directions), channel detection time (2 s from start of incumbent transmission), probability of detection (0.9 @ sensitivity),

probability of false alarm (0.1 @ sensitivity) and etc. To detect the incumbent systems, network-wide quiet periods are scheduled by BS for sensing time of both BS and UEs. Besides 802.22 protocol, there are several other ones aiming to work on TV white space, i.e. 802.11af, ECMA and etc. All their spectrum sensing parts are similar standardized according to regulation rules.

## *3.4 MAC-Based Coexistence Mechanism*

The MAC protocol of a system can contain functionality that enables spectrum sharing especially. The key essence is to avoid possible collision which may occur between different systems or operators but can't guarantee that there is no interference at all. The MAC is designed to alleviate the collision as much as possible, i.e. 'best-effort' collision avoidance. So this tool is more suitable for horizontal sharing. Examples of this approach include listen before talk (LBT)-based CSMA/CA protocol in IEEE 802.11 and LAA-LTE, frequency hopping in Bluetooth and etc.

A particularly interesting instance of MAC behavior for coexistence is the LBT-based CSMA/CA. One example using this protocol is IEEE 802.11 WiFi systems. As shown in Fig. 7, the basic protocol is applied by all stations, including the access point, i.e. in both downlink and uplink. A station that wishes to transmit a packet first senses the medium. If the medium is sensed idle for a certain minimum time, a so-called Distributed Inter Frame Space (DIFS, 34 μs for 11n), the packet is transmitted. If the medium is busy, the station first defers until the medium is sensed idle. When this occurs, the station does not transmit immediately, which would lead to collisions if more than one station was deferring. Instead, the station sets a backoff timer to a random number, and does not transmit until this timer has expired. The backoff timer is only decreased when the medium is sensed idle, whereas whenever the medium is sensed busy, a deferral state is entered where the backoff timer is not decreased. When the backoff timer expires, the packet is transmitted. Another example is LAA-LTE in unlicensed band which is using LBT-based CSMA/CA to deal with coexistence with WiFi and between different operators as well. The listen-before-talk mechanism is similar with WiFi and the

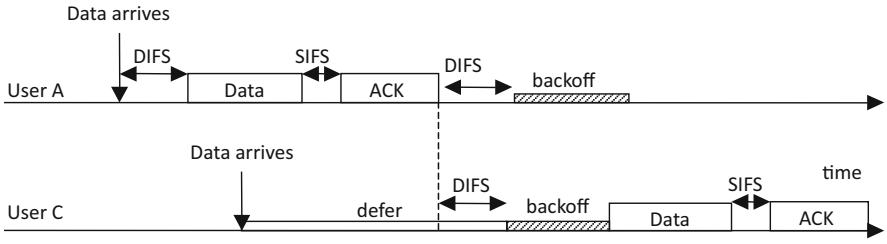


**Fig. 7** Illustration of IEEE 802.11 MAC

most difference is that LAA-LTE systems can't transmit the data any time since the data should be scheduled by downlink control channel in certain resource part. One solution is to insert redundancy signal until control channel boundary when the channel is determined to be idle. Both the simulation and demo results show that LAA-LTE is a good neighbor to WiFi systems and can make both systems working better than before [16].

Bluetooth operates at 2.4 GHz short-range radio frequency band as WiFi does. But it uses a different radio technology called frequency-hopping spread spectrum to deal with coexistence problem. Bluetooth divides transmitted data into packets, and transmits each packet on one of 79 designated Bluetooth channels. Each channel has a bandwidth of 1 MHz. Bluetooth 4.0 uses 2 MHz spacing, which accommodates 40 channels. The first channel starts at 2402 MHz and continues up to 2480 MHz in 1 MHz steps. It usually performs 1600 hops per second, with Adaptive Frequency-Hopping (AFH) enabled.

### *3.5 Conclusion*

It is clearly seen that various spectrum sharing techniques are already used in existing wireless systems but not for cellular systems. However, the application of them is different according to the features. Coordination protocol is more feasible for horizontal sharing between the same technologies since connection between different systems is very difficult to be established. GLDB and cognitive radio are two important techniques to enable vertical sharing between different systems. GLDB is more reliable to guarantee the protection to primary systems but centralized structure and database information collection of other systems are needed which is a lot of overhead. Cognitive radio is flexible as a distributed way for coexistence with primary systems although it may not be reliable when spectrum sensing. MAC-based coexistence mechanism such as CSMA is very promising for sharing between systems in unlicensed band. To conclude, these spectrum sharing techniques in existing wireless systems will provide a good reference for the application in 5G cellular systems.

## 4 5G Spectrum Sharing Considerations

### *4.1 5G Spectrum Sharing Enablers*

5G systems have some natural characteristics to facilitate the above-mentioned spectrum sharing technique, which is explained in detail as follows:

- *Vertical sharing* can be a good way to unlock many bands for 5G systems as soon as possible. As mentioned in this chapter, the candidate bands for 5G systems are

not "clean" even for frequency above 6 GHz. Almost all candidate bands have other existing allocations and already in use, e.g. fixed services or radar systems. Coexistence with them may not in every case be easy and thus one way to use these bands for 5G systems is to repurpose them for licensed cellular usage. In some cases this approach may require a lot of effort and time to succeed. In order to make them available for 5G use in relatively short time, one possible compromise is to introduce LSA. 5G systems would then be operating under the constraint of protecting the incumbent service. To allow an efficient, spectrum database or spectrum sensing technique may be employed and details will be described in later sections.

- *Homogenous horizontal sharing* could be feasible for 5G systems particularly in high frequency band, i.e. above 6 GHz. Using these bands might not target to provide ubiquitous coverage but non-continual coverage islands. Further, different operators may be required to serve different traffic volumes at the same instant. In such case, spectrum sharing techniques can bring potential gain from statistical multiplexing. Besides, propagation conditions at high frequencies combined with high-gain beamforming for 5G systems is a promising venue for making good use of such frequencies.
- *Heterogeneous horizontal sharing* will likely only be employed in unlicensed spectrum and can be used to boost capacity for 5G systems. The use of unlicensed spectrum in 5 GHz for 4G systems is being discussed in 3GPP in combination of licensed spectrum for critical control signaling. It is very likely that unlicensed mode will continue to be a complementary method of using spectrum for 5G systems, especially for indoor enterprise solutions.

## *4.2 5G General Architecture for Spectrum Sharing*

The impact of spectrum sharing on 5G system architecture discussed in this section aims at providing basic logical architecture support for implementation of 5G spectrum sharing in all scenarios foreseen by all spectrum sharing techniques. For specific scenario, only parts of functions are enabled according to the needs.

In order to meet the above functionality requirements, new logical entities need to be introduced in the overall logical 5G spectrum sharing architecture. The general view of additional logical entities overlaid to the physical structure is given in Fig. 8.

### 4.2.1 Spectrum Management Client (SMaC)

As shown in Fig. 8, SMaC is implemented in each Access Node (AN) to handle spectrum-related functions. Basically, there will be one spectrum information database which provides a list of supported frequency bands from hardware point of view, comprising of frequency range, band type (i.e. primary user mode, unlicensed

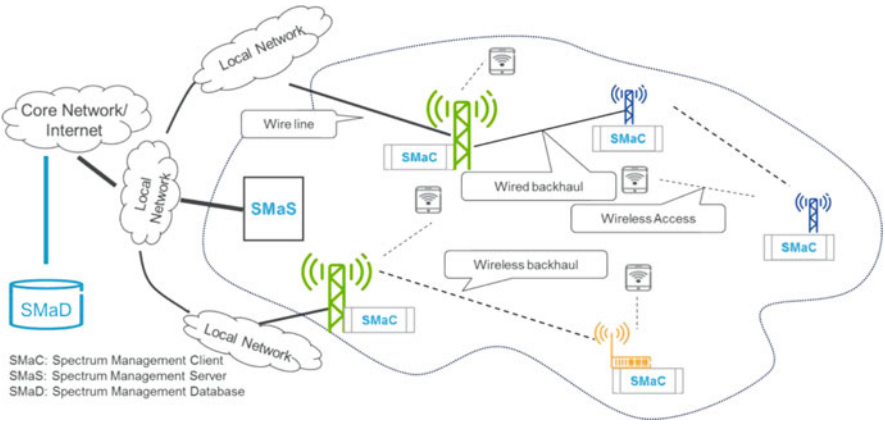


**Fig. 8** Logical spectrum management entities (*blue*) over physical 5G architecture

and LSA) and potential coexisting system type (e.g. Radar, Wi-Fi, other 5G systems and etc.). The basic function for SMaC is to check the band information from spectrum information database which may be stored in memory part of AN. In other words, when AN is willing to operate or operating in certain band, the implemented SMaC inside is able to know which spectrum sharing scenarios are needed by checking the stored band information in its database.

For example, if 5G system is operating on TVWS, it will know that this is unlicensed band with primary users. Thus the following spectrum sharing scenarios listed in Fig. 2 are needed in this band: *vertical sharing* as secondary user and *horizontal sharing* including heterogeneous and homogeneous. Additionally, it may be also known that there will be TV systems as primary user and WiFi system as coexisting system. Then to enable coexistence with different systems in different spectrum sharing scenario, different tool box should be selected (e.g. GLDB for coexistence with TV systems in the scenario of vertical coexistence as secondary user). SMaC is an import entity introduced to support implementation of spectrum sharing tools, which comprises of the following possible support functions (SF):

- **SF for GLDB**: request spectrum availability information and receive spectrum information directly from GLDB or indirectly from the Spectrum Management Server (SMaS introduced below);
- **SF for Cognitive/DSA and MAC-based coexistence**: local spectrum sensing (energy, feature detection or virtual sensing) for primary systems (e.g. Radar etc.) or horizontal systems (e.g. Wi-Fi systems and etc.);
- **SF for Cognitive/DSA**: send or receive the spectrum sensing related information to other logical functionalities (e.g. SMaS or other SMaCs) for cooperative spectrum sensing;
- **SF for Coordination Protocol**: local interference measurement information collection from other 5G systems in the same operating channel and exchange information with other SMaCs located in neighbor ANs if needed;

- **SF for all**: make its own decision on which part of the spectrum is available for use based on collected information or receive centralized decision or recommendation information for band availability from other logical functionalities (e.g. SMaS).

Note that not all the SFs listed above should be implemented in SMaC. It depends on which spectrum sharing tool is selected and how it is implemented.

### 4.2.2 Spectrum Management Server (SMaS)

The spectrum management server (SMaS) is an operational assisting logical entity for centralized solution in part of 5G networks. Here "centralized" means it can collect information from other distributed entities (i.e. SMaC) and may provide a centralized decision or recommendation for spectrum availability based on gathered information (e.g. sensing). Having a SMaS as a centralized logical entity is optional, i.e. depending on whether the spectrum sharing tool is implemented in a centralized manner. One benefit of this is to provide more accurate estimation on the spectrum availability. For example, SMaS can gather sensing information from multiple SMaCs for cooperative spectrum sensing, which will avoid hidden node problem and improve spectrum discovery performance. The SMaS can be implemented in a centralized control node if existing. The SMaS may have one or more of the following functions:

- **SF for GLDB support**: If applicable (i.e.in scenarios where this is required), send request for spectrum status and receive UDN-wide spectrum information directly from GLDB/spectrum broker (i.e. SMaS can act as a proxy for the SMaCs);
- **SF for Cognitive/DSA and MAC-based coexistence**: Receive the spectrum sensing related information to other logical functionalities (e.g. SMaS or other SMaCs) for cooperative spectrum sensing;
- **SF for all**: Make centralized recommendation on spectrum availability information for its connected SMaCs and Send centralized recommendation on spectrum availability information to its connected SMaCs.

Note that SMaS is a purely logical entity to form the logical spectrum sharing architecture. Even when there is a SMaS in one 5G systems, only part of the listed functions may be selected to be implemented for a specific solution.

### 4.2.3 Spectrum Management Database (SMaD)

A SMaD is the global centralized logical functionality to provide spectrum information about primary systems. It is typically located outside the 5G networks. Obviously GLDB is a typical SMaD. The detailed functions are listed as follows:

- Keep track of the spectrum usage status of the primary system and estimate the spectrum availability and corresponding requirements for different areas and keep the information updated;
- Receive spectrum requests and related information from SMaS or SMaCs;
- Send spectrum information (availability status and constraint or grant) to other logical functionalities (e.g. SMaS or SMaC).

Obviously most of the functions listed in above section require logical connections between the different logical entities. For example, the implementation of the first listed function for the SMaC needs the connection between SMaC and SMaD or SMaS. In addition, a protocol running over those interfaces needs to be defined as well to ensure that the information transmitted on the connection can be understood by the communicating logical entities. Therefore, a logical architecture can be established by including logical entities, connections and protocols for spectrum sharing. Given different spectrum sharing requirements, different logical architectures may be designed. In order to provide a complete view for spectrum sharing support, our proposed architecture here is aiming to support the implementation of all spectrum sharing toolbox components in the different spectrum sharing scenarios from Fig. 2.

From Fig. 9, it can be seen that for different kinds of logical connections, there will be different kinds of spectrum sharing information exchange between the end points of the logical connection. Therefore, different protocols should be developed to make the communication understandable. In Fig. 10, it is exemplified by detailing some behavior by the various flavors of the Spectrum Management Protocol (SMP):

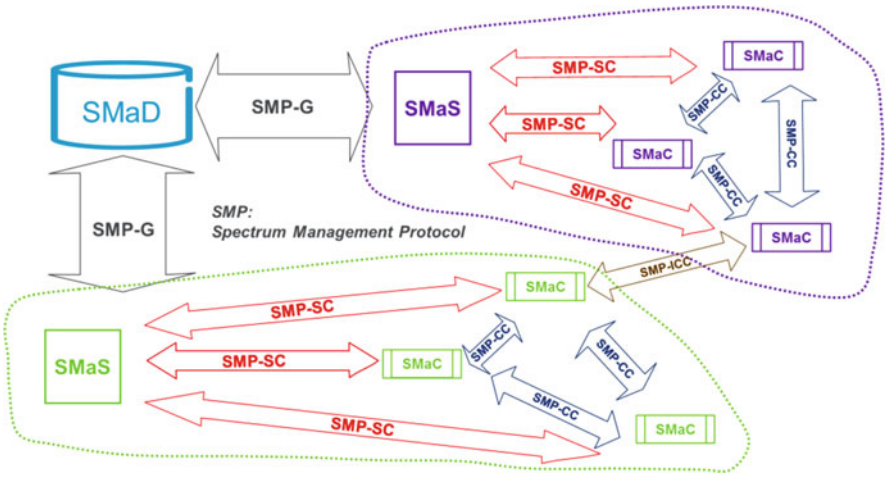


**Fig. 9** Logical 5G spectrum sharing architecture

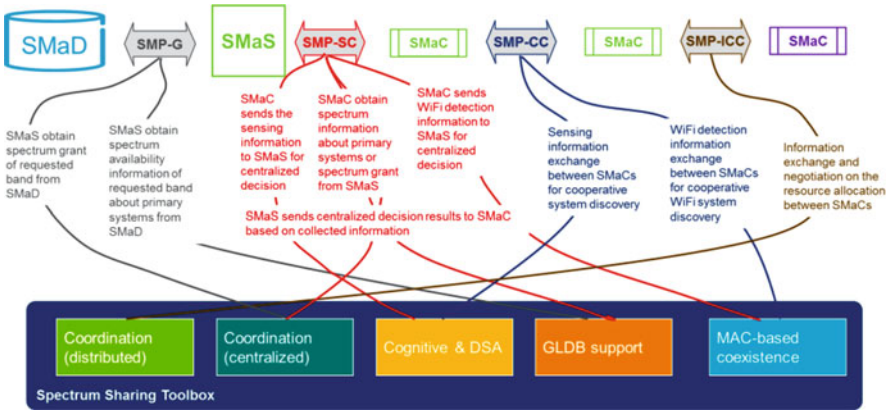


**Fig. 10** Illustration of spectrum management protocols

- SMP-G (SMP-Global): Protocol used to communicate with external services such as a geo-location database or a spectrum coordinator or broker entity
- SMP-SC (SMP-Server Client): Protocol used for communication between the SMaS and the SMaC functionalities
- SMP-CC (SMP-Client Client): Protocol used for communication between SMaC functional entities residing in different nodes.
- SMP-ICC (SMP-Inter Cluster Coordination): Protocol used for information exchange and negotiating between different node clusters.

## *4.3 Spectrum Sharing Techniques for 5G*

For 5G, one main difference is that the operating frequency will go even higher especially beyond 6 GHz. However, current application spectrum sharing techniques are employed in low frequency band, i.e. below 6 GHz. So sharing high frequency band for 5G systems will bring significant impact to current existing spectrum sharing techniques described in Sect. 3. The detailed impacts are analyzed below as a starting point for 5G spectrum sharing applications.

### 4.3.1 Coordination Protocol for 5G

For 5G systems, coordination protocol can be one feasible candidate tool to achieve homogenous horizontal sharing between different sharing networks or operators. It is expected that antenna directivity with massive MIMO, which is

one important feature for 5G, could alleviate inter-network interference which enables aggressive spatial resource reuse between different networks. At the same time, it brings the difference for interference coordination design from original node pair-based in omni-mode to future link pair-based in high-gain beamforming mode. A 'coordination context' may be introduced for interfering links that require coordination for optimized performance. The objective is a solution that coordinates radio resource usage on a "per link" basis, i.e. only transmission links that actually create non-negligible interference (beyond a given threshold) to particular other links should be subject to coordination. Here, coordination means that transmissions of an inter-network interference link pair (one link in one 5G system and one in the other) are subject to scheduling constraints so that they cannot be scheduled on the same (time/frequency) radio resources. The two interfering networks negotiate an agreement on such a resource partitioning and record it in the form of a Coordination Context (CC) which is stored in a Coordination Context Database (CCDB) in each network. The CC thus represents a constraint that needs to be considered by the scheduler in each network. For example as shown in Fig. 11, A1 and B1 is an interfering link pair between different 5G systems. As a result of coordination, they will be coordinated to that the blue resources are reserved for A1 and the red resources for B1. In general, the resource division for one interference link pair can be in time, frequency or both.

To enable such interference coordination, reliable connections between different networks or operators are needed. This may be more practical for LSA spectrum since one centralized controller may be available for providing spectrum information to achieve vertical sharing. This LSA controller will be a good bridge to handle such kind of coordination between different networks if horizontal sharing is valid.

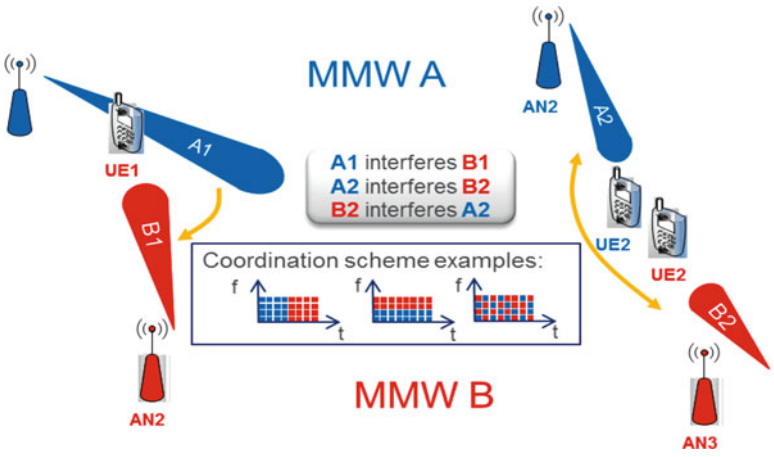


**Fig. 11** Illustration of coordination context concept

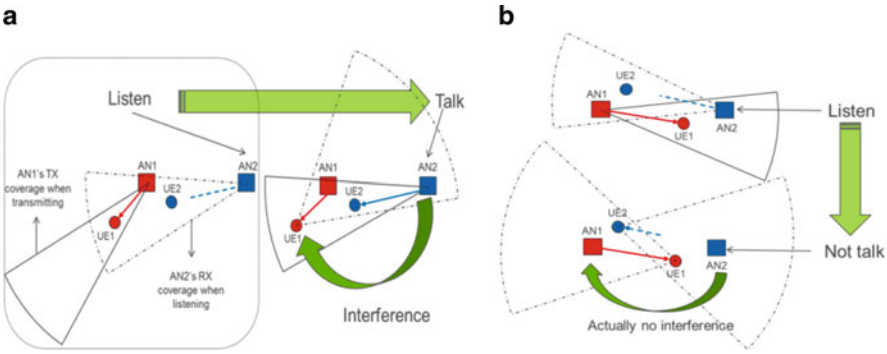


**Fig. 12** Problem illustration of directional LBT: (**a**) hidden node problem; (**b**) exposed node problem

### 4.3.2 MAC-Based Coexistence for 5G

In high frequency band, high-gain beamforming is available for 5G system design, which will bring severe hidden node problem and exposed node problem for traditional LBT. On one aspect, hidden node problem is that transmitter of one link can't hear the current link transmission but there will be interference when they are transmitting simultaneously. One example of hidden node problem is given in Fig. 12a. In left side, AN1 is transmitting data to UE1 and AN2 is listening. But AN2 is not in TX coverage of AN1 and thus considers the channel is available. Then in right side, AN2 starts to transmit data to UE2. Obviously UE1 is interfered by AN2's transmission. On the other aspect, exposed node problem is that transmitter of one link hears the current link transmission but actually there will be not interference when they are transmitting simultaneously. As shown in Fig. 12b, AN1 is transmitting to UE1 and AN2 hear it. So AN2 determines the carrier/channel is busy and will not transmit. But actually there will be no interference between these two links because AN1 is not in RX coverage of UE2 and AN2 is not in RX coverage of UE1.

Based on our simple evaluation and analysis, it can be concluded that traditional 'listen', i.e., physical carrier sensing, is not effective to avoid transmission collision in directional transmission case especially with narrow TX&RX beam width, Particularly there exists severe hidden node problem that almost 100 % real interference case can't be detected and exposed node problem that almost 100 % detected interference is not real interference. The key reason is that transmitter listen to transmitter doesn't represent interference in directional transmission case. Solution directions to this are to involve receiver into LBT process and two different schemes are proposed. One is enhanced LBT with smart listening and notification, i.e. transmitter listens to directional CTS and receiver listens to directional RTS. The other is 'listen after talk', whose key idea is 'talk' first anyway and then collision solving. It is well known that the default mode of LBT for transmitter is 'not to send' and data is sent only when it is confirmed that the channel is idle by listening. Listen

after talk concept adopts opposite logic as follows: the default mode for transmitter is 'to send' and data is not sent only when it is confirmed that channel is occupied by interfering transmissions. This is motivated that low interference situation exists for high frequency band case with high gain beamforming. In such situation, listen after talk will reduce overhead compared to LBT due to no waste of backoff time in most cases.

### 4.3.3 GLDB & Cognitive Radio for Vertical Sharing for 5G

As mentioned in Sects. 2 and 3, current vertical sharing primary systems are TV systems or Radar systems in low frequency band. But from spectrum allocation given by the ITU-R shows that mobile allocation above 6 GHz is almost always allocated together with fixed link services. Above 6 GHz, fixed link services are currently utilizing many spectrum bands. If 5G system intend to use that spectrum, then mechanism to protect the fixed services will be required. Fixed link applications are usually point to point, with very directive/high gain antennas. GLDB and cognitive radio may be employed but additional features should be developed. For GLDB support, the antenna pattern, transmission direction and etc. may also be included in the information database besides location and power information. Furthermore, smart exclusive zone generation should be performed according to coexistence evaluation between fixed link system and 5G systems. For spectrum sensing-based cognitive radio, the hidden node problem and exposed node problem also exists, cf. Fig. 12. More advanced spectrum sensing algorithms are needed to solve these problems of high-gain beamforming case in high frequency band.

## 5 Summary

First, spectrum sharing becomes more and more important for 5G systems, which will comprise of multiple spectrum type with different scenarios, i.e., vertical sharing and horizontal sharing. Next, spectrum sharing techniques mapping into different scenarios are introduced, i.e. coordination protocol, GLDB support, cognitive/DSA and MAC-based coexistence. Besides, current applications of these techniques in real systems are described. Finally spectrum sharing directions for 5G systems are analyzed for different spectrum sharing techniques.

In general, it should be clear that applying these spectrum sharing techniques to practice need additional functions. These may involve signaling overhead and implementation limits which means more work is needed. Licensed dedicated spectrum will continue to be the dominant spectrum usage method for 5G systems due to the possibility to control interference and guarantee coverage, while other spectrum sharing scenarios will act as complementary spectrum usage methods when beneficial.

## References


1. T. Irnich, J. Kronander, Y. Selén, G. Li, Spectrum sharing scenarios and resulting technical requirements for 5G systems, in *Workshop on Spectrum Sharing Strategies for Wireless Broadband Services, PIMRC 2013*, London, UK, 8–11 September 2013
2. T. Wang, G. Li et al., 5G Spectrum, is China Ready?, published on IEEE Communication Magazine 5G Spectrum: Enabling the Future Mobile Landscape, July 2015
3. M. Xing, Y. Peng, T. Xia, H. Long, K. Zheng, Adaptive spectrum sharing of LTE co-existing with WLAN in unlicensed frequency bands, in *IEEE 81st Vehicular Technology Conference 2015-Spring*, Glasgow, Scotland, 11–14 May 2015
4. IEEE Standard for Local and metropolitan area networks Part 16. Air Interface for Broadband Wireless Access Systems Amendment 2: Improved Coexistence Mechanisms for License-Exempt Operation. IEEE Std 802.16h-2010 (Amendment to IEEE Std 802.16-2009), July 30, 2010. doi:10.1109/IEEESTD.2010.5538195
5. Ericsson White Paper on Spectrum Sharing. http://www.ericsson.com/res/docs/whitepapers/wp-spectrum-sharing.pdf, Oct. 2013
6. Cognitive Radio and Dynamic Spectrum Access, Lars Berlemann Deutsche Telekom, Germany and Stefan Mangold Swisscom, Switzerland
7. Cognitive Radio An Integrated Agent Architecture for Software Defined Radio-Dissertation, Doctor of Technology Joseph Mitola III, Royal Institute of Technology (KTH), ISSN 1403-5286, ISRN KTH/IT/AVH-00/01-SE, 8 May 2000
8. Federal Communications Commission, Cognitive Radio Technologies Proceeding (CRTP) ET Docket No. 03-108, March 2005. http://www.fcc.gov/oet/cognitiveradio/
9. F.F. Digham, M.-S. Alouini, M.K. Simon, On the energy detection of unknown signals over fading channels, in *IEEE ICC 2003*, Washington, May 2003
10. Y.C. Liang, Y. Zeng, E.C.Y. Peh, A.T. Hoang, Sensing-throughput tradeoff for cognitive radio networks. IEEE Trans. Wireless Commun. **7**(4), 1326–1337 (2008)
11. M.K. Steveny, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory* (Prentice Hall PTR, 1998)
12. H.-S. Chen, W. Gao, D.G. Daut, Signature based spectrum sensing algorithms for IEEE 802.22 WRAN, in *IEEE International Conference on Communications (ICC)*, June 2007, pp. 6487–6492
13. Y. Zeng, Y.C. Liang, Spectrum-sensing algorithms for cognitive radio based on statistical covariances. IEEE Trans. Veh. Technol. **58**(4), 1804–1815 (2009)
14. A. Dandawate, G.B. Giannakis, Statistical tests for presence of cyclostationarity. IEEE Trans. Sig. Process. **42**, 2355–2369 (1994)
15. D. Birru, K. Challapali, B. Dong, Detection of the presence of television signals embedded in noise using cyclostationary toolbox. U.S. Patent 2010/0157066, 2010
16. A. Mukherjee, J.F. Cheng et al., System architecture and coexistence evaluation of licensed-assisted access LTE with IEEE 802.11, in *IEEE ICC 2015*, London, June 2015

# Part II
# Transmission and Design Techniques for 5G Networks

# Massive MIMO Communications

**Trinh Van Chien and Emil Björnson**

**Abstract** Every new network generation needs to make a leap in area data throughput, to manage the growing wireless data traffic. The Massive MIMO technology can bring at least ten-fold improvements in area throughput by increasing the spectral efficiency (bit/s/Hz/cell), while using the same bandwidth and density of base stations as in current networks. These extraordinary gains are achieved by equipping the base stations with arrays of a hundred antennas to enable spatial multiplexing of tens of user terminals. This chapter explains the basic motivations and communication theory behind the Massive MIMO technology, and provides implementation-related design guidelines.

## 1 Introduction

Much higher area data throughput is required in future cellular networks, since the global demand for wireless data traffic is continuously growing. This goal can be achieved without the need for more bandwidth or additional base stations if the spectral efficiency (measured in bit/s/Hz/cell) is improved. This chapter explains why the Massive MIMO (multiple-input multiple-output) communication technology, where multi-antenna base stations spatially multiplex a multitude of user terminals over the entire bandwidth, is well-suited for this purpose. The rationale behind the Massive MIMO concept and its transmission protocol is explained from a historical perspective in Sect. 2. Next, Sect. 3 provides a basic communication theoretic performance analysis. Closed-form spectral efficiency expressions are derived and the key properties and performance limitations of Massive MIMO are highlighted. The chapter is concluded by Sect. 4 where implementation-related design guidelines are given, particularly regarding power allocation and the reuse of pilot sequences for efficient channel estimation. Multi-cell simulations are provided to showcase that the Massive MIMO technology can provide tenfold or even 50-fold improvements in spectral efficiency over contemporary technology, without

T. Van Chien • E. Björnson (✉)
Department of Electrical Engineering (ISY), Linköping University,
SE-581 83 Linköping, Sweden
e-mail: trinh.van.chien@liu.se; emil.bjornson@liu.se



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_4

the need for advanced signal processing or network coordination. Finally, the full mathematical details are provided in Appendix at the end of this chapter.

## 2 Importance of Improving the Spectral Efficiency

The wireless information traffic has doubled every two and a half years since the beginning of wireless communications, as observed by Martin Cooper at ArrayComm in the nineties. Different technologies and use cases have dominated in different periods, but the exponential increase is currently driven by wireless data traffic in cellular and local area networks. There are no indications that this trend will break anytime soon; in fact, a slightly faster traffic growth is predicted in the well-reputed Cisco Visual Networking Index and Ericsson Mobility Report.

To keep up with the rapid traffic growth, a key goal of the 5G technologies is to improve the area throughput by orders of magnitude; 100× and even 1000× higher throughput are regularly mentioned as 5G design goals. The area throughput of a wireless network is measured in bit/s/km$^2$ and can be modeled as follows:

$$\text{Area throughput (bit/s/km}^2\text{)} =$$

$$\text{Bandwidth (Hz)} \times \text{Cell density (cells/km}^2\text{)} \times \text{Spectral efficiency (bit/s/Hz/cell)}.$$

This simple formula reveals that there are three main components that can be improved to yield higher area throughput: (1) more bandwidth can be allocated for 5G services; (2) the network can be densified by adding more cells with independently operating access points; and (3) the efficiency of the data transmissions (per cell and for a given amount of bandwidth) can be improved.

The improvements in area throughput in previous network generations have greatly resulted from cell densification and allocation of more bandwidth. In urban environments, where contemporary networks are facing the highest traffic demands, cellular networks are nowadays deployed with a few hundred meters inter-site distances and wireless local area networks (WLANs) are available almost everywhere. Further cell densification is certainly possible, but it appears that we are reaching a saturation point. Moreover, the most valuable frequency bands are below 6 GHz because these frequencies can provide good network coverage and service quality, while higher bands might only work well under short-range line-of-sight conditions. In a typical country like Sweden, the cellular and WLAN technologies have in total been allocated more than 1 GHz of bandwidth in the interval below 6 GHz and thus we cannot expect any major bandwidth improvements either.

In contrast, the spectral efficiency (SE) has not seen any major improvements in previous network generations. Hence, it might be a factor that can be greatly improved in the future and possibly become the primary way to achieve high area throughput in 5G networks. In this chapter, we describe the rationale and background of the physical-layer technology Massive multiple-input multiple-output (MIMO), which provides the means to improve the SE of future networks by one or two orders of magnitude.

## 2.1 Multi-User MIMO Communication

The SE of a single-input single-output (SISO) communication channel, from a single-antenna transmitter to a single-antenna receiver, is upper bounded by the Shannon capacity, which has the form $\log_2(1 + \overline{SNR})$ bit/s/Hz for additive white Gaussian noise (AWGN) channels. The SISO capacity is thus a logarithmic function of the signal-to-noise ratio (SNR), denoted here as $\overline{SNR}$. To improve the SE we need to increase the SNR, which corresponds to increasing the power of the transmitted signal. For example, suppose we have a system that operates at 2 bit/s/Hz and we would like to double its SE to 4 bit/s/Hz, then this corresponds to improving the SNR by a factor 5, from 3 to 15. The next doubling of the SE, from 4 to 8 bit/s/Hz, requires another 17 times more power. In other words, the logarithm of the SE expression forces us to increase the transmit power exponentially fast to achieve a linear increase in the SE of the SISO channel. This is clearly a very inefficient and non-scalable way to improve the SE, and the approach also breaks down when there are interfering transmissions in other cells that scale their transmit powers in the same manner. We therefore need to identify another way to improve the SE of cellular networks.

Each base station (BS) in a cellular network serves a multitude of user terminals. Traditionally, the time/frequency resources have been divided into resource blocks and only one of the user terminals was active per block. This terminal can then receive a single data stream with an SE quantified as $\log_2(1 + \overline{SNR})$. The efficient way to increase the SE of a cellular network is to have multiple parallel transmissions. If there are $\mathbb{G}$ parallel and independent transmissions, the sum SE becomes $\mathbb{G}\log_2(1 + \overline{SNR})$ where $\mathbb{G}$ acts as a multiplicative pre-log factor. Parallel transmissions can be realized by having multiple transmit antennas and multiple receive antennas. There are two distinct cases:

1. Point-to-point MIMO [39], where a BS with multiple antennas communicates with a single user terminal having multiple antennas.
2. Multi-user MIMO [34], where a BS with multiple antennas communicates with multiple user terminals, each having one or multiple antennas.

There are many reasons why multi-user MIMO is the most scalable and attractive solution [17]. Firstly, the wavelength is 5–30 cm in the frequency range of cellular communication (1–6 GHz). This limits the number of antennas that can be deployed in a compact user terminal for point-to-point MIMO, while one can have almost any number of spatially separated single-antenna terminals in multi-user MIMO. This is an important distinction since the number of simultaneous data streams that can be separated by MIMO processing equals the minimum of the number of transmit and receive antennas. Secondly, the wireless propagation channel to a user terminal is likely to have only a few dominating paths, which limits the ability to convey multiple parallel data streams to a terminal in point-to-point MIMO. The corresponding restriction on multi-user MIMO is that the users need to be, say, a few meters apart to have sufficiently different channel characteristics, which is

a very loose restriction that is true in most practical scenarios. Thirdly, advanced signal processing is needed at the terminals in point-to-point MIMO to detect the multiple data streams, while each terminal in multi-user MIMO only needs to detect a single data stream.

The canonical multi-user MIMO system consists of a BS with $M$ antennas that serves $K$ single-antenna terminals; see Fig. 1 for a schematic illustration. The BS multiplexes one data stream per user in the downlink and receives one stream per user in the uplink. Simply speaking, the BS uses its antennas to direct each signal towards its desired receiver in the downlink, and to separate the multiple signals received in the uplink. If the terminal is equipped with multiple antennas, it is often beneficial to use these extra antennas to mitigate interference and improve the SNR rather than sending multiple data streams [6]. For the ease of exposition, this chapter concentrates on single-antenna terminals. In this case, $\min(M, K)$ represents the maximal number of data streams that can be simultaneously transmitted in the cell, while still being separable in the spatial domain. The number $\min(M, K)$ is referred to as the *multiplexing gain* of a multi-user MIMO system.

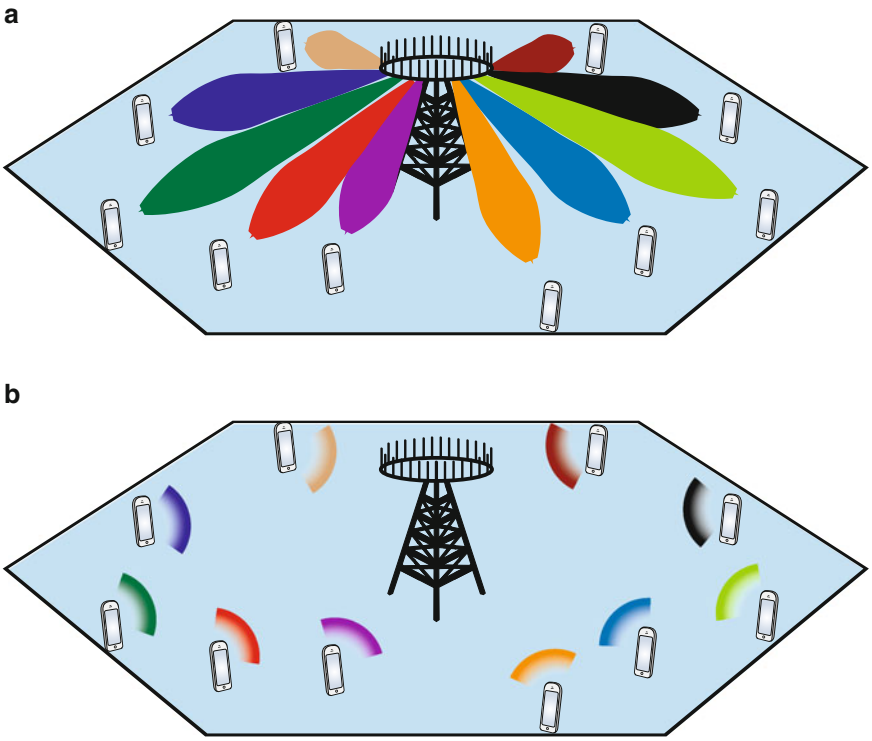


**Fig. 1** Illustration of the downlink and uplink transmission in a multi-user MIMO system, where the BS is equipped with $M$ antennas and serves $K$ user terminals simultaneously. This illustration focuses on line-of-sight propagation where the downlink signals can be viewed as angular beams, but multi-user MIMO works equally well in non-line-of-sight conditions. (**a**) Downlink in multi-user MIMO. (**b**) Uplink in multi-user MIMO

## 2.2 Lessons Learned

The research on multi-user MIMO, particularly with multi-antenna BSs, has been going on for decades. Some notable early works are the array processing papers [1, 38, 44, 47], the patent [36] on spatial division multiple access (SDMA), and the seminal information-theoretic works [11, 18, 42, 43, 46] that characterized the achievable multi-user capacity regions, assuming that perfect channel state information (CSI) is available in the system. In this section, we summarize some of the main design insights that have been obtained over the years.

Capacity-achieving transmission schemes for multi-user MIMO are based upon non-linear signal processing; for example, the dirty-paper coding (DPC) scheme that achieves the downlink capacity and the successive interference cancelation (SIC) scheme that achieves the uplink capacity. The intuition behind these schemes is that the inter-user interference needs to be suppressed, by interference-aware transmit processing or iterative interference-aware receive processing, to achieve the optimal performance. These non-linear schemes naturally require extensive computations and accurate CSI, because otherwise the attempts to subtract interference cause more harm than good.

How large are the gains of optimal non-linear processing (e.g., DPC and SIC) over simplified linear processing schemes where each user terminal is treated separately? To investigate this, let us provide a numerical example where $K = 10$ user terminals are simultaneously served by a BS with $M$ antennas. For simplicity, each user is assumed to have an average SNR of 5 dB, there is perfect CSI available everywhere, and the channels are modeled as uncorrelated Rayleigh fading (this is defined in detail in Sect. 3). Figure 2 shows the average sum SE, as a function of $M$, achieved by sum capacity-achieving non-linear processing and a simplified linear processing scheme called zero-forcing (ZF), which attempts to suppress all interference. The results are representative for both uplink and downlink transmissions.

This simulation shows that the non-linear processing greatly outperforms linear ZF when $M \approx K$. The operating point $M = K$ makes particular sense from a multiplexing perspective since the multiplexing gain $\min(M, K)$ does not improve if we let $M$ increase for a fixed $K$. Nevertheless, Fig. 2 shows that there are other reasons to consider $M > K$; the capacity increases and the performance with linear ZF processing approaches the capacity. Already at $M = 20$ (i.e., $M/K = 2$) there is only a small gap between optimal non-linear processing and linear ZF. In fact, both schemes also approach the upper curve in Fig. 2 which represents the upper bound where the interference between the users is neglected. This shows that we can basically serve all the $K$ users as if each one of them was alone in the cell.

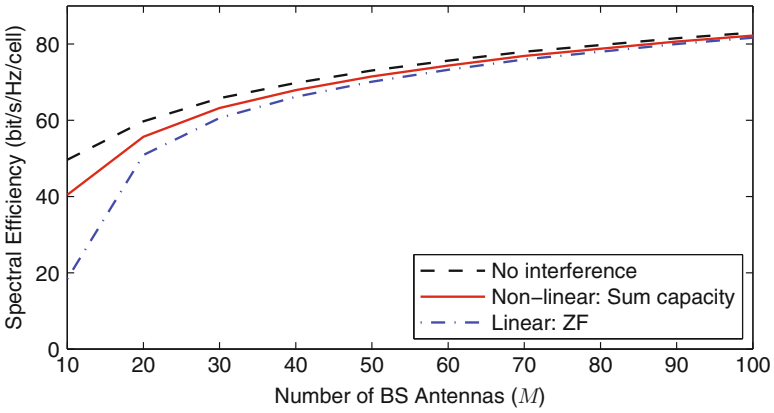


**Fig. 2** Average spectral efficiency in a multi-user MIMO system with $K = 10$ users and varying number of BS antennas. Each user has an average SNR of 5 dB and the channels are Rayleigh fading. The sum capacity is compared with the performance of linear ZF processing and the upper bound when neglecting all interference. The results are representative for both uplink and downlink

**First lesson learned**: Linear processing, such as ZF, provides a sum spectral efficiency close to the sum capacity when $M \gg K$.

The performance analysis and optimization of linear processing schemes have received much attention from academic researchers. While non-linear schemes are hard to implement but relatively easy to analyze and optimize, linear processing schemes have proved to have the opposite characteristics. In particular, computing the optimal downlink linear precoding is an NP-hard problem in many cases [27], which requires monotonic optimization tools to solve; see for example [5]. Nevertheless, the suboptimal ZF curve in Fig. 2 was generated without any complicated optimization, thus showing that the optimal linear processing obtained in [5] can only bring noticeable gains over simple ZF for $M \approx K$, which is the regime where we have learnt not to operate.

As mentioned earlier, the BS needs CSI in multi-user MIMO systems to separate the signals associated with the different users. Perfect CSI can typically not be achieved in practice, since the channels are changing over time and frequency, and thus must be estimated using limited resources. The channel estimation of a frequency-selective channel can be handled by splitting the frequency resources into multiple independent frequency-flat subchannels that can be estimated separately. A known pilot sequence is transmitted over each such subchannel and the received signal is used to estimate the channel response. In order to explore all spatial channel dimensions, this sequence must at least have the same length as the number of

transmit antennas [4]. This means that a pilot sequence sent by the BS needs to have the length $M$, while the combined pilot sequence sent by the single-antenna user terminals needs to have the length $K$.

There are two ways of implementing the downlink and uplink transmission over a given frequency band. In frequency division duplex (FDD) mode the bandwidth is split into two separate parts: one for the uplink and one for the downlink. Pilot sequences are needed in both the downlink and the uplink due to the frequency-selective fading, giving an average pilot length of $(M+K)/2$ per subchannel. There is an alternative time-division duplex (TDD) mode where the whole bandwidth is used for both downlink and uplink transmission, but separated in time. If the system switches between downlink and uplink faster than the channels are changing, then it is sufficient to learn the channels in only one of the directions. This leads to an average pilot length of $\min(M, K)$ per subchannel, if we send pilots only in the most efficient direction. In the preferable operating regime of $M \gg K$, we note that TDD systems should send pilots only in the uplink and the pilot length becomes $\min(M, K) = K$. We conclude that TDD is the preferable mode since it not only requires shorter pilots than FDD, but is also highly scalable since the pilot length is independent of the number of BS antennas.

We give a concrete numerical example in Fig. 3 for downlink transmission with $K = 10$ users, an SNR of 5 dB, and uncorrelated Rayleigh fading channels. Two linear precoding schemes are considered; (a) maximum ratio (MR) and (b) zero-forcing (ZF). These schemes are later defined mathematically in Sect. 3. This simulation compares the SE obtained when having perfect CSI with the performance when having CSI estimated with pilot sequences of length $\tau_p$. The SE is shown as a function of the number of BS antennas, $M$, and we compare TDD mode using $\tau_p = K = 10$ with FDD mode using either $\tau_p = 10$, $\tau_p = M$, or $\tau_p = \min(M, 50)$, where the latter models an arbitrarily chosen maximum pilot length of 50 (e.g., motivated by pilot overhead constraints).

In TDD mode there is a visible performance loss in Fig. 3 as compared to having perfect CSI. The loss with MR precoding is very small, which shows that it is robust to estimation errors. The performance loss is larger for ZF precoding, since estimation errors make it harder to suppress interference, but we notice that ZF anyway provide higher performance than MR for all considered $M$. We notice that the performance losses are substantially constant irrespective of the number of BS antennas, thus TDD systems always benefit from adding more antennas. In contrast, FDD systems only benefits from adding more antennas if the pilot sequences are also made longer, as in the case $\tau_p = M$. With $\tau_p = 10$ there is no benefit from having more than ten antennas, while the performance saturates at 50 antennas when $\tau_p = \min(M, 50)$. In summary, TDD operation is fully scalable with respect to the number of BS antennas, while FDD operation can only handle more antennas by also increasing the pilot overhead. It is practically feasible to deploy FDD systems with many antennas, particularly for slowly varying channels where we can accept a large pilot overhead, but TDD is always the better choice in this respect.

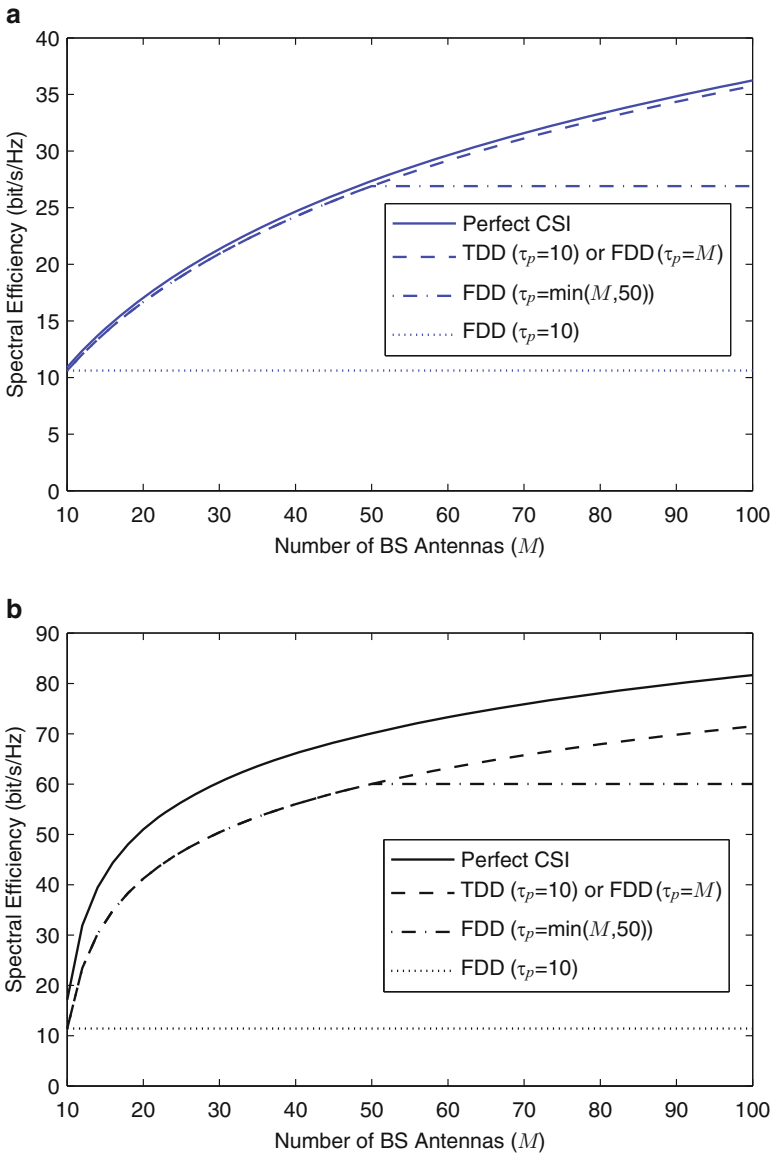


**Fig. 3** Average downlink spectral efficiency, as a function of the number of BS antennas, with different processing schemes and different types of CSI available at the BS. (**a**) Downlink simulation with maximum ratio precoding. (**b**) Downlink simulation with zero-forcing precoding

**Second lesson learned**: The channel estimation is simplified when operating in TDD mode, since the pilot sequences only need to be of length $K$ irrespective of the number of BS antennas $M$.

Note that the uplink works in the same way in the TDD and FDD modes, while the distinct benefit of TDD in terms of scalability appears in the downlink.

### 2.2.1 Favorable Propagation

Recall from Fig. 2 that by adding more BS antennas, both the sum capacity-achieving non-linear processing and the simplified linear ZF processing approached the case without interference. This is not a coincidence but a fundamental property that is referred to as *favorable propagation*.

Let $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{C}^M$ represent the channel responses between a BS and two different user terminals. If these vectors are non-zero and orthogonal in the sense that

$$\mathbf{h}_1^H \mathbf{h}_2 = 0, \tag{1}$$

where $(\cdot)^H$ denotes the conjugate transpose, then the BS can completely separate the signals $s_1, s_2$ transmitted by the users when it observes $\mathbf{y} = \mathbf{h}_1 s_1 + \mathbf{h}_2 s_2$. By simply computing the inner product between $\mathbf{y}$ and $\mathbf{h}_1$, the BS obtains

$$\mathbf{h}_1^H \mathbf{y} = \mathbf{h}_1^H \mathbf{h}_1 s_1 + \mathbf{h}_1^H \mathbf{h}_2 s_2 = \|\mathbf{h}_1\|^2 s_1 \tag{2}$$

where the inter-user interference disappeared due to (1). The same thing can be done for the second user: $\mathbf{h}_2^H \mathbf{y} = \|\mathbf{h}_2\|^2 s_2$. Note that the BS needs perfect knowledge of $\mathbf{h}_1$ and $\mathbf{h}_2$ to compute these inner products. The channel orthogonality in (1) is called favorable propagation, since the two users can communicate with the BS without affecting each other.

Is there a chance that practical channels offer favorable propagation? Probably not according to the strict definition that $\mathbf{h}_1^H \mathbf{h}_2 = 0$, but an approximate form of favorable propagation is achieved in non-line-of-sight scenarios with rich scattering:

**Lemma 1.** *Suppose that* $\mathbf{h}_1 \in \mathbb{C}^M$ *and* $\mathbf{h}_2 \in \mathbb{C}^M$ *have independent random entries with zero mean, identical distribution, and bounded fourth-order moments, then*

$$\frac{\mathbf{h}_1^H \mathbf{h}_2}{M} \to 0 \tag{3}$$

*almost surely as* $M \to \infty$.

*Proof.* This is a consequence of the law of large numbers. A direct proof is provided along with Theorem 3.7 in [14]. ■

This lemma shows that the inner product between $\mathbf{h}_1$ and $\mathbf{h}_2$, if normalized with the number of BS antennas, goes asymptotically to zero as $M$ increases. We refer to this as *asymptotic favorable propagation* and note that this phenomenon explains the behaviors in Fig. 2; the difference between having no inter-user interference and suppressing the interference by ZF becomes smaller and smaller as the number of antennas increases, because the loss in desired signal gain when using ZF reduces when the user channels become more orthogonal.

One special case in which Lemma 1 holds is $\mathbf{h}_1, \mathbf{h}_2 \sim \mathbb{C}\mathcal{N}(\mathbf{0}, \mathbf{I}_M)$, where $\mathbb{C}\mathcal{N}(\cdot, \cdot)$ denotes a multi-variate circularly symmetric complex Gaussian distribution and $\mathbf{I}_M$ is the $M \times M$ identity matrix. This is known as *uncorrelated Rayleigh fading* and in this case one can even prove that the variance of the inner product in (3) is $1/M$ and thus decreases linearly with the number of antennas [32]. Many academic works on Massive MIMO systems consider Rayleigh fading channels, due to the analytic tractability of Gaussian distributions. Nevertheless, Lemma 1 shows that asymptotic favorable propagation holds for other random channel distributions as well. This mathematical result can be extended to also include correlation between the elements in a channel vector. One can also derive similar analytic results for line-of-sight propagation [32] and behaviors that resemble asymptotic favorable propagation have been observed also in the real-world multi-user MIMO channel measurements presented in [16, 19].

**Third lesson learned**: Most wireless channels seem to provide asymptotic favorable propagation.

This lesson is yet another reason to design multi-user MIMO systems with $M \gg K$. It is, however, important to note that $(\mathbf{h}_1^H \mathbf{h}_2)/M \to 0$, as $M \to \infty$, does not imply that $\mathbf{h}_1^H \mathbf{h}_2 \to 0$. Strict favorable propagation is unlikely to appear in practical or theoretical channels. In fact, the inner product $\mathbf{h}_1^H \mathbf{h}_2$ grows roughly as $\sqrt{M}$ for Rayleigh fading channels. The key point is that this correlation has a negligible impact, since the SE depends on $(\mathbf{h}_1^H \mathbf{h}_2)/M$ which goes to zero roughly as $1/\sqrt{M}$. Moreover, the main suppression of inter-user interference appears already at relatively small number of antennas due to the square root.

## 2.3 Massive MIMO Concept

The Massive MIMO concept was proposed in the seminal paper [28] and described in the patent [29], both of which have received numerous scientific awards. Massive MIMO takes multi-user MIMO communications to a new level by designing a highly scalable communication protocol that utilizes the three lessons described in Sect. 2.2. The basic information and communication theoretic limits of this 5G technology were established in early works such as [7, 20, 21, 23, 30]. In this chapter we define Massive MIMO as follows:

**Massive MIMO** is a multi-user MIMO system with $M$ antennas and $K$ users per BS. The system is characterized by $M \gg K$ and operates in TDD mode using linear uplink and downlink processing.

This definition does not manifest any particular ratio between $M$ and $K$, or any particular orders of magnitude that these parameters should have. One attractive example is a system with $M$ in the range of 100–200 antennas, serving between $K = 1$ and $K = 40$ users depending on the data traffic variations. The first public real-time implementation of Massive MIMO is the LuMaMi testbed described in [41], which features $M = 100$ and $K = 10$. We stress that other definitions of Massive MIMO are available in other works and can both be more restrictive (e.g., require certain dimensionality of $M$ and $K$) and looser (e.g., also include FDD mode), but in this chapter we only consider the definition above.

The BS antenna array typically consists of $M$ dipole antennas, each having an effective size $\lambda/2 \times \lambda/2$, where $\lambda$ is the wavelength. This means that an array area of 1 m$^2$ can fit 100 antennas at a 1.5 GHz carrier frequency and 400 antennas at 3 GHz. Each antenna is attached to a separate transceiver chain, so that the system can access the individual received signal at each antenna and select the individual signals to be transmitted from each antenna. The array can have any geometry; linear, rectangular, cylindrical, and distributed arrays are described in [25]. It is important to note that no model of the array geometry is exploited in the Massive MIMO processing, thus the antennas can be deployed arbitrarily without any geometrical array calibration.

The basic Massive MIMO transmission protocol is illustrated in Fig. 4. The time–frequency resources are divided into blocks of size $B_c$ Hz and $T_c$ s, with the purpose of making each user channel approximately frequency-flat and static within a block. Hence, the bandwidth $B_c$ is selected to be smaller or equal to the anticipated channel coherence bandwidth among the users, while $T_c$ is smaller or equal to the anticipated channel coherence time of the users. For this particular reason, each

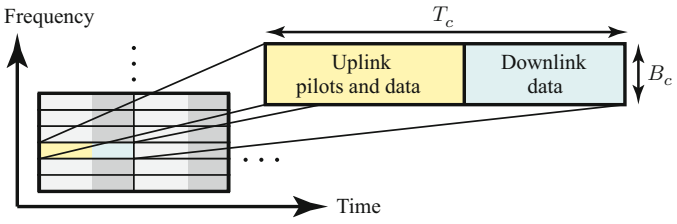


**Fig. 4** Illustration of the basic Massive MIMO transmission protocol, where the time–frequency resources are divided into coherence intervals, each containing $\tau_c = B_c T_c$ transmission symbols. Each coherence interval contains uplink pilot sequences and can be used for both uplink and downlink payload data transmission based on TDD operation

block is referred to as a *coherence interval*. The number of transmission symbols that fit into a coherence interval is given by $\tau_c = B_c T_c$, due to the Nyquist-Shannon sampling theorem. The dimensionality of the coherence interval depends greatly on the anticipated system application. For example, a coherence interval of $\tau_c = 200$ symbols can be obtained with $B_c = 200\,\text{kHz}$ and $T_c = 1\,\text{ms}$, which supports highway user velocities in urban environments at 2 GHz carrier frequencies. Much larger coherence intervals (e.g., $\tau_c$ at the order of $10^3$ or $10^4$) can be obtained by limiting the application to scenarios with low user mobility and short delay spread.

Each coherence interval is operated in TDD mode and can contain both downlink and uplink payload transmissions. To enable channel estimation at the BS, $\tau_p$ of the symbols in each coherence interval are allocated for uplink transmission of pilot sequences (where $\tau_p \geq K$), while the remaining $\tau_c - \tau_p$ symbols can be allocated arbitrarily between uplink and downlink payload data transmissions.

We let $\gamma^{\text{UL}}$ and $\gamma^{\text{DL}}$ denote the fractions of uplink and downlink payload transmission, respectively. This means that the uplink contains $\gamma^{\text{UL}}(\tau_c - \tau_p)$ data symbols and the downlink contains $\gamma^{\text{DL}}(\tau_c - \tau_p)$ data symbols per coherence interval. Naturally, these fractions satisfies $\gamma^{\text{UL}} + \gamma^{\text{DL}} = 1$ and $\gamma^{\text{UL}}, \gamma^{\text{DL}} \geq 0$. Notice that no downlink pilots are assumed in this protocol, since the effective precoded channels converge to their mean values when the BS has many antennas (due to the law of large numbers). It is certainly possible to also send a small amount of downlink pilots, particularly for estimating the small fading variations of the effective precoded channels, but the additional gains from doing this appears to be small in many relevant Massive MIMO cases [31].

Based on this definition of Massive MIMO, the next sections analyze how large SEs that the transmission protocol can offer in 5G cellular networks.

## 3 Performance Analysis

In this section, we describe the uplink detection and downlink precoding of a Massive MIMO network, and analyze the achievable system performance. We consider a basic Massive MIMO network comprising $L$ cells, each consisting of a BS with $M$ antennas and $K$ single-antenna user terminals.

The channel response between the $l$th BS and user $k$ in the $i$th cell is denoted by $\mathbf{h}_{i,k}^{l} = [h_{i,k,1}^{l} \ \ldots \ h_{i,k,M}^{l}]^T \in \mathbb{C}^M$, where $(\cdot)^T$ denotes the transpose. These channel vectors are ergodic random variables that are assumed to take new independent realizations in each coherence interval; recall the Massive MIMO protocol described in Sect. 2.3. To show that the general concept of Massive MIMO is applicable in any propagation environment, we keep the performance analysis general by only defining the basic statistical channel properties: the mean value and variance of each channel coefficient $h_{i,k,m}^{l}$ (note that $m$ stands for the $m$th antenna at BS $l$, for $m = 1, \ldots, M$). We let

$$\bar{\mathbf{h}}_{i,k}^{l} = \mathbb{E}\{\mathbf{h}_{i,k}^{l}\} = [\bar{h}_{i,k,1}^{l} \ \ldots \ \bar{h}_{i,k,M}^{l}]^{T} \tag{4}$$

denote the vector of mean values. The variance of the $m$th coefficient of $\mathbf{h}_{i,k}^{l}$ is denoted by

$$\beta_{i,k}^{l} = \mathbb{V}\{h_{i,k,m}^{l}\}, \tag{5}$$

which is independent of the antenna index $m$ (assuming that the large-scale fading is stationary over the BS array). We also assume that each BS and user can keep perfect track of these long-term statistical properties, and that the user channels are statistically independent.

Using these channel properties, we now analyze the uplink and the downlink.

## 3.1 Uplink with Linear Detection

For each uplink symbol, the received baseband signal $\mathbf{y}_l \in \mathbb{C}^M$ at the $l$th BS is modeled as

$$\mathbf{y}_l = \sum_{i=1}^{L} \sum_{k=1}^{K} \mathbf{h}_{i,k}^{l} \sqrt{p_{i,k}} x_{i,k} + \mathbf{n}_l, \tag{6}$$

where $x_{i,k}$ is the normalized transmission symbol (with $\mathbb{E}\{|x_{i,k}|^2\} = 1$) and $p_{i,k}$ is the transmit power of user $k$ in cell $i$. The receiver hardware at the BS is contaminated by additive white noise, as modeled by the vector $\mathbf{n}_l \in \mathbb{C}^M$ which is zero-mean circularly symmetric complex Gaussian distributed with variance $\sigma_{\text{UL}}^2$; that is, $\mathbf{n}_l \sim \mathbb{C}\mathcal{N}(\mathbf{0}, \sigma_{\text{UL}}^2 \mathbf{I}_M)$.

The matrix notations $\mathbf{H}_i^l = [\mathbf{h}_{i,1}^l \ \ldots \ \mathbf{h}_{i,K}^l] \in \mathbb{C}^{M \times K}$, $\mathbf{P}_i = \text{diag}(p_{i,1}, \ldots, p_{i,K}) \in \mathbb{C}^{K \times K}$, and $\mathbf{x}_i = [x_{i,1} \ \ldots \ x_{i,K}]^T \in \mathbb{C}^K$ can be used to write the multi-cell multi-user MIMO system model from (6) in a compact matrix form:

$$\mathbf{y}_l = \sum_{i=1}^{L} \mathbf{H}_i^l \mathbf{P}_i^{1/2} \mathbf{x}_i + \mathbf{n}_l. \tag{7}$$

The channels $\mathbf{h}_{i,k}^l$ need to be estimated at BS $l$ to perform good detection and this is done in the uplink by letting each user transmit a sequence of $\tau_p$ pilot symbols; see Fig. 4. We let $\tau_p = fK$ for some positive integer $f$ (e.g., $1, 2, \ldots$) which is called the *pilot reuse factor*. This allows for linear independence between a total of $\tau_p$ different pilot sequences. This is, by design, sufficient to allocate independent pilot sequence to the $K$ users in each cell and to also divide the $L$ cells into $f$ disjoint cell groups having fully independent pilot sequences. The benefit of having multiple cell groups is reduced interference during the pilot transmission and the corresponding gains in estimation quality are quantified below.

The uplink received signal $\mathbf{Y}_l^{\text{pilot}} \in \mathbb{C}^{M \times \tau_p}$ at the $l$th BS during pilot transmission is

$$\mathbf{Y}_l^{\text{pilot}} = \sum_{i=1}^{L} \mathbf{H}_i^l \mathbf{P}_i^{1/2} \boldsymbol{\Phi}_i^H + \mathbf{N}_l \tag{8}$$

and collects the received signal from (7) over the $\tau_p$ pilot symbols. Here, $\boldsymbol{\Phi}_i = [\boldsymbol{\phi}_{i,1} \ldots \boldsymbol{\phi}_{i,K}] \in \mathbb{C}^{\tau_p \times K}$ denotes the pilot matrix used by the $K$ users in the $i$th cell, where $\boldsymbol{\phi}_{i,k} \in \mathbb{C}^{\tau_p}$ is the pilot sequence used by the $k$th user in that cell. The pilot matrix satisfies $\boldsymbol{\Phi}_i^H \boldsymbol{\Phi}_i = \tau_p \mathbf{I}_K$. Moreover, $\boldsymbol{\Phi}_l^H \boldsymbol{\Phi}_i = \tau_p \mathbf{I}_K$ if cell $l$ and cell $j$ belong to the same cell group (i.e., use the same set of pilots), while $\boldsymbol{\Phi}_l^H \boldsymbol{\Phi}_i = \mathbf{0}$ if the two cells belong to different cell groups. For notational convenience, we let $\mathscr{P}_l \subset \{1, \ldots, L\}$ denote the set of cell indices that belong to the same cell group as cell $l$, including $l$ itself. Some particular examples are given later in Fig. 7.

By using the channel mean and variances, defined in the beginning of Sect. 3, we can use the linear minimum mean square error (LMMSE) estimator to separately acquire each element of $\mathbf{h}_{i,k}^l$ from the received pilot signal (8), which was proposed in [37] as a low-complexity estimation scheme. The channel estimate $\hat{\mathbf{h}}_{i,k}^l$ related to the true channel response $\mathbf{h}_{i,k}^l$ is given by the following lemma.

**Lemma 2.** *Suppose that BS $l$ estimates each channel coefficient separately from its received signal (8) using an LMMSE estimator. BS $l$ can then estimate the channel to the $k$th user in the $j$th cell as*

$$\hat{\mathbf{h}}_{j,k}^l = \bar{\mathbf{h}}_{j,k}^l + \frac{\sqrt{p_{j,k}}\beta_{j,k}^l}{\sum_{i \in \mathscr{P}_j} p_{i,k}\tau_p \beta_{i,k}^l + \sigma_{\text{UL}}^2} \left( \mathbf{Y}_l^{\text{pilot}} \boldsymbol{\phi}_{j,k} - \sum_{i \in \mathscr{P}_j} \sqrt{p_{i,k}}\tau_p \bar{\mathbf{h}}_{i,k}^l \right). \tag{9}$$

*Each element of the uncorrelated estimation error $\mathbf{e}_{j,k}^l = \mathbf{h}_{j,k}^l - \hat{\mathbf{h}}_{j,k}^l$ has zero mean and the variance*

$$\text{MSE}_{j,k}^l = \beta_{j,k}^l \left( 1 - \frac{p_{j,k}\tau_p \beta_{j,k}^l}{\sum_{i \in \mathscr{P}_j} p_{i,k}\tau_p \beta_{i,k}^l + \sigma_{\text{UL}}^2} \right). \tag{10}$$

*Proof.* The proof is available in Appendix at the end of this chapter. ■

It is worth emphasizing that the estimation error variance in (10) is independent of $M$, thus the estimation quality per channel coefficient is not affected by adding more antennas at the BS. Note that Lemma 2 holds for any correlation between the channel coefficients, since each coefficient is estimated separately. If the channel coefficients are correlated, with a known correlation structure and distribution, the estimation quality would improve with the number of antennas if the estimator is modified appropriately [4]. We also stress that the estimation error is only affected

by noise and interference from the users in the same cell group that are allocated the same pilot sequence. In addition, we notice that the estimate in (9) can be computed using elementary linear algebra operations, with low computational complexity.

Using the channel estimates derived in Lemma 2, in this chapter we analyze the performance of a Massive MIMO network with non-cooperative BSs. During uplink payload data transmission this means that the BS in cell $l$ only utilizes its own received signal $\mathbf{y}_l$ in (6) and only targets to detect the signals sent by its own $K$ users. Signals coming from users in other cells are perceived as inter-cell interference and eventually treated as additional noise. The BS in cell $l$ discriminates the signal transmitted by its $k$th user from the interference by multiplying the received signal in (6) with a linear detection vector $\mathbf{v}_{l,k} \in \mathbb{C}^M$ as follows:

$$\begin{aligned}\mathbf{v}_{l,k}^H \mathbf{y}_l &= \sum_{i=1}^{L} \sum_{t=1}^{K} \mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l \sqrt{p_{i,t}} x_{i,t} + \mathbf{v}_{l,k}^H \mathbf{n}_l \\ &= \underbrace{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l \sqrt{p_{l,k}} x_{l,k}}_{\text{Desired signal}} + \underbrace{\sum_{\substack{t=1 \\ t \neq k}}^{K} \mathbf{v}_{l,k}^H \mathbf{h}_{l,t}^l \sqrt{p_{l,t}} x_{l,t}}_{\text{Intra-cell interference}} + \underbrace{\sum_{\substack{i=1 \\ i \neq l}}^{L} \sum_{t=1}^{K} \mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l \sqrt{p_{i,t}} x_{i,t}}_{\text{Inter-cell interference}} + \underbrace{\mathbf{v}_{l,k}^H \mathbf{n}_l}_{\text{Residual noise}}\end{aligned} \tag{11}$$

where $x_{i,t}$ is the transmitted data symbol from user $t$ in cell $i$. As seen from (11), the processed received signal is the superposition of four parts: the desired signal, intra-cell interference, inter-cell interference, and residual noise. Since the linear detection vector $\mathbf{v}_{l,k}$ appears in all these terms, it can be used to amplify the desired signal, suppress the interference, and/or suppress the noise. More precisely, by gathering the detection vectors at BS $l$ in matrix form as $\mathbf{V}_l = [\mathbf{v}_{l,1} \ldots \mathbf{v}_{l,K}] \in \mathbb{C}^{M \times K}$, there are two main schemes being considered in the Massive MIMO literature: maximum ratio (MR) and zero-forcing (ZF). These are given by

$$\mathbf{V}_l = \begin{cases} \hat{\mathbf{H}}_l^l, & \text{for MR,} \\ \hat{\mathbf{H}}_l^l \left( (\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l \right)^{-1}, & \text{for ZF.} \end{cases} \tag{12}$$

MR detection exploits the $M$ observations in $\mathbf{y}_l$ to maximize the ratio between the average signal gain in (11) and the norm of the detection vector:

$$\mathbb{E}\left\{ \frac{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l}{\|\mathbf{v}_{l,k}\|} \right\} = \frac{\mathbf{v}_{l,k}^H \hat{\mathbf{h}}_{l,k}^l}{\|\mathbf{v}_{l,k}\|} \leq \|\hat{\mathbf{h}}_{l,k}^l\| \tag{13}$$

where the expectation is computed with respect to the zero-mean channel estimation error. The inequality in (13) is satisfied with equality by $\mathbf{v}_{l,k} = \hat{\mathbf{h}}_{l,k}^l$ (leading to MR detection with $\mathbf{V}_l = \hat{\mathbf{H}}_l^l$). In contrast, the ZF detection matrix utilizes the $M$ observations over the antennas to minimize the average intra-cell interference, while retaining the desired signals:

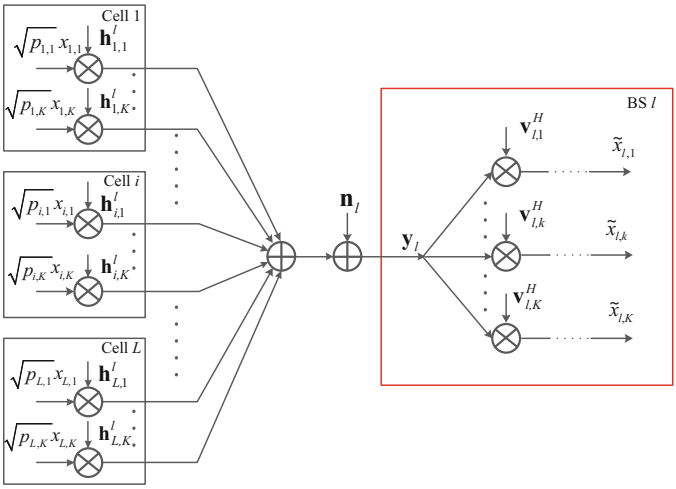


**Fig. 5** Block diagram of the uplink transmission with linear detection in a multi-cell multi-user MIMO network, where BS $l$ receives a linear combination of the signals transmitted from all $K$ users in all $L$ cells

$$\mathbb{E}\{\mathbf{V}_l^H \mathbf{H}_l \mathbf{P}_l^{1/2} \mathbf{x}_l\} = \mathbf{V}_l^H \hat{\mathbf{H}}_l^l \mathbf{P}_l^{1/2} \mathbf{x}_l = \left( (\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l \right)^{-1} \left( (\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l \right) \mathbf{P}_l^{1/2} \mathbf{x}_l = \mathbf{P}_l^{1/2} \mathbf{x}_l \quad (14)$$

where the expectation is computed with respect to the zero-mean channel estimation error and the second equality follows from the ZF detection matrix definition. The average processed signal becomes $\mathbf{P}_l^{1/2} \mathbf{x}_l = [\sqrt{p_{l,1}} x_{l,1} \ \ldots \ \sqrt{p_{l,K}} x_{l,K}]^T$, which contains no intra-cell interference. Note that the inverse of the $K \times K$ matrix $(\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l$ only exists if $M \geq K$. There are also multi-cell variants of ZF detection that can be used to cancel out inter-cell interference; see for example [2] and [9].

A block diagram of the uplink transmission with linear detection is provided in Fig. 5. The purpose of the detection is to make the detected signal $\tilde{x}_{l,k}$ at BS $l$ equal to the true signal $x_{l,k}$, at least up to a scaling factor. Due to noise and estimation errors, there is always a mismatch between the signals which is why the communication link has a limited capacity. If the true signal $x_{l,k}$ originates from a discrete constellation set $\mathcal{X}$ (e.g., a quadrature amplitude modulation (QAM)), $\tilde{x}_{l,k}$ is selected based on $\mathbf{v}_{l,k}^H \mathbf{y}_l$ by finding the minimum distance over all the candidates $x \in \mathcal{X}$:

$$\tilde{x}_{l,k} = \min_{x \in \mathcal{X}} \left| \mathbf{v}_{l,k}^H \mathbf{y}_l - \mathbf{v}_{l,k}^H \hat{\mathbf{h}}_{l,k}^l \sqrt{p_{l,k}} x \right|^2. \quad (15)$$

This expression can be utilized to compute bit error rates and similar uncoded performance metrics. Since modern communication systems apply channel coding over relatively long data blocks, to protect against errors, the ergodic channel capacity is a more appropriate performance metric in 5G networks. It merits to

note that the ergodic capacities of the individual communication links are hard to characterize exactly, particularly under imperfect channel knowledge, but tractable lower bounds are obtained by the following theorem.

**Theorem 1.** *In the uplink, a lower bound on the ergodic capacity of an arbitrary user $k$ in cell $l$ is*

$$R_{l,k}^{\mathrm{UL}} = \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \mathrm{SINR}_{l,k}^{\mathrm{UL}}\right), \tag{16}$$

*where the signal-to-interference-and-noise ratio (SINR) is*

$$\mathrm{SINR}_{l,k}^{\mathrm{UL}} = \frac{p_{l,k} \left|\mathbb{E}\left\{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l\right\}\right|^2}{\sum\limits_{i=1}^{L} \sum\limits_{t=1}^{K} p_{i,t} \mathbb{E}\left\{\left|\mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l\right|^2\right\} - p_{l,k} \left|\mathbb{E}\left\{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l\right\}\right|^2 + \sigma_{\mathrm{UL}}^2 \mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}}. \tag{17}$$

*Proof.* The proof is available in Appendix at the end of this chapter. ■

Theorem 1 demonstrates that the achievable SE of an arbitrary user $k$ in cell $l$ in a Massive MIMO network can be described by an SINR term $\mathrm{SINR}_{l,k}^{\mathrm{UL}}$ that contains expectations with respect to the small-scale channel fading. The numerator contains the gain of the desired signal, while the denominator contains three different terms. The first term is the average power of all the signals, including both multi-user interference and the desired signal, while the second term subtracts the part of the desired signal power that is usable for decoding. The third term is the effective noise power. The pre-log factor $(1 - \frac{\tau_p}{\tau_c})$ compensates for the fact that $\tau_p/\tau_c$ of the transmission symbols contain pilots instead of payload data. The SE is also multiplied by $\gamma^{\mathrm{UL}}$, which was defined earlier as the fraction of uplink data. Clearly, MR detection aims at maximizing the numerator of $\mathrm{SINR}_{l,k}^{\mathrm{UL}}$, while ZF detection tries to minimize the intra-cell interference.

The expectations in Theorem 1 can be computed numerically for any channel distribution and any detection scheme. In the case of MR detection, the desired signal gain $|\mathbb{E}\{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l\}|^2$ grows as $M^2$ for most channel distributions, while the noise term $\sigma_{\mathrm{UL}}^2 \mathbb{E}\{\|\mathbf{v}_{l,k}\|^2\}$ only grows as $M$ and thus becomes less significant the more antennas are deployed at the BS. This property is known as the *array gain* from coherent detection. The behavior of the multi-user interference terms greatly depends on the channel distribution, but typically these terms will also have the slower scaling of $M$ [32], except for users that interfered with each other during pilot transmission (i.e., appeared in each other's expressions (10) for the estimation error variance). The latter is a phenomenon called *pilot contamination* and is further discussed later in this chapter.

To demonstrate these properties in detail, we now consider the special case in which the channel between BS $l$ and user $k$ in cell $i$ is uncorrelated Rayleigh fading:

$$\mathbf{h}_{i,k}^l \sim \mathbb{C}\mathcal{N}\left(\mathbf{0}, \beta_{i,k}^l \mathbf{I}_M\right). \tag{18}$$

Hence, $\bar{\mathbf{h}}_{i,k}^{l} = \mathbb{E}\{\mathbf{h}_{i,k}^{l}\} = \mathbf{0}$, which means that there is no line-of-sight channel component. This special case is relevant in rich-scattering environments where the channel does not have any statistically dominating directivity.

Subsequently, the LMMSE estimate in Lemma 2 simplifies to

$$\hat{\mathbf{h}}_{j,k}^{l} = \frac{\sqrt{p_{j,k}}\beta_{j,k}^{l}}{\sum\limits_{i\in\mathscr{P}_j} p_{i,k}\tau_p\beta_{i,k}^{l} + \sigma_{\mathrm{UL}}^2}\mathbf{Y}_{l}^{\mathrm{pilot}}\boldsymbol{\phi}_{j,k} \tag{19}$$

and becomes circularly-symmetric complex Gaussian distributed:

$$\hat{\mathbf{h}}_{j,k}^{l} \sim \mathbb{C}\mathcal{N}\left(\mathbf{0}, (\beta_{j,k}^{l} - \mathrm{MSE}_{j,k}^{l})\mathbf{I}_M\right). \tag{20}$$

There is an important relationship between the two estimated channels $\hat{\mathbf{h}}_{l,k}^{l}$ and $\hat{\mathbf{h}}_{i,k}^{l}$ for cell indices $i$ and $l$ such as $i \in \mathscr{P}_l$, expressed by

$$\hat{\mathbf{h}}_{i,k}^{l} = \frac{\sqrt{p_{i,k}}\beta_{i,k}^{l}}{\sqrt{p_{l,k}}\beta_{l,k}^{l}}\hat{\mathbf{h}}_{l,k}^{l}. \tag{21}$$

This equation shows that BS $l$ cannot tell apart the channels of users that send the same pilot sequence; the estimates are the same up to a scaling factor. This fact is the cause of pilot contamination and will have a key impact on the performance, as shown later.

Moreover, the LMMSE estimator in (19) is also the MMSE estimator in the special case of Rayleigh fading, since the channels are Gaussian distributed [24]. By using these key properties, the ergodic SE in Theorem 1 can be computed in closed form for MR and ZF detection, as shown by the following corollary.

**Corollary 1.** *In the uplink, if all channels are uncorrelated Rayleigh fading, the lower bound on the ergodic capacity of user k in cell l stated in Theorem 1 becomes*

$$R_{l,k}^{\mathrm{UL}} = \gamma^{\mathrm{UL}}\left(1 - \frac{\tau_p}{\tau_c}\right)\log_2\left(1 + \mathrm{SINR}_{l,k}^{\mathrm{UL}}\right), \tag{22}$$

*where the SINR is*

$$\mathrm{SINR}_{l,k}^{\mathrm{UL}} = \frac{Gp_{l,k}\beta_{l,k}^{l}\frac{p_{l,k}\tau_p\beta_{l,k}^{l}}{\sum_{i'\in\mathscr{P}_l}p_{i',k}\tau_p\beta_{i',k}^{l}+\sigma_{\mathrm{UL}}^2}}{G\sum\limits_{i\in\mathscr{P}_l\setminus\{l\}}p_{i,k}\beta_{i,k}^{l}\frac{p_{i,k}\tau_p\beta_{i,k}^{l}}{\sum_{i'\in\mathscr{P}_l}p_{i',k}\tau_p\beta_{i',k}^{l}+\sigma_{\mathrm{UL}}^2} + \sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K}p_{i,t}z_{i,t}^{l} + \sigma_{\mathrm{UL}}^2} \tag{23}$$

*and the parameters G and $z_{i,t}^{l}$ depend on the choice of detection scheme. MR gives $G = M$ and $z_{i,t}^{l} = \beta_{i,t}^{l}$, while ZF gives $G = M - K$ and*

$$z_{i,t}^{l} = \begin{cases} \mathrm{MSE}_{i,t}^{l}, & \text{for } i \in \mathscr{P}_l, \\ \beta_{i,t}^{l}, & \text{otherwise.} \end{cases}$$

*Proof.* The proof is available in Appendix at the end of this chapter. ■

The closed-form achievable SE expressions in Corollary 1 provide many insights on the advantages of spatial multi-user multiplexing and the effects of channel estimation. Firstly, the desired signal term in the numerator of (23) scales with the number of BS antennas, proportionally to $M$ and $M - K$ with MR and ZF, respectively. This array gain is multiplied with the average received signal power per antenna, $p_{l,k}\beta_{l,k}^l$, and the relative channel estimation quality

$$\frac{p_{l,k}\tau_p\beta_{l,k}^l}{\sum_{i' \in \mathscr{P}_l} p_{i',k}\tau_p\beta_{i',k}^l + \sigma_{\text{UL}}^2}, \tag{24}$$

which is a number between 0 and 1 (where 1 is perfect CSI and 0 is no CSI).

Secondly, we notice that the first term of the denominator in (23) has a similar structure as the desired signal and represents the coherent pilot contamination—interference that is amplified along with the desired signals due the BS's inability to tell apart users that use the same pilot sequence. The pilot contamination degrades the SINR by adding additional interference that scales as $M$ or $M - K$, depending on the detection scheme. However, since pilot contamination only arises at BS $l$ from the interfering user in cell $i$ in $\mathscr{P}_l$, the network can suppress pilot contamination by increasing the pilot reuse factor $f$ and by designing the cell groups appropriately. To understand how to suppress pilot contamination, we have a look at the ratio between the pilot contamination term and the signal term in (23):

$$\frac{G \sum\limits_{i \in \mathscr{P}_l \setminus \{l\}} p_{i,k}\beta_{i,k}^l \frac{p_{i,k}\tau_p\beta_{i,k}^l}{\sum_{i' \in \mathscr{P}_l} p_{i',k}\tau_p\beta_{i',k}^l + \sigma_{\text{UL}}^2}}{G p_{l,k}\beta_{l,k}^l \frac{p_{l,k}\tau_p\beta_{l,k}^l}{\sum_{i' \in \mathscr{P}_l} p_{i',k}\tau_p\beta_{i',k}^l + \sigma_{\text{UL}}^2}} = \sum_{i \in \mathscr{P}_l \setminus \{l\}} \left( \frac{p_{i,k}\beta_{i,k}^l}{p_{l,k}\beta_{l,k}^l} \right)^2. \tag{25}$$

This ratio represents the relative strength of the pilot contamination and (25) should preferably be small. The pilot contamination caused by UE $k$ in cell $i$ is small whenever $\beta_{i,k}^l / \beta_{l,k}^l$ is small, which occurs when either $\beta_{l,k}^l$ is large (i.e., the desired user is close to its serving BS) or $\beta_{i,k}^l$ is small (i.e., the interfering cell is far away). The cell groups should be designed based on these properties, and this issue is further discussed in Sect. 4.2.

Thirdly, the performance in Corollary 1 is also affected by classical noise and interference. Since MR focuses only on maximizing the SNR, the interference term $\sum_{i=1}^{L} \sum_{t=1}^{K} p_{i,t}\beta_{i,t}^l$ is simply the average signal power received at any antenna of BS $l$. In contrast, ZF pays attention to the intra-cell interference and takes no notice of the noise. The interference suppression replaces the full channel variance $\beta_{i,t}^l$ in the aforementioned interference summation with the estimation error variance $\text{MSE}_{i,t}^l$ for cells $i \in \mathscr{P}_l$. Due to the imperfect CSI (i.e., $\text{MSE}_{i,t}^l > 0$) not all intra-cell interference can be removed by ZF. However, the pilot contamination also has the positive effect that not only intra-cell interference is suppressed, but also the inter-cell interference coming from other users in the same cell group (which use

the same pilots as in cell $l$). The fact that the interference and noise terms are independent of $M$, while the desired signal scales with $M$, is a consequence of the asymptotic favorable propagation that was described in Sect. 2.2.1.

If we limit the scope to a single-cell network, achievable SE expressions can be obtained directly from Corollary 1 by simply setting $\mathscr{P}_l = \{l\}$ and removing the interference from all other cells $j \in \{1, \ldots, L\} \setminus \{l\}$. For simplicity of exposition, we leave out the cell index $l$ in this special case.

**Corollary 2.** *In the single-cell uplink, if all channels are uncorrelated Rayleigh fading, a lower bound on the ergodic SE of an arbitrary user k is given by*

$$R_k^{\mathrm{UL}} = \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \frac{G p_k^2 \tau_p \beta_k^2}{\left(p_k \tau_p \beta_k + \sigma_{\mathrm{UL}}^2\right)\left(\sum\limits_{t=1}^{K} p_t z_t + \sigma_{\mathrm{UL}}^2\right)}\right). \tag{26}$$

*Here, the parameters G and $z_t$ depend on the detection scheme. MR gives $G = M$ and $z_t = \beta_t$, while ZF gives $G = M - K$ and $z_t = \frac{\beta_t \sigma_{\mathrm{UL}}^2}{p_t \tau_p \beta_t + \sigma_{\mathrm{UL}}^2}$.*

This corollary shows that the spatial multi-user multiplexing capability is even greater in isolated single-cell networks. The most notable difference compared to a multi-cell network is the lack of inter-cell interference, both during data and pilot transmission. In other words, the interference only originates from users within the own cell, while pilot contamination vanishes thanks to the orthogonality of all pilot sequences in the cell. The SE per cell is therefore higher in single-cell networks than in the multi-cell networks—at least if the cell geometry is the same and we only neglect inter-cell interference. The motivation of having multiple cells is, of course, to cover a larger area and thereby achieve much higher total SE. The scenarios when the interference suppression of ZF is beneficial as compared to MR can be identified from Corollary 2 as the cases when

$$\frac{M - K}{\sum\limits_{t=1}^{K} \frac{p_t \beta_t \sigma_{\mathrm{UL}}^2}{p_t \tau_p \beta_t + \sigma_{\mathrm{UL}}^2} + \sigma_{\mathrm{UL}}^2} > \frac{M}{\sum\limits_{t=1}^{K} p_t \beta_t + \sigma_{\mathrm{UL}}^2}. \tag{27}$$

To summarize, we have derived uplink SE expressions for Massive MIMO networks, for general channel distributions in Theorem 1 and for Rayleigh fading in Corollary 1. In the latter case, the expressions are in closed form and can thus be computed and analyzed directly, without having to simulate any channel fading realizations. These expressions are used in Sect. 4 to illustrate the anticipated performance of Massive MIMO networks.

### 3.2 Downlink with Linear Precoding

Next, we consider the downlink of a Massive MIMO network where the BSs are transmitting signals to their users. For an arbitrary BS $l$, we let $\mathbf{x}_l \in \mathbb{C}^M$ denote the transmitted signal vector intended for its $K$ users. We consider linear precoding where this vector is computed as

$$\mathbf{x}_l = \sum_{t=1}^{K} \sqrt{\rho_{l,t}} \mathbf{w}_{l,t} s_{l,t}, \tag{28}$$

where the payload symbol $s_{l,t}$ intended for user $t$ in cell $l$ has unit transmit power $\mathbb{E}\{|s_{l,t}|^2\} = 1$ and $\rho_{l,t}$ represents the transmit power allocated to this particular user. Moreover, $\mathbf{w}_{l,t} \in \mathbb{C}^M$, for $t = 1, \ldots, K$, are the corresponding linear precoding vectors that determine the spatial directivity of the signal sent to each user.

The received signal $y_{l,k} \in \mathbb{C}$ at user $k$ in cell $l$ is modeled as

$$y_{l,k} = \sum_{i=1}^{L} (\mathbf{h}_{l,k}^{i})^H \mathbf{x}_i + n_{l,k}, \tag{29}$$

where $n_{l,k} \sim \mathbb{C}\mathcal{N}(0, \sigma_{\text{DL}}^2)$ is the additive white noise with variance $\sigma_{\text{DL}}^2$. Notice that $\mathbf{h}_{l,k}^{i}$ is the same channel response as in the uplink, due to the reciprocity of physical propagation channels (within a coherence interval).[1] A block diagram of the downlink transmission is provided in Fig. 6. Since there are no downlink pilots in the Massive MIMO protocol described in Sect. 2.3, the users are assumed to only know the channel statistics. The lack of instantaneous CSI would greatly reduce the performance of small MIMO systems, but Massive MIMO works well without it since the effective precoded channels quickly approach their mean as more antennas are added. Hence, coherent downlink reception is possible using only statistical CSI. This leads to a low-complexity communication solution where all the intelligence is placed at the BS. Since the ergodic capacity is hard to characterize in this case, the following theorem derives a lower bound on the capacity between user $k$ in cell $l$ and its serving BS.

**Theorem 2.** *In the downlink, a lower bound on the ergodic rate an arbitrary user $k$ in cell $l$ is*

$$R_{l,k}^{\text{DL}} = \gamma^{\text{DL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \text{SINR}_{l,k}^{\text{DL}}\right), \tag{30}$$

[1] In fact, the reciprocal channel is $(\mathbf{h}_{l,k}^{i})^T$, using the regular transpose instead of the conjugate transpose as in (29), but since the only difference is a complex conjugation we can characterize the performance using (29) without loss of generality. The reason to use the conjugate transpose is that the notation becomes easier and the relation to the uplink is clearer.

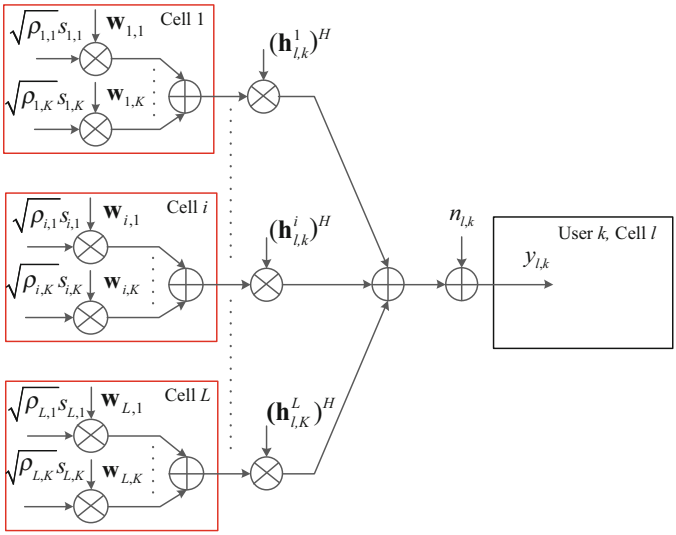


**Fig. 6** Block diagram of the downlink transmission with linear precoding in a multi-cell MIMO system, where BSs equipped with $M$ antennas are transmitting signals that reach user $k$ in cell $l$

*where the SINR is*

$$\mathrm{SINR}_{l,k}^{\mathrm{DL}} = \frac{\rho_{l,k} \left| \mathbb{E}\left\{ (\mathbf{h}_{l,k}^{l})^H \mathbf{w}_{l,k} \right\} \right|^2}{\sum\limits_{i=1}^{L} \sum\limits_{t=1}^{K} \rho_{i,t} \mathbb{E}\left\{ |(\mathbf{h}_{l,k}^{i})^H \mathbf{w}_{i,t}|^2 \right\} - \rho_{l,k} \left| \mathbb{E}\left\{ (\mathbf{h}_{l,k}^{l})^H \mathbf{w}_{l,k}^{l} \right\} \right|^2 + \sigma_{\mathrm{DL}}^2}. \tag{31}$$

*Proof.* The proof is available in Appendix at the end of this chapter. ■

The downlink achievable SE provided in Theorem 2 holds for any channel distributions and choice of precoding vectors. Since the uplink and downlink channels are reciprocal, it would make sense if the uplink and downlink performance were also somehow connected. The downlink achievable SE in Theorem 2 indeed bears much similarity with the corresponding uplink expression in Theorem 1. The desired signal terms are the same, except for the potentially different transmit power parameters and the fact that the detection vector is replaced by the corresponding precoding vector. The interference terms have a similar structure, but the indices are swapped between the channel vector and the processing vector. This is because the uplink interference arrives through different channels for different users while all the downlink interference from a particular cell comes through the same channel from the BS. These observations lead to the following *uplink–downlink duality* [9, 10]:

**Lemma 3.** *Suppose that the downlink precoding vectors are selected as*

$$\mathbf{w}_{l,k} = \frac{\mathbf{v}_{l,k}}{\sqrt{\mathbb{E}\{\|\mathbf{v}_{l,k}\|^2\}}} \tag{32}$$

*based on the uplink detection vectors* $\mathbf{v}_{l,k}$*, for all* $l$ *and* $k$*. For any given uplink powers* $p_{i,t}$ *(for* $i = 1, \ldots, L$ *and* $t = 1, \ldots, K$*), there exist a corresponding set of downlink powers* $\rho_{i,t}$ *(for* $i = 1, \ldots, L$ *and* $t = 1, \ldots, K$*) such that*

$$\mathrm{SINR}_{l,k}^{\mathrm{UL}} = \mathrm{SINR}_{l,k}^{\mathrm{DL}} \tag{33}$$

*for all* $l$ *and* $k$*, and*

$$\frac{\sum_{i=1}^{L}\sum_{t=1}^{K} p_{i,t}}{\sigma_{\mathrm{UL}}^2} = \frac{\sum_{i=1}^{L}\sum_{t=1}^{K} \rho_{i,t}}{\sigma_{\mathrm{DL}}^2}. \tag{34}$$

*Proof.* The proof is available in Appendix at the end of this chapter. ■

This lemma shows that the same performance can be achieved in both the uplink and the downlink, if the downlink power is allocated in a particular way based on the uplink powers and the precoding vectors are selected based on the detection vectors as in (32). The downlink powers are computed according to (72), which is given in Appendix at the end of this chapter since the important thing for now is that there exist a collection of downlink powers that give exactly the same performance in both directions. If $\sigma_{\mathrm{UL}}^2 = \sigma_{\mathrm{DL}}^2$, then the same total transmit power is used in both directions of the Massive MIMO network; however, the power will generally be distributed differently over the users.

Motivated by the uplink–downlink duality, it makes sense to consider MR and ZF precoding as the main downlink precoding schemes. These are defined as

$$\mathbf{w}_{l,k} = \begin{cases} \dfrac{\hat{\mathbf{h}}_{l,k}^{l}}{\sqrt{\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^2\right\}}}, & \text{for MR,} \\ \dfrac{\hat{\mathbf{H}}_{l}^{l}\mathbf{r}_{l,k}}{\sqrt{\mathbb{E}\left\{\|\hat{\mathbf{H}}_{l}^{l}\mathbf{r}_{l,k}\|^2\right\}}}, & \text{for ZF,} \end{cases} \tag{35}$$

where $\mathbf{r}_{l,k}$ denotes the $k$th column of $((\hat{\mathbf{H}}_{l}^{l})^H \hat{\mathbf{H}}_{l}^{l})^{-1}$.

Similar to the uplink performance analysis, we now compute the downlink SE in closed form for uncorrelated Rayleigh fading channels, as defined in (18). Because of the channel reciprocity, the channel estimates obtained at the BSs in the uplink can also be used in the downlink. In particular, the channel estimates $\hat{\mathbf{h}}_{i,k}^{i}$ and $\hat{\mathbf{h}}_{l,k}^{i}$ for cell indices $i$ and $l$ with $l \in \mathscr{P}_i$ are still related as

$$\hat{\mathbf{h}}_{l,k}^{i} = \frac{\sqrt{p_{l,k}}\beta_{l,k}^{i}}{\sqrt{p_{i,k}}\beta_{i,k}^{i}}\hat{\mathbf{h}}_{i,k}^{i}, \tag{36}$$

thus showing that pilot contamination exists also in the downlink; that is, BS $i$ cannot precode signals toward its user $k$ without also precode the signal towards user $k$ in cell $i \in \mathscr{P}_l$. The next corollary specializes Theorem 2 for Rayleigh fading channels.

**Corollary 3.** *In the downlink, if all channels are uncorrelated Rayleigh fading, the lower bound on the ergodic capacity of user k stated in Theorem 2 becomes*

$$R_{l,k}^{\mathrm{DL}} = \gamma^{\mathrm{DL}}\left(1 - \frac{\tau_p}{\tau_c}\right)\log_2\left(1 + \mathrm{SINR}_{l,k}^{\mathrm{DL}}\right), \tag{37}$$

*where the SINR is*

$$\mathrm{SINR}_{l,k}^{\mathrm{DL}} = \frac{G\rho_{l,k}\beta_{l,k}^{l}\frac{p_{l,k}\tau_p\beta_{l,k}^{l}}{\sum_{i'\in\mathscr{P}_l}p_{i',k}\tau_p\beta_{i',k}^{l}+\sigma_{\mathrm{UL}}^2}}{G\sum\limits_{i\in\mathscr{P}_l\setminus\{l\}}\rho_{i,k}\beta_{l,k}^{i}\frac{p_{l,k}\tau_p\beta_{l,k}^{i}}{\sum_{i'\in\mathscr{P}_l}p_{i',k}\tau_p\beta_{i',k}^{i}+\sigma_{\mathrm{UL}}^2} + \sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K}\rho_{i,t}z_{l,k}^{i} + \sigma_{\mathrm{DL}}^2}. \tag{38}$$

*The parameters $G$ and $z_{l,k}^{i}$ are specified by the precoding scheme. MR precoding gives $G = M$ and $z_{l,k}^{i} = \beta_{l,k}^{i}$, while ZF precoding gives $G = M - K$ and*

$$z_{l,k}^{i} = \begin{cases} \mathrm{MSE}_{l,k}^{i}, & \text{for } i \in \mathscr{P}_l, \\ \beta_{l,k}^{i}, & \text{otherwise.} \end{cases}$$

*Proof.* The proof is available in Appendix at the end of this chapter. ■

For Rayleigh fading channels, Corollary 3 shows that the array gain, pilot contamination, and all other attributes of MR and ZF precoding are very similar to the uplink counterparts. Hence, the same kind of observations can be made from Corollary 3 as previously done for Corollary 1.

In the single-cell scenario, the SE expression in Corollary 3 simplifies to the following result.

**Corollary 4.** *In the single-cell downlink, if all channels are uncorrelated Rayleigh fading, a lower bound on the ergodic SE of an arbitrary user k is given by*

$$R_{k}^{\mathrm{DL}} = \gamma^{\mathrm{DL}}\left(1 - \frac{\tau_p}{\tau_c}\right)\log_2\left(1 + \frac{G\rho_k p_k\tau_p\beta_k^2}{\left(p_k\tau_p\beta_k + \sigma_{\mathrm{UL}}^2\right)\left(z_k\sum_{t=1}^{K}\rho_t + \sigma_{\mathrm{DL}}^2\right)}\right). \tag{39}$$

*The parameters $G$ and $z_k$ depend on the precoding scheme. MR gives $G = M$ and $z_k = \beta_k$, while ZF obtains $G = M - K$ and $z_k = \frac{\beta_k\sigma_{\mathrm{UL}}^2}{p_k\tau_p\beta_k + \sigma_{\mathrm{UL}}^2}$.*

We conclude the analytical part of this chapter by recalling that the uplink and downlink spectral efficiencies with Massive MIMO can be easily computed from Theorems 1 and 2 for any channel distributions and processing schemes. In the uncorrelated Rayleigh fading case there are even closed-form expressions. The same SINR performance can be achieved in the uplink and downlink, based on what is known as uplink–downlink duality. The intuition is that the downlink precoding and uplink detection vectors should be the same, but that the power allocation needs to be adapted differently in the two cases.

## 4 Design Guidelines and Anticipated Spectral Efficiency Gains

In this section, we provide some basic design guidelines for Massive MIMO networks and showcase the SEs that the technology can deliver to 5G networks according to the theory developed in Sect. 3. For illustrative purposes, we consider a classic cellular network topology with hexagonal cells, where each cell can be illustrated as in Fig. 1. In other words, the BS is deployed in the center of the cell, while the $K$ users are distributed over the cell area. When many cells of this type are placed next to each other, the cellular network has the shape showed in Fig. 7. While conventional cellular networks use sectorization to split each cell into, say, three static sectors, this is not assumed here. This is because the spatial transceiver processing at the BS in Massive MIMO basically creates $K$ virtual sectors, adapted dynamically to the positions of the current set of users.

### *4.1 Power Allocation*

The average transmit power of user $k$ in cell $j$ is denoted by $p_{j,k}$ in the uplink and by $\rho_{j,k}$ in the downlink. These are important design parameters that determine the SEs of the users; see Theorem 1 (for the uplink) and Theorem 2 (for the downlink). Since inter-user interference is an important factor in any multi-user MIMO system, each transmit power coefficient affects not only the strength of the desired signal at the desired user, but also the amount of interference caused to all the other users in the network (although the interference is most severe within a cell and between neighboring cells). The selection of these transmit power coefficients is referred to as *power allocation* and needs to be addressed properly.

A key property of Massive MIMO is that the small-scale fading in time and the frequency-selective fading variations are negligible, since they essentially average out over the many antennas at each BS. For example, the SE expressions for Rayleigh fading channels in Corollaries 1–4 only depend on the channel variances $\beta_{i,k}^{l}$ and not on the instantaneous realizations of the corresponding channel

vectors $\mathbf{h}_{i,k}^{l}$. Therefore, there is no need to change the power allocation between each coherence interval, but only over the longer time frame where the channel variances change, due to modifications in the large-scale propagation behaviors (e.g., caused by user mobility). This is a substantial increase of the time frame in which power allocation decisions are to be made, from milliseconds to seconds. This fact makes it possible to optimize and coordinate the power allocation across cells, in ways that have not been possible in the past due computational or delay limitations.

A structured approach to power allocation is to find the transmit powers that jointly maximize the network utility functions $U^{\mathrm{UL}}(\{R_{l,k}^{\mathrm{UL}}\})$ and $U^{\mathrm{DL}}(\{R_{l,k}^{\mathrm{DL}}\})$ in the uplink and downlink, respectively. These utilities are increasing functions of the users' SEs, where $\{R_{l,k}^{\mathrm{UL}}\}$ and $\{R_{l,k}^{\mathrm{DL}}\}$ denote the sets of all SEs. Some particular examples of network utility functions are [8]

$$U(\{R_{l,k}\}) = \begin{cases} \sum_{l=1}^{L} \sum_{k=1}^{K} R_{l,k}, & \text{Sum utility,} \\ \prod_{l=1}^{L} \prod_{k=1}^{K} R_{l,k}, & \text{Proportional fairness,} \\ \min_{l \in \{1,\ldots,L\},\, k \in \{1,\ldots,K\}} R_{l,k}, & \text{Max-min fairness,} \end{cases} \tag{40}$$

where we have omitted the uplink/downlink superscripts since the same type of utility function can be utilized in both cases. These utilities are often maximized with respect to a given power budget per user (in the uplink) and per BS (in the downlink). For brevity, we will not provide any further mathematical details, but briefly outline what is known around power allocation for Massive MIMO.

Maximization of the sum utility (SU) provides high SEs to users with good average channel conditions, at the expense of low SE for users with bad average channel conditions. In contrast, max-min fairness (MMF) enforces that each user should get equal SE, which effectively means that users with good channels reduce their SEs to cause less interference to the users with bad channels. Proportional fairness (PF) can be shown to lie in between these extremes. The SU achieves the highest sum SE, since this is really what is optimized by this utility function, while MMF trades some of the sum SE to obtain a uniform user experience. The choice of network utility function is a matter of subjective taste, since there is no objectively optimal utility function [8]. Nevertheless, there seems to be a trend towards more fairness-emphasizing utilities in the Massive MIMO literature [9, 33, 45], motivated by the fact that contemporary networks are designed to provide high peak rates, while the cell edge performance is modest and needs to be improved in 5G. In the uplink, another important aspect to consider in the power allocation is the fact that a BS cannot simultaneously receive desired user signals of very different power levels, since then the weak signals will then drown in the quantization noise caused by the analog-to-digital conversion. Hence, even if the channel attenuation might differ by 50 dB within a cell, these variations need to be brought down to, say, 10 dB by the uplink power allocation.

From a numerical optimization perspective, the downlink power allocation problem (for fixed uplink power allocation) has the same mathematical structure as the seemingly different scenario of single-antenna multi-cell communications

with perfect CSI. The downlink utility optimization can therefore be solved using the techniques described in [3, 27, 35] and references therein. In general, the PF and MMF utilities give rise to convex optimization problems that can be solved efficiently with guaranteed convergence to the global optimum. These algorithms can also be implemented in a distributed fashion [3]. The SU problem is, in contrast, provable non-convex and hard to solve [27], which means that the optimal solution cannot be found under any practical constraints on complexity.

The uplink power allocation is more complicated than power allocation in the downlink; for example, because the SE expression in Corollary 1 contains both $p_{j,k}$ and $p_{j,k}^2$ (while the downlink SE expressions only contain the linear term $\rho_{j,k}$). Nevertheless, there are several efficient algorithms that maximize the MMF utility [12, 13, 45], and the approach in [13] can also maximize the SU and PF utilities with MR and ZF detection. The work [26] provides an alternative methodology to maximize an approximation of the SU metric for other detection methods.

In summary, power allocation is used in Massive MIMO to distribute the sum SE over the individual users. There are plenty of algorithms that can be used to optimize the power allocation, depending on the utility function that is used in the system.

## *4.2 Non-Universal Pilot Reuse*

An important insight from the theoretical analysis in Sect. 3 is that the SE of a particular cell $j$ is influenced by the pilot signaling carried out in other cells. The degradations in CSI estimation quality and pilot contaminated interference are caused only by the interfering cells in $\mathscr{P}_j$ that use the same pilot sequences as cell $j$. Since the channel attenuation of the interference increases with distance, one would like these interfering cells to be as far away from cell $j$ as possible—and the same is desirable for all cells in $\mathscr{P}_j$.

Recall that the pilot reuse factor $f = \tau_p/K$ was assumed to be an integer in Sect. 3, which leads to a division of the $L$ cells into $f$ disjoint cell groups. The case $f = 1$ is known as universal pilot reuse and $f > 1$ is called non-universal pilot reuse. Since the hexagonal cell topology has a multiple of six cells in each tier of interfering cells, the smallest pilot reuse factors that give rise to symmetric pilot reuse patterns are $f = 1, f = 3$, and $f = 4$ [15]. Examples of such reuse patterns are given in Fig. 7, where cells with different colors use different subsets of the pilot sequences. The cells with the same color use exactly the same subset of pilots and therefore cause pilot contamination to each other, while there is no contamination between cells with different colors. If the center cell in Fig. 7 has index $j$, then $\mathscr{P}_j$ is the set of all cells having the same color. By increasing the pilot reuse factor, there are more colors and therefore fewer interfering cells in each group. We note that with a pilot reuse factor of $f = 4$, one can either divide the cells into four different disjoint groups (as in the lower left example in Fig. 7) or divide each cell into two subcells: cell edge and cell center (as in the lower right example in Fig. 7). The latter

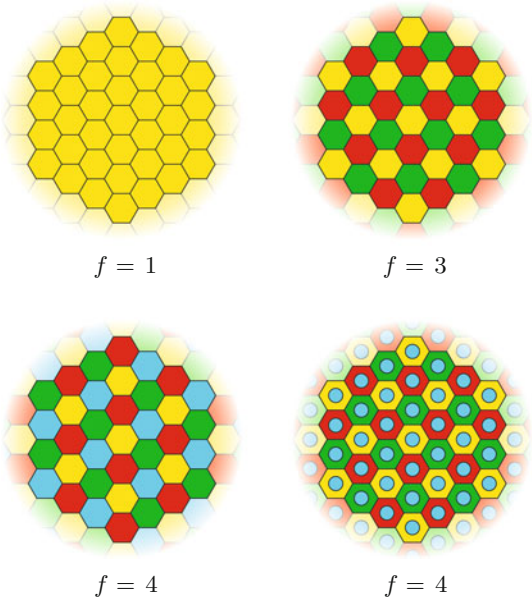


**Fig. 7** Illustration of potential symmetric reuse patterns created by three different pilot reuse factors, $f$, in a cellular network with hexagonal cells. In the *lower right* case, each cell is divided into two sub-cells with different sets of pilots. If $j$ is the index of a particular cell, then $\mathscr{P}_j$ is the index set of all cells having the same color. Only the cells with the same *color* use the same pilot sequences, and thereby degrade each other's CSI estimation quality and cause pilot contaminated interference

is known as *fractional pilot reuse* and can be used to have less frequent pilot reuse at the cell edges than in the cell centers [2], because it is users at the cell edges that are most sensitive to pilot contamination.

To give a concrete example, consider a Massive MIMO scenario with $M = 200$ BS antennas and a coherence interval of $\tau_c = 400$ symbols. The users are assumed to be uniformly distributed in the cell, except for the 10 % cell center, and the channels are modeled as uncorrelated Rayleigh fading with a distant-dependent channel attenuation with pathloss exponent 3.7. We consider the pilot reuse factors $f \in \{1, 3, 4\}$, but not the fractional pilot reuse case. Recall from the uplink–downlink duality in Lemma 3 that the same SE is achievable in the uplink and the downlink, thus it is sufficient to study the uplink. We assume a simple power allocation policy

$$p_{j,k} = \frac{\delta}{\beta_{j,k}^j} \quad j = 1, \ldots, L, \quad k = 1, \ldots, K, \tag{41}$$

where $\delta \geq 0$ is a design parameter that determines the SNR achieved at each BS antenna: $p_{j,k}\beta_{j,k}^j / \sigma_{\text{BS}}^2 = \delta / \sigma_{\text{BS}}^2$.[2] This is called statistical channel inversion power allocation.

[2]This explicit uplink power allocation policy is very similar to what the uplink MMF utility would give [3], but has the benefit of not requiring any numerical optimization.

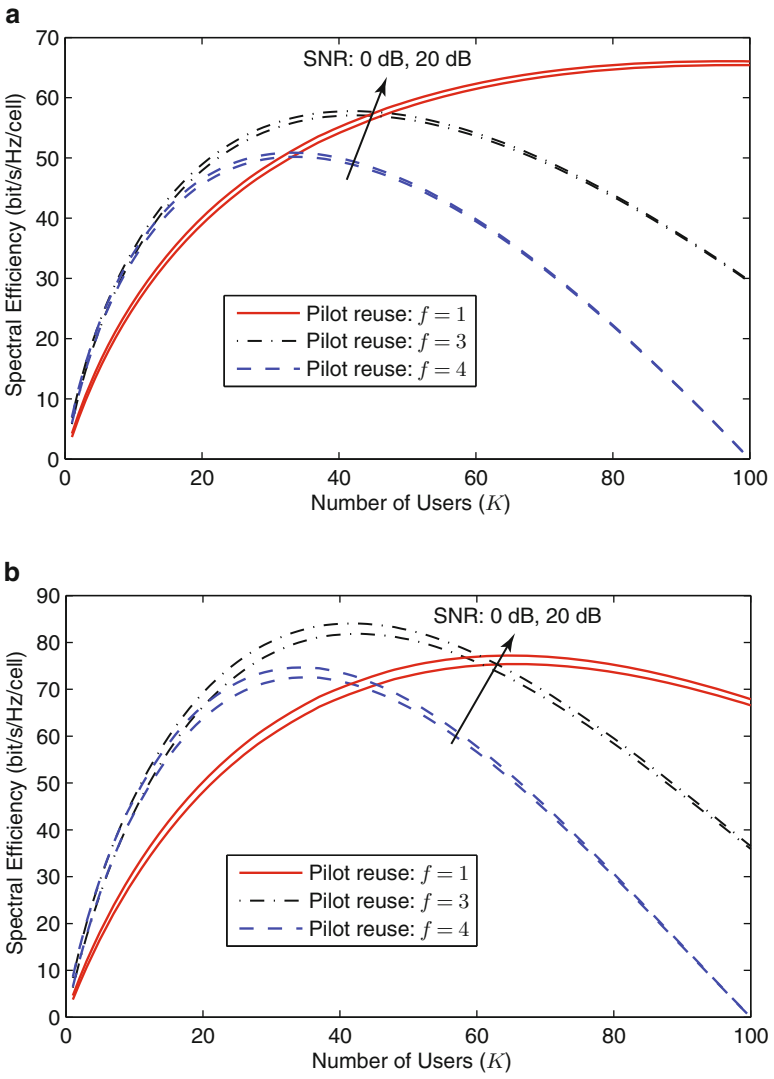


**Fig. 8** Average spectral efficiency, as a function of the number of users, with different processing schemes and pilot reuse factors. Two different SNR levels are considered: $\delta/\sigma^2_{\text{BS}} = 0$ dB or 20 dB. (**a**) Maximum ratio detection. (**b**) Zero-forcing detection

Figure 8 shows the average SE for different number of users, for both MR and ZF detection. The first observation from Fig. 8 is that the two SNR levels, $\delta/\sigma^2_{\text{BS}} = 0$ dB and 20 dB, give essentially the same performance. This shows that Massive MIMO works equally well at high and low SNRs, since the array gain makes the SE interference-limited and not noise-limited. Next, we notice that different pilot reuse factors are desirable at different user loads (i.e., number of users $K$). A pilot reuse

of $f = 3$ is desired at low load, while $f = 1$ is needed to reduce the prelog factor $(1 - fK/\tau_c)$ when $K$ is large. By selecting $f$ properly, one can always operate on the top curve in Fig. 8 and then Massive MIMO can provide a high SE over a wide range of different number of users. In fact, the technology provides a relatively stable SE for any $K > 10$. This removes the need for intricate scheduling in Massive MIMO networks, because all active users can basically be served simultaneously in every coherence interval (or at least up to $\tau_c/2$ users, to leave half of the coherence interval for data, which is a number that is typically more than a hundred [9]); the high sum SE is then shared between all the users.

Another observation is that the difference in SE between ZF and MR is relatively small in Fig. 8; ZF only gives a performance gain of between 3 and 45 %, depending on the number of users. This should be compared with the single-cell simulation in Fig. 3, where ZF provided more than twice the SEs as MR. The reason for the more modest performance gap is that also ZF suffers from interference in the multi-cell case, since the pilot contamination and many inter-cell interferers make it impossible to cancel all interference.

In summary, the pilot reuse factor is an important design parameter in Massive MIMO networks and the best choice depends on the user load. As shown in [3], it also depends on the propagation environment and the number of BS antennas.

### *4.3 How High Spectral Efficiency can Massive MIMO Deliver?*

We conclude this chapter by showcasing the SE that the Massive MIMO technology can deliver in the uplink and downlink of 5G networks—which is the same due to the uplink–downlink duality. We continue the previous simulation example from Fig. 8, but focus on ZF processing with pilot reuse $f = 3$ and a power allocation policy that gives an SNR of 0 dB to everyone. Note that these design choices are motivated by the previous simulation results.

Figure 9 shows the SE as a function of the number of BS antennas $M$. The number of active users is optimized for each $M$ to get the highest SE, and the optimal user numbers are also shown in the figure. A reasonable performance baseline is the IMT-Advanced requirements for 4G networks, provided in [22]. This document specifies spectral efficiencies in the range of 2–3 bit/s/Hz/cell, depending on the simulation scenario. In comparison, the Massive MIMO network considered in Fig. 9 achieves 52 bit/s/Hz/cell using $M = 100$ antennas, which is a 17× to 26× improvement over IMT-Advanced. With $M = 400$ antennas the Massive MIMO system achieves 114 bit/s/Hz/cell, which is an incredible 38× to 57× improvement over IMT-Advanced. These improvements are between one and two orders of magnitude!

It is important to notice that the number of active users increase alongside the SE in Fig. 9. If one divides the top curve with the bottom curve, this gives the SE per user. Interestingly, the SE per user lies in the modest range of 1–2.5 bit/s/Hz.

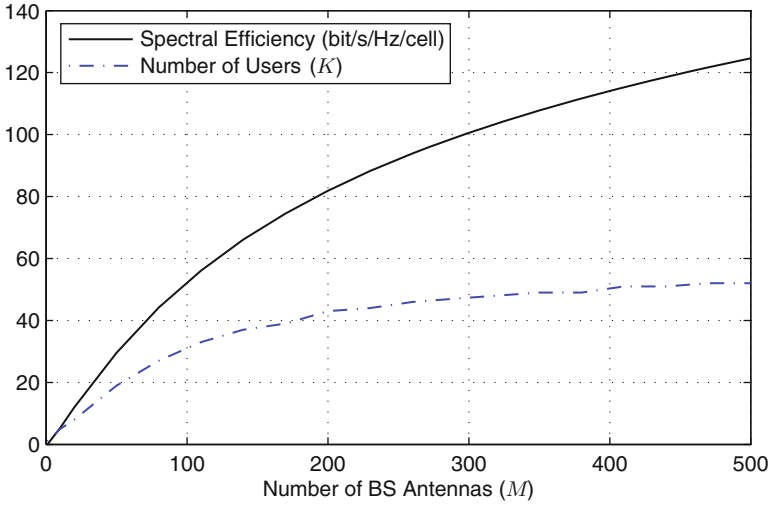


**Fig. 9** Average spectral efficiency, as a function of the number of BS antennas, with ZF processing, a pilot reuse factor $f = 3$, and an SNR of 0 dB. The number of users is optimized for each antenna number to yield the highest SE, and the corresponding number of users is also shown

Such spectral efficiencies can be achieved in practice using conventional modulation schemes, such as 16-QAM with appropriate channel coding.

In conclusion, Massive MIMO can theoretically provide tenfold or even 50-fold improvements in SE over IMT-Advanced. These huge improvements are mainly achieved by serving many users simultaneously, while the SE per user is in the conventional range. Huge gains are achieved already at $M = 100$ or fewer BS antennas. These are indeed very encouraging results indicating that the Massive MIMO technology is key to not only improve the SE, but can also be the driving force towards achieving orders of magnitude higher area throughput in 5G technologies.

**Acknowledgements** The authors of this chapter have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 641985 (5Gwireless). The authors are also supported by ELLIIT and CENIIT. We would also like to thank Daniel Verenzuela and Meysam Sadeghi for helping out with the proof-reading.

# Appendix

## *Proof of Lemma 2*

Recall that $\boldsymbol{\phi}_{j,k} \in \mathbb{C}^{\tau_p}$ is the pilot sequence used by the $k$th user in the $j$th cell, where $\boldsymbol{\Phi}_j = [\boldsymbol{\phi}_{j,1} \ldots \boldsymbol{\phi}_{j,K}]$. Since the desired channel $\mathbf{h}_{j,k}^l$ only appears as $\mathbf{h}_{j,k}^l \boldsymbol{\phi}_{j,k}^H$ in (8), a sufficient statistic for estimating this channel is given by

$$\begin{aligned} \mathbf{Y}_l^{\text{pilot}} \boldsymbol{\phi}_{j,k} &= \sum_{i=1}^{L} \mathbf{H}_i^l \mathbf{P}_i \boldsymbol{\Phi}_i^H \boldsymbol{\phi}_{j,k} + \mathbf{N}_l \boldsymbol{\phi}_{j,k} \\ &= \sum_{i \in \mathscr{P}_j} \sqrt{p_{i,k}} \tau_p \mathbf{h}_{i,k}^l + \tilde{\mathbf{n}}_{l,j,k} \end{aligned} \tag{42}$$

where $\tilde{\mathbf{n}}_{l,j,k} = \mathbf{N}_l \boldsymbol{\phi}_{j,k} = [\tilde{n}_{l,j,k,1} \ldots \tilde{n}_{l,j,k,M}]^T \sim \mathbb{C}\mathscr{N}(\mathbf{0}, \tau_p \sigma_{\text{UL}}^2 \mathbf{I}_M)$. The second equality follows from the assumed orthogonality of the pilot sequences.

Based on (42), we compute a separate LMMSE estimate of each element of $\mathbf{h}_{j,k}^l$. If $y_{l,j,k,m} \in \mathbb{C}$ denotes the $m$th row of the vector in (42), then

$$y_{l,j,k,m} = \sum_{i \in \mathscr{P}_j} \sqrt{p_{i,k}} \tau_p h_{i,k,m}^l + \tilde{n}_{l,j,k,m}. \tag{43}$$

By the definition of LMMSE estimation [24], the LMMSE estimate of $h_{j,k,m}^l$ is given by

$$\hat{h}_{j,k,m}^l = \mathbb{E}\left\{h_{j,k,m}^l\right\} + \frac{\text{Cov}\left\{h_{j,k,m}^l, y_{l,j,k,m}\right\}}{\mathbb{V}\left\{y_{l,j,k,m}\right\}} \left(y_{l,j,k,m} - \mathbb{E}\left\{y_{l,j,k,m}\right\}\right) \tag{44}$$

where we recall that $\mathbb{E}\left\{h_{j,k,m}^l\right\} = \bar{h}_{j,k,m}^l$ by definition and $\text{Cov}\{\cdot,\cdot\}$ stands for covariance. Moreover, we have

$$\mathbb{E}\left\{y_{l,j,k,m}\right\} = \sum_{i \in \mathscr{P}_j} \sqrt{p_{i,k}} \tau_p \bar{h}_{i,k,m}^l, \tag{45}$$

$$\text{Cov}\left\{h_{j,k,m}^l, y_{l,j,k,m}\right\} = \sqrt{p_{j,k}} \tau_p \beta_{j,k}^l, \tag{46}$$

$$\mathbb{V}\left\{y_{l,j,k,m}\right\} = \sum_{i \in \mathscr{P}_j} p_{i,k} \tau_p^2 \beta_{i,k}^l + \tau_p \sigma_{\text{UL}}^2. \tag{47}$$

The estimation expression in (9) is obtained by substituting (45)–(47) into (44) and writing the result in vector form. The variance of the estimate is then given by

$$\mathbb{V}\left\{\hat{h}_{j,k,m}^l\right\} = \frac{\left|\text{Cov}\left\{h_{j,k,m}^l, y_{l,j,k,m}\right\}\right|^2}{\mathbb{V}\left\{y_{l,j,k,m}\right\}} = \frac{p_{j,k} \tau_p (\beta_{j,k}^l)^2}{\sum_{i \in \mathscr{P}_j} p_{i,k} \tau_p \beta_{i,k}^l + \sigma_{\text{UL}}^2}, \tag{48}$$

while the estimation error variance in (10) is obtained as $\beta_{j,k}^l - \mathbb{V}\left\{\hat{h}_{j,k,m}^l\right\}$ since the LMMSE estimate and its error are uncorrelated [24].

## *Proof of Theorem 1*

The ergodic capacity $C_{l,k}^{\mathrm{UL}}$ with linear detection and pilot-based channel estimation is the supremum of the mutual information between the input signal $x_{l,k}$ and the output signal $\mathbf{v}_{l,k}^H \mathbf{y}_l$ in (11). The supremum is taken over the distribution of the unit-variance input signal $x_{l,k}$, thus a lower bound is obtained by assuming that $x_{l,k} \sim \mathbb{C}\mathcal{N}(0,1)$. Let $\widehat{\mathscr{H}}$ denote the channel estimates available as side-information at the receiver. We then have that

$$\begin{aligned} C_{l,k}^{\mathrm{UL}} &\geq \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) I(x_{l,k}; \mathbf{v}_{l,k}^H \mathbf{y}_l, \widehat{\mathscr{H}}) \\ &= \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \left(h(x_{l,k}) - h(x_{l,k} | \mathbf{v}_{l,k}^H \mathbf{y}_l, \widehat{\mathscr{H}})\right) \\ &= \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \left(\log_2(\pi e) - h(x_{l,k} | \mathbf{v}_{l,k}^H \mathbf{y}_l, \widehat{\mathscr{H}})\right) \end{aligned} \tag{49}$$

where $I(\cdot;\cdot)$ denotes the mutual information under the suboptimally assumed Gaussian signal distribution and $h(\cdot)$ is the differential entropy function. The first equality follows from the definition of mutual information and the second equality uses the entropy expression for complex Gaussian random variables. The factor $\gamma^{\mathrm{UL}}(1-\frac{\tau_p}{\tau_c})$ is the fraction of transmission symbols used for uplink data in each coherence interval. It remains to characterize $h(x_{l,k}|\mathbf{v}_{l,k}^H \mathbf{y}_l, \widehat{\mathscr{H}})$, which is done by finding a tractable upper bound on this term:

$$\begin{aligned} h(x_{l,k} | \mathbf{v}_{l,k}^H \mathbf{y}_l, \widehat{\mathscr{H}}) &= h(x_{l,k} - \alpha \mathbf{v}_{l,k}^H \mathbf{y}_l | \mathbf{v}_{l,k}^H \mathbf{y}_l, \widehat{\mathscr{H}}) \\ &\leq h(x_{l,k} - \alpha \mathbf{v}_{l,k}^H \mathbf{y}_l) \\ &\leq \log_2 \left(\pi e \mathbb{E}\{|x_{l,k} - \alpha \mathbf{v}_{l,k}^H \mathbf{y}_l|^2\}\right) \end{aligned} \tag{50}$$

where the equality follows from the fact that subtracting a known variable $\alpha \mathbf{v}_{l,k}^H \mathbf{y}_l$, for some deterministic scalar $\alpha$, does not change the entropy. The first inequality follows from dropping the knowledge of $\mathbf{v}_{l,k}^H \mathbf{y}_l$ and $\widehat{\mathscr{H}}$ which increases the entropy, while the second inequality follows from exploiting the fact that the highest entropy is obtained when $x_{l,k} - \alpha \mathbf{v}_{l,k}^H \mathbf{y}_l$ is a zero-mean complex Gaussian random variable with the same second-order moment as the original variable has.

The last step of the proof is to select $\alpha$ to get the tightest upper bound in (50), which corresponds to the minimization problem

$$\min_{\alpha} \ \mathbb{E}\left\{|x_{l,k} - \alpha \mathbf{v}_{l,k}^H \mathbf{y}_l|^2\right\} = \frac{1}{1 + \mathrm{SINR}_{l,k}^{\mathrm{UL}}}, \tag{51}$$

which is solved by first computing the expectation with respect to the independent signals $x_{i,t}$, for all $i$ and $t$, then finding the optimal $\alpha$ by equating the first derivative of the expression (with respect to $\alpha$) to zero, and substituting it back into the expression. From (49) to (51) we now have

$$\begin{aligned} C_{l,k}^{\mathrm{UL}} &\geq \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \left(\log_2(\pi e) - \log_2 \left(\pi e \frac{1}{1 + \mathrm{SINR}_{l,k}^{\mathrm{UL}}}\right)\right) \\ &= \gamma^{\mathrm{UL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \mathrm{SINR}_{l,k}^{\mathrm{UL}}\right), \end{aligned} \tag{52}$$

which is the result stated in the theorem.

## *Proof of Corollary 1*

Before computing the SINR expression in (23) for Rayleigh fading channels, we recall that

$$\mathbb{V}\left\{h_{i,t,m}^{l}\right\} = \beta_{i,t}^{l}, \tag{53}$$

$$\mathbb{V}\left\{\hat{h}_{i,t,m}^{l}\right\} = \frac{p_{i,t}\tau_p(\beta_{i,t}^{l})^2}{\sum_{i'\in\mathscr{P}_l} p_{i',t}\tau_p\beta_{i',t}^{l} + \sigma_{\mathrm{UL}}^2}, \tag{54}$$

$$\mathrm{MSE}_{i,t}^{l} = \beta_{i,t}^{l}\left(1 - \frac{p_{i,t}\tau_p\beta_{i,t}^{l}}{\sum_{i'\in\mathscr{P}_l} p_{i',t}\tau_p\beta_{i',t}^{l} + \sigma_{\mathrm{UL}}^2}\right), \tag{55}$$

for the channel between an arbitrary user $t$ in cell $i$ ($i = 1,\ldots,L$) and BS $l$. Note that $m$ is used as an arbitrary antenna index since the channel variance is the same for all antennas. The corollary is first proved in the case of MR detection, where $\mathbf{v}_{l,k} = \hat{\mathbf{h}}_{l,k}^{l}$, in which case $\mathrm{SINR}_{l,k}^{\mathrm{UL}}$ in Theorem 1 becomes

$$\mathrm{SINR}_{l,k}^{\mathrm{MR,UL}} = \frac{p_{l,k}\left|\mathbb{E}\left\{(\hat{\mathbf{h}}_{l,k}^{l})^H\mathbf{h}_{l,k}^{l}\right\}\right|^2}{\sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|(\hat{\mathbf{h}}_{l,k}^{l})^H\mathbf{h}_{i,t}^{l}|^2\right\} - p_{l,k}\left|\mathbb{E}\left\{(\hat{\mathbf{h}}_{l,k}^{l})^H\mathbf{h}_{l,k}^{l}\right\}\right|^2 + \sigma_{\mathrm{UL}}^2\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^2\right\}}. \tag{56}$$

It remains to compute the expectations in the numerator and denominator of (56). Since $\mathbf{h}_{l,k}^{l} = \hat{\mathbf{h}}_{l,k}^{l} + \mathbf{e}_{l,k}^{l}$, as stated in Lemma 2, the numerator is computed as

$$p_{l,k}\left|\mathbb{E}\left\{(\hat{\mathbf{h}}_{l,k}^{l})^H\mathbf{h}_{l,k}^{l}\right\}\right|^2 = M^2 p_{l,k}\left(\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\right)^2. \tag{57}$$

When computing the denominator, we decompose its first term into three parts based on the pilot reuse; the first two parts contain the cells that use the same pilot sequences as cell $l$ (i.e., all cells in $\mathscr{P}_l$) and the third part contains the remaining cells. We then observe that

$$\begin{aligned}
&\sum_{i=1}^{L}\sum_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|(\hat{\mathbf{h}}_{l,k}^{l})^H\mathbf{h}_{i,t}^{l}|^2\right\} - p_{l,k}\left|\mathbb{E}\left\{(\hat{\mathbf{h}}_{l,k}^{l})^H\mathbf{h}_{l,k}^{l}\right\}\right|^2 + \sigma_{\mathrm{UL}}^2\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^2\right\} \\
&= \sum_{i\in\mathscr{P}_l} p_{i,k}\mathbb{E}\left\{|(\hat{\mathbf{h}}_{l,k}^{l})^H\mathbf{h}_{l,k}^{l}|^2\right\} + \sum_{i\in\mathscr{P}_l}\sum_{\substack{t=1\\ t\neq k}}^{K} p_{i,t}\mathbb{E}\left\{|(\hat{\mathbf{h}}_{l,k}^{l})^H\mathbf{h}_{i,t}^{l}|^2\right\} \\
&\quad + \sum_{i\notin\mathscr{P}_l}\sum_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|(\hat{\mathbf{h}}_{l,k}^{l})^H\mathbf{h}_{i,t}^{l}|^2\right\} - p_{l,k}\left|\mathbb{E}\left\{(\hat{\mathbf{h}}_{l,k}^{l})^H\mathbf{h}_{l,k}^{l}\right\}\right|^2 + \sigma_{\mathrm{UL}}^2\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^2\right\} \\
&= M^2\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\sum_{i\in\mathscr{P}_l\setminus\{l\}} p_{i,k}\mathbb{V}\left\{\hat{h}_{i,k,m}^{l}\right\} + M\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\sum_{i\in\mathscr{P}_l}\sum_{t=1}^{K} p_{i,t}\mathbb{V}\left\{h_{i,t,m}^{l}\right\} \\
&\quad + M\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\sum_{i\notin\mathscr{P}_l}\sum_{t=1}^{K} p_{i,t}\mathbb{V}\left\{h_{i,t,m}^{l}\right\} + M\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\sigma_{\mathrm{UL}}^2.
\end{aligned} \tag{58}$$

The first term in the second expression of (58) demonstrates the effect of pilot contamination and is computed by using (21) and the independence between the MMSE estimate and its estimation error.

Besides, we handle the expectation $\mathbb{E}\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^4\}$ by virtue of Lemma 2.9 in [40], since $\hat{\mathbf{h}}_{l,k}^{l}(\hat{\mathbf{h}}_{l,k}^{l})^H$ is an $M \times M$ central complex Wishart matrix with one degree of freedom:

$$\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^4\right\} = \mathbb{E}\left\{\mathrm{tr}^2\left(\hat{\mathbf{h}}_{l,k}^{l}(\hat{\mathbf{h}}_{l,k}^{l})^H\right)\right\} = M(M+1)\left(\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}\right)^2, \tag{59}$$

where $\mathrm{tr}(\cdot)$ stands for the trace of a matrix. In contrast, the second term of the middle expression of (58) is computed by the fact that the remaining users in $\mathscr{P}_l$ use pilot sequences that are orthogonal to the pilot sequence of user $k$. The third term in (58) is computed based on the independence between the channel estimates in cell $l$ and the channels in other cells not belong to $\mathscr{P}_l$, while the last term follows from the fact that $\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^2\right\} = M\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}$.

Substituting (57) and (58) into (56), the SINR expression with MR detection becomes

$$\mathrm{SINR}_{l,k}^{\mathrm{MR,UL}} = \frac{Mp_{l,k}\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}}{M\sum\limits_{i\in\mathscr{P}_l\setminus\{l\}} p_{i,k}\mathbb{V}\left\{\hat{h}_{i,k,m}^{l}\right\} + \sum\limits_{i\in\mathscr{P}_l}\sum\limits_{t=1}^{K} p_{i,t}\mathbb{V}\left\{h_{i,t,m}^{l}\right\} + \sigma_{\mathrm{UL}}^2} \tag{60}$$

which equals the expression in the corollary by further substituting (53)–(55) into (60).

In case of ZF detection, the channel inversion structure yields the property

$$\mathbb{E}\left\{\mathbf{v}_{l,k}^H\mathbf{h}_{l,k}^{l}\right\} = 1. \tag{61}$$

Additionally, the noise term in (17) is computed as

$$\sigma_{\mathrm{UL}}^2\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\} = \sigma_{\mathrm{UL}}^2\mathbb{E}\left\{\mathrm{tr}\left[\left((\hat{\mathbf{H}}_{l}^{l})^H\hat{\mathbf{H}}_{l}^{l}\right)^{-1}\right]_{k,k}\right\} = \frac{\sigma_{\mathrm{UL}}^2}{(M-K)\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}} \tag{62}$$

by utilizing the fact that $(\hat{\mathbf{H}}_{l}^{l})^H\hat{\mathbf{H}}_{l}^{l}$ is a $K \times K$ central complex Wishart matrix with $M$ degrees of freedom and applying Lemma 2.10 in [40] to compute the trace of the inverse. Note that $[\cdot]_{k,k}$ is used here to denote the $k$th diagonal element of a matrix.

Substituting (61) and (62) into (17), we achieve the SINR expression

$$\mathrm{SINR}_{l,k}^{\mathrm{ZF,UL}} = \frac{p_{l,k}}{\sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|\mathbf{v}_{l,k}^H\mathbf{h}_{i,t}^{l}|^2\right\} - p_{l,k} + \frac{\sigma_{\mathrm{UL}}^2}{(M-K)\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}}}. \tag{63}$$

To compute the remaining expectations, we utilize the pilot reuse patterns together with the ZF properties to decompose the expectation term in (63) into three terms:

$$\begin{aligned}
&\sum_{i=1}^{L}\sum_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|\mathbf{v}_{l,k}^H\mathbf{h}_{i,t}^{l}|^2\right\} \\
&= \sum_{i\in\mathscr{P}_l} p_{i,k}\mathbb{E}\left\{|\mathbf{v}_{l,k}^H\hat{\mathbf{h}}_{i,k}^{l}|^2\right\} + \sum_{i\in\mathscr{P}_l}\sum_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|\mathbf{v}_{l,k}^H\mathbf{e}_{i,t}^{l}|^2\right\} + \sum_{i\notin\mathscr{P}_l}\sum_{t=1}^{K} p_{i,t}\mathbb{E}\left\{|\mathbf{v}_{l,k}^H\mathbf{h}_{i,t}^{l}|^2\right\} \\
&= \sum_{i\in\mathscr{P}_l}\frac{p_{i,k}^2(\beta_{i,k}^{l})^2}{p_{l,k}(\beta_{l,k}^{l})^2} + \sum_{i\in\mathscr{P}_l}\sum_{t=1}^{K}\frac{p_{i,t}\mathrm{MSE}_{i,t}^{l}}{(M-K)\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}} + \sum_{i\notin\mathscr{P}_l}\sum_{t=1}^{K}\frac{p_{i,t}\mathbb{V}\left\{h_{i,t,m}^{l}\right\}}{(M-K)\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}}.
\end{aligned} \tag{64}$$

In the last equality of (64), the first term is obtained by utilizing the relationship between user channels for cells in $\mathscr{P}_l$ as stated in (21). The second and third terms follow directly from the independence between the ZF detection vector, the estimation errors for channels in $\mathscr{P}_l$ and the complete channels for cells not in $\mathscr{P}_l$. Moreover, Lemma 2.10 in [40] is again used to compute the expectation of the inverse of the central complex Wishart matrix $(\hat{\mathbf{H}}_l^l)^H \hat{\mathbf{H}}_l^l$. Substituting (64) back into (63) and utilizing the properties in (53)–(55), the final SINR expression for ZF is obtained.

## *Proof of Theorem 2*

Substituting (28) into (29), the received signal at user $k$ in cell $l$ is

$$\begin{aligned} y_{l,k} &= \sum_{i=1}^{L} (\mathbf{h}_{l,k}^i)^H \sum_{t=1}^{K} \sqrt{\rho_{i,t}} \mathbf{w}_{i,t} s_{i,t} + n_{l,k} \\ &= \underbrace{\sqrt{\rho_{l,k}} (\mathbf{h}_{l,k}^l)^H \mathbf{w}_{l,k} s_{l,k}}_{\text{Desired signal}} + \underbrace{\sum_{\substack{t=1 \\ t \neq k}}^{K} \sqrt{\rho_{l,t}} (\mathbf{h}_{l,k}^l)^H \mathbf{w}_{l,t} s_{l,t}}_{\text{Intra-cell interference}} + \underbrace{\sum_{\substack{i=1 \\ i \neq l}}^{L} \sum_{t=1}^{K} \sqrt{\rho_{i,t}} (\mathbf{h}_{l,k}^i)^H \mathbf{w}_{i,t} s_{i,t}}_{\text{Inter-cell interference}} + \underbrace{n_{l,k}}_{\text{Noise}}. \end{aligned} \tag{65}$$

The last row of (65) shows that $s_{l,k}$ is the desired signal that we want to detect, under additive noise, intra-cell, and inter-cell interference. Similar to Theorem 1, if $C_{l,k}^{\mathrm{DL}}$ is the ergodic capacity with linear precoding, then we compute a lower bound on the mutual information between $s_{l,k}$ and $y_{l,k}$ by considering a potentially suboptimal Gaussian signal distribution, $s_{l,k} \sim \mathscr{CN}(0,1)$, and bounding the corresponding conditional mutual information $I(s_{l,k}; y_{l,k})$ as follows:

$$\begin{aligned} C_{l,k}^{\mathrm{DL}} &\geq \gamma^{\mathrm{DL}} \left(1 - \frac{\tau_p}{\tau_c}\right) I(s_{l,k}; y_{l,k}) \\ &\geq \gamma^{\mathrm{DL}} \left(1 - \frac{\tau_p}{\tau_c}\right) \left(\log_2(\pi e) - h(s_{l,k}|y_{l,k})\right) \\ &\geq \log_2\left(1 + \mathrm{SINR}_{l,k}^{\mathrm{DL}}\right) \end{aligned} \tag{66}$$

where the inequalities follow from the same procedures as in (50)–(52). The lower bound on the ergodic capacity $R_{l,k}^{\mathrm{DL}}$ in (30) is then obtained. Note that in contrast to the proof of Theorem 1, the receiver does not have any side-information with channel estimates in the downlink.

## *Proof of Lemma 3*

Let $\xi_{l,1}, \ldots, \xi_{l,K}$ be the uplink SINRs of the $K$ users in cell $l$ that are achieved by Theorem 1 for the given detection vectors and uplink power coefficients, such that the equations $\mathrm{SINR}_{l,k}^{\mathrm{UL}} = \xi_{l,k}$ hold for $l = 1, \ldots, L$ and $k = 1, \ldots, K$. From this condition we get

$$\xi_{l,k} \frac{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}}{\left|\mathbb{E}\left\{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l\right\}\right|^2} = \frac{p_{l,k}}{\sum\limits_{i=1}^{L} \sum\limits_{t=1}^{K} p_{i,t} \frac{\mathbb{E}\left\{|\mathbf{v}_{l,k}^H \mathbf{h}_{i,t}^l|^2\right\}}{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}} - p_{l,k} \frac{\left|\mathbb{E}\left\{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l\right\}\right|^2}{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}} + \sigma_{\mathrm{UL}}^2}, \tag{67}$$

by multiplying each side of the equation $\mathrm{SINR}_{l,k}^{\mathrm{UL}} = \xi_{l,k}$ with $\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\} / |\mathbb{E}\{\mathbf{v}_{l,k}^H \mathbf{h}_{l,k}^l\}|^2$.

The goal is to prove that also $\text{SINR}_{l,k}^{\text{DL}} = \xi_{l,k}$ holds if the downlink precoding vectors in (32) are used and the downlink transmit power coefficients are selected appropriately. According to the definition of the downlink precoding vectors, the equation $\text{SINR}_{l,k}^{\text{DL}} = \xi_{l,k}$ can be written as

$$\xi_{l,k} \frac{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}}{\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H \mathbf{v}_{l,k}\right\}\right|^2} = \frac{\rho_{l,k}}{\sum\limits_{i=1}^{L}\sum\limits_{t=1}^{K} \rho_{i,t} \frac{\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H \mathbf{v}_{i,t}|^2\right\}}{\mathbb{E}\left\{\|\mathbf{v}_{i,t}\|^2\right\}} - \rho_k \frac{\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H \mathbf{v}_{l,k}\right\}\right|^2}{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}} + \sigma_{\text{DL}}^2}, \tag{68}$$

by multiplying each side of the equation with $\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\} / |\mathbb{E}\{(\mathbf{h}_{l,k}^l)^H \mathbf{v}_{l,k}\}|^2$.

Let us define a diagonal matrix $\mathbf{D}$ and a matrix $\boldsymbol{\Psi}$, both of size $KL \times KL$. Let $\mathbf{D}_l \in \mathbb{C}^{K \times K}$ be the $l$th diagonal block of $\mathbf{D}$ and let $\boldsymbol{\Psi}_{l,i} \in \mathbb{C}^{K \times K}$ be the $(l,i)$th block of $\boldsymbol{\Psi}$. The elements of these blocks are defined as

$$[\mathbf{D}_l]_{k,k} = \frac{\xi_{l,k} \mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}}{\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H \mathbf{v}_{l,k}\right\}\right|^2}, \tag{69}$$

$$\left[\boldsymbol{\Psi}_{l,i}\right]_{k,t} = \begin{cases} \frac{\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H \mathbf{v}_{i,t}|^2\right\}}{\mathbb{E}\left\{\|\mathbf{v}_{i,t}\|^2\right\}} - \frac{\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^l)^H \mathbf{v}_{l,k}\right\}\right|^2}{\mathbb{E}\left\{\|\mathbf{v}_{l,k}\|^2\right\}}, & \text{for } l = i \text{ and } t = k, \\ \frac{\mathbb{E}\left\{|(\mathbf{h}_{l,k}^i)^H \mathbf{v}_{i,t}|^2\right\}}{\mathbb{E}\left\{\|\mathbf{v}_{i,t}\|^2\right\}}, & \text{otherwise.} \end{cases} \tag{70}$$

Using this notation, the $KL$ equations in (67) and (68) respectively become

$$\frac{\mathbf{p}}{\sigma_{\text{UL}}^2} = \left(\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi}^T\right)^{-1} \mathbf{D}\mathbf{1}_{KL}, \tag{71}$$

$$\frac{\boldsymbol{\rho}}{\sigma_{\text{DL}}^2} = \left(\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi}\right)^{-1} \mathbf{D}\mathbf{1}_{KL}, \tag{72}$$

where $\mathbf{p} = [\mathbf{p}_1^T \ldots \mathbf{p}_L^T]^T$ and $\mathbf{p}_i = [p_{i,1} \ldots p_{i,K}]^T$ contain the uplink transmit powers, $\boldsymbol{\rho} = [\boldsymbol{\rho}_1^T \ldots \boldsymbol{\rho}_L^T]^T$ and $\boldsymbol{\rho}_i = [\rho_{i,1} \ldots \rho_{i,K}]^T$ contain the downlink transmit powers, $\mathbf{1}_{KL}$ is a $KL \times 1$ vector with only ones, and $\mathbf{I}_{KL}$ is the $KL \times KL$ identity matrix. These equations give the uplink and downlink transmit powers that provide the SINRs $\xi_{l,1}, \ldots, \xi_{l,K}$ in cell $l$, but only if the inverses $(\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi}^T)^{-1}$ and $(\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi})^{-1}$ exist.

Since $\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi}^T$ and $\mathbf{I}_{KL} - \mathbf{D}\boldsymbol{\Psi}$ have the same eigenvalues, either both or none of the inverses exist. Recall that we have selected $\xi_{l,k}$ (for $l = 1, \ldots, L$, $k = 1, \ldots, K$) as the SINRs that were actually achieved in the uplink, thus the inverses must exist and (72) gives the downlink transmit powers that achieves the same SINRs in the downlink as in the uplink. It is also straightforward to show that

$$\frac{\mathbf{1}_{KL}^T \mathbf{p}}{\sigma_{\text{UL}}^2} = \frac{\boldsymbol{\rho}^T \mathbf{1}_{KL}}{\sigma_{\text{DL}}^2}, \tag{73}$$

which corresponds to the relationship between the total transmit power in the uplink and downlink stated in the lemma.

## *Proof of Corollary 3*

The proof follows along the same lines as the proof of Corollary 1, because the same expectations are involved, thus the variances summarized in (53)–(55) are still useful. We briefly summarize the proof of Corollary 3 as follows.

We need to compute all the expectations in (31). MR precoding gives the desired signal power

$$\rho_{l,k}\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^{l})^{H}\mathbf{w}_{l,k}\right\}\right|^{2}=\frac{\rho_{l,k}}{M\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}}\left|\mathbb{E}\left\{\|\hat{\mathbf{h}}_{l,k}^{l}\|^{2}\right\}\right|^{2}=M\rho_{l,k}\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\} \tag{74}$$

and the denominator is computed as

$$\begin{aligned}
&\sum_{i=1}^{L}\sum_{t=1}^{K}\rho_{i,t}\mathbb{E}\left\{|(\mathbf{h}_{l,k}^{i})^{H}\mathbf{w}_{i,t}|^{2}\right\}-\rho_{l,k}\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^{l})^{H}\mathbf{w}_{l,k}^{l}\right\}\right|^{2}+\sigma_{\mathrm{DL}}^{2}\\
&=\sum_{i\in\mathscr{P}_{l}}\rho_{i,k}\mathbb{E}\left\{|(\mathbf{h}_{l,k}^{i})^{H}\mathbf{w}_{i,k}|^{2}\right\}+\sum_{i\in\mathscr{P}_{l}}\sum_{\substack{t=1\\t\neq k}}^{K}\rho_{i,t}\mathbb{E}\left\{|(\mathbf{h}_{l,k}^{i})^{H}\mathbf{w}_{i,t}|^{2}\right\}\\
&\quad+\sum_{i\notin\mathscr{P}_{l}}\sum_{t=1}^{K}\rho_{i,t}\mathbb{E}\left\{|(\mathbf{h}_{l,k}^{i})^{H}\mathbf{w}_{i,t}|^{2}\right\}-\rho_{l,k}\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^{l})^{H}\mathbf{w}_{l,k}^{l}\right\}\right|^{2}+\sigma_{\mathrm{DL}}^{2}\\
&=M\sum_{i\in\mathscr{P}_{l}\setminus\{l\}}\rho_{i,k}\mathbb{V}\left\{\hat{h}_{i,k,m}^{i}\right\}+\sum_{i=1}^{L}\sum_{t=1}^{K}\rho_{i,t}\mathbb{V}\left\{h_{l,k,m}^{i}\right\}+\sigma_{\mathrm{DL}}^{2}.
\end{aligned} \tag{75}$$

Substituting (74) and (75) into (31), yields the SINR expression stated for MR in the corollary.

Next, we consider ZF precoding for which we notice that

$$\mathbb{E}\left\{\left\|\hat{\mathbf{H}}_{i}^{i}\mathbf{r}_{i,t}\right\|^{2}\right\}=\mathbb{E}\left\{\left[\left((\hat{\mathbf{H}}_{i}^{i})^{H}\hat{\mathbf{H}}_{i}^{i}\right)^{-1}\right]_{t,t}\right\}=\frac{1}{(M-K)\mathbb{V}\left\{\hat{h}_{i,t,m}^{i}\right\}} \tag{76}$$

by utilizing the fact that $(\hat{\mathbf{H}}_{i}^{i})^{H}\hat{\mathbf{H}}_{i}^{i}$ is a central complex Wishart matrix with $M$ degrees of freedom. Hence, the ZF precoding vector becomes

$$\mathbf{w}_{i,t}=\sqrt{(M-K)\mathbb{V}\left\{\hat{h}_{i,t,m}^{i}\right\}}\hat{\mathbf{H}}_{i}^{i}\mathbf{r}_{i,t}. \tag{77}$$

Using this precoding vector, we compute the numerator and denominator of (31) as follows:

$$\rho_{l,k}\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^{l})^{H}\mathbf{w}_{l,k}^{l}\right\}\right|^{2}=(M-K)\rho_{l,k}\mathbb{V}\left\{\hat{h}_{l,k,m}^{l}\right\}, \tag{78}$$

$$\begin{aligned}
&\sum_{i=1}^{L}\sum_{t=1}^{K}\rho_{i,t}\mathbb{E}\left\{|(\mathbf{h}_{l,k}^{i})^{H}\mathbf{w}_{i,t}|^{2}\right\}-\rho_{l,k}\left|\mathbb{E}\left\{(\mathbf{h}_{l,k}^{l})^{H}\mathbf{w}_{l,k}\right\}\right|^{2}+\sigma_{\mathrm{DL}}^{2}\\
&=(M-K)\sum_{i\in\mathscr{P}_{l}\setminus\{l\}}\rho_{i,k}\mathbb{V}\left\{\hat{h}_{l,k,m}^{i}\right\}+\sum_{i\in\mathscr{P}_{l}}\sum_{t=1}^{K}\rho_{i,t}\mathrm{MSE}_{l,k}^{i}+\sum_{i\notin\mathscr{P}_{l}}\sum_{t=1}^{K}\rho_{i,t}\mathbb{V}\{h_{i,t,m}\}+\sigma_{\mathrm{DL}}^{2}.
\end{aligned} \tag{79}$$

Substituting (78) and (79) into (31), yields the SINR expression stated for ZF in the corollary.

## References


1. S. Anderson, M. Millnert, M. Viberg, B. Wahlberg, An adaptive array for mobile communication systems. IEEE Trans. Veh. Technol. **40**(1), 230–236 (1991)
2. I. Atzeni, J. Arnau, M. Debbah, Fractional pilot reuse in massive MIMO systems, in *Proceedings of the IEEE ICC*, 2015
3. E. Björnson, E. Jorswieck, Optimal resource allocation in coordinated multi-cell systems. Found. Trends Commun. Inf. Theor. **9**(2–3), 113–381 (2013)
4. E. Björnson, B. Ottersten, A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance. IEEE Trans. Signal Process. **58**(3), 1807–1820 (2010)
5. E. Björnson, G. Zheng, M. Bengtsson, B. Ottersten, Robust monotonic optimization framework for multicell MISO systems. IEEE Trans. Signal Process. **60**(5), 2508–2523 (2012)
6. E. Björnson, M. Kountouris, M. Bengtsson, B. Ottersten, Receive combining vs. multi-stream multiplexing in downlink systems with multi-antenna users. IEEE Trans. Signal Process. **61**(13), 3431–3446 (2013)
7. E. Björnson, J. Hoydis, M. Kountouris, M. Debbah, Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits. IEEE Trans. Inf. Theory **60**(11), 7112–7139 (2014)
8. E. Björnson, E. Jorswieck, M. Debbah, B. Ottersten, Multi-objective signal processing optimization: the way to balance conflicting metrics in 5G systems. IEEE Signal Process. Mag. **31**(6), 14–23 (2014)
9. E. Björnson, E. Larsson, M. Debbah, Massive MIMO for maximal spectral efficiency: how many users and pilots should be allocated? IEEE Trans. Wireless Commun. **15**(2), 1293–1308 (2016)
10. H. Boche, M. Schubert, A general duality theory for uplink and downlink beamforming, in *Proceedings of the IEEE VTC-Fall* (2002), pp. 87–91
11. G. Caire, S. Shamai, On the achievable throughput of a multiantenna Gaussian broadcast channel. IEEE Trans. Inf. Theory **49**(7), 1691–1706 (2003)
12. H.V. Cheng, E. Björnson, E.G. Larsson, Uplink pilot and data power control for single cell massive MIMO systems with MRC, in *Proceedings of the IEEE ISWCS*, 2015
13. H.V. Cheng, E. Björnson, E.G. Larsson, Optimal pilot and payload power control in single-cell massive MIMO systems. IEEE Trans. Signal Process (submitted)
14. R. Couillet, M. Debbah, *Random Matrix Methods for Wireless Communications* (Cambridge University Press, Cambridge, 2011)
15. D. Cox, Cochannel interference considerations in frequency reuse small-coverage-area radio systems. IEEE Trans. Commun. **30**(1), 135–142 (1982)
16. X. Gao, O. Edfors, F. Rusek, F. Tufvesson, Linear pre-coding performance in measured very-large MIMO channels, in *Proceedings of the IEEE VTC Fall*, 2011
17. D. Gesbert, M. Kountouris, R. Heath, C.B. Chae, T. Sälzer, Shifting the MIMO paradigm. IEEE Signal Process. Mag. **24**(5), 36–46 (2007)
18. A. Goldsmith, S. Jafar, N. Jindal, S. Vishwanath, Capacity limits of MIMO channels. IEEE J. Sel. Areas Commun. **21**(5), 684–702 (2003)
19. J. Hoydis, C. Hoek, T. Wild, S. ten Brink, Channel measurements for large antenna arrays, in *International Symposium on Wireless Communications Systems (ISWCS)*, 2012
20. J. Hoydis, S. ten Brink, M. Debbah, Massive MIMO in the UL/DL of cellular networks: How many antennas do we need? IEEE J. Sel. Areas Commun. **31**(2), 160–171 (2013)
21. H. Huh, G. Caire, H. Papadopoulos, S. Ramprashad, Achieving "massive MIMO" spectral efficiency with a not-so-large number of antennas. IEEE Trans. Wireless Commun. **11**(9), 3226–3239 (2012)
22. ITU: Requirements related to technical performance for IMT-advanced radio interface(s). Tech. rep., ITU-R M.2134 (2008)

23. J. Jose, A. Ashikhmin, T.L. Marzetta, S. Vishwanath, Pilot contamination and precoding in multi-cell TDD systems. IEEE Trans. Commun. **10**(8), 2640–2651 (2011)
24. S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory* (Prentice Hall, Englewood Cliffs, NJ, 1993)
25. E.G. Larsson, F. Tufvesson, O. Edfors, T.L. Marzetta, Massive MIMO for next generation wireless systems. IEEE Commun. Mag. **52**(2), 186–195 (2014)
26. X. Li, E. Björnson, E.G. Larsson, S. Zhou, J. Wang, Massive MIMO with multi-cell MMSE processing: exploiting all pilots for interference suppression. IEEE Trans. Wireless Commun. (2015, submitted). Available: http://arxiv.org/abs/1505.03682
27. Y.F. Liu, Y.H. Dai, Z.Q. Luo, Coordinated beamforming for MISO interference channel: complexity analysis and efficient algorithms. IEEE Trans. Signal Process. **59**(3), 1142–1157 (2011)
28. T.L. Marzetta, Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans. Wireless Commun. **9**(11), 3590–3600 (2010)
29. T.L. Marzetta, A. Ashikhmin, MIMO system having a plurality of service antennas for data transmission and reception and method thereof, US Patent 8,594,215, 2011
30. H. Ngo, E. Larsson, T. Marzetta, Energy and spectral efficiency of very large multiuser MIMO systems. IEEE Trans. Commun. **61**(4), 1436–1449 (2013)
31. H.Q. Ngo, E.G. Larsson, T.L. Marzetta, Massive MU-MIMO downlink TDD systems with linear precoding and downlink pilots. IEEE Trans. Commun. **61**(4), 1436–1449 (2013)
32. H. Ngo, E. Larsson, T. Marzetta, Aspects of favorable propagation in massive MIMO, in *Proceedings of the EUSIPCO*, 2014
33. H.Q. Ngo, A.E. Ashikhmin, H. Yang, E.G. Larsson, T.L. Marzetta, Cell-free massive MIMO: uniformly great service for everyone, in *Proceedings of the IEEE SPAWC*, 2015
34. A. Paulraj, C. Papadias, Space-time processing for wireless communications. IEEE Signal Process. Mag. **14**(6), 49–83 (1997)
35. L. Qian, Y. Zhang, J. Huang, MAPEL: achieving global optimality for a non-convex wireless power control problem. IEEE Trans. Wireless Commun. **8**(3), 1553–1563 (2009)
36. R. Roy, B. Ottersten, Spatial division multiple access wireless communication systems. US Patent 5,515,378, 1991
37. N. Shariati, E. Björnson, M. Bengtsson, M. Debbah, Low-complexity polynomial channel estimation in large-scale MIMO with arbitrary statistics. IEEE J. Sel. Top. Signal Process. **8**(5), 815–830 (2014)
38. S.C. Swales, M.A. Beach, D.J. Edwards, J.P. McGeehan, The performance enhancement of multibeam adaptive base-station antennas for cellular land mobile radio systems. IEEE Trans. Veh. Technol. **39**(1), 56–67 (1990)
39. E. Telatar, Capacity of multi-antenna Gaussian channels. Eur. Trans. Telecom. **10**(6), 585–595 (1999)
40. A.M. Tulino, S. Verdú, Random matrix theory and wireless communications. Found. Trends Commun. Inf. Theor. **1**(1), 1–182 (2004)
41. J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I.C. Wong, V. Öwall, O. Edfors, F. Tufvesson, A flexible 100-antenna testbed for massive MIMO, in *Proceedings of the IEEE Globecom Workshop - Massive MIMO: From Theory to Practice*, 2014
42. P. Viswanath, D. Tse, Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality. IEEE Trans. Inf. Theor. **49**(8), 1912–1921 (2003)
43. H. Weingarten, Y. Steinberg, S. Shamai, The capacity region of the Gaussian multiple-input multiple-output broadcast channel. IEEE Trans. Inf. Theory **52**(9), 3936–3964 (2006)
44. J. Winters, Optimum combining for indoor radio systems with multiple users. IEEE Trans. Commun. **35**(11), 1222–1230 (1987)
45. H. Yang, T.L. Marzetta, A macro cellular wireless network with uniformly high user throughputs, in *Proceedings of the IEEE VTC-Fall*, 2014
46. W. Yu, Uplink-downlink duality via minimax duality. IEEE Trans. Inf. Theory **52**(2), 361–374 (2006)
47. P. Zetterberg, B. Ottersten, The spectrum efficiency of a base station antenna array system for spatially selective transmission. IEEE Trans. Veh. Technol. **44**(3), 651–660 (1995)

# Millimeter-Wave Mobile Communications

**Yi Wang and Zhenyu Shi**

**Abstract** Mobile communications in millimeter wave (mmWave) bands have recently attained a wide range of research due to the available ultra-broad spectrum bands. In this chapter, we introduce the key technologies of mmWave communications based on pioneering researches. Channel measurement and modeling as a fundamental issue is presented in Sect. 2. Beam-tracking technique based on large-scale antenna array is studied in Sect. 3. Network architecture, particularly considering unified access and backhaul, is presented in Sect. 4. Current prototypes are introduced in Sect. 5. Finally we summarize the chapter in Sect. 6.

## 1 Background

The development of wireless cellular networks has been growing rapidly during the past decades. There are still exploding demands for high quality data communications in applications such as smart phones, tablets and video streaming. The capacity of cellular networks in 2020 is envisioned to be as much as 1000 times compared to current fourth-generation (4G) technology. Meanwhile, the cellular network is expected to be capable of connecting 50 billion devices for wireless services.

Recently, mmWave communication has been investigated because of its ultra broad spectrum band [1, 2]. Typical mmWave frequencies range from 6 GHz to 100 GHz including 6 GHz, 15 GHz, 28 GHz [3, 4], 38 GHz [3], 60 GHz [5], and E-band (71–76 GHz, 81–86 GHz) [1, 4]. From channel measurements, the propagation loss for mmWave transmission is quite large. Its transmission range is thus limited. Therefore, mmWave transmission is more suitable for small cells for data rate and dense user scenarios. Owing to these features, mmWave transmission is tailored for a hotspot scenario and has become a promising candidate for the fifth generation (5G) [7–9, 16, 18, 22].

Y. Wang (✉) • Z. Shi
Huawei Technologies Co., Ltd., Jinsu Road 200, Shanghai, P. R. China
e-mail: yi.wang@huawei.com; zhenyu.shi@huawei.com



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_5

## 2 mmWave Channel Measurements and Modeling

Channel propagation property is a fundamental topic of mmWave communications [17]. The channel model will affect the spectrum allocation in the International Telecomm Union (ITU) and the World Radio Conference (WRC) as well as system design and performance evaluation in 3rd Generation Partnership Project (3GPP) standards.

### *2.1 Ultra-Wide Band Channel Sounder*

An ultra wide band channel measurement sounder for transmitter and receiver has been developed by Huawei Chengdu branch, see Fig. 1. The front-end can be changed in order to match different carrier frequencies. The transmitter has a signal generator for signal generation, and the receiver uses a signal analyzer for measurement data acquisition. A transmitter (TX) and receiver (RX) synchronized spectrum sweeping solution is adopted. In this way, 2 GHz or even larger bandwidth can be measured. The signal generator at TX sites and the signal analyzer at RX sites are controlled remotely by a personal computer through network.

**Fig. 1** Channel sounder at E-band, transmitter (*left*) and receiver (*right*)

Frequency bands from 71 to 73 GHz were measured. Vertically polarized horn antennas with 25 dBi gain and 10° half power beam width (HPBW) are used for both outdoor and indoor measurements. At the receiver, the channel information from four different neighboring angles of arrival can be caught by a four-channel RF-front together with four horn antennas.

## 2.2 Channel Measurement

In this section we present current measurements in E-band. The typical scenarios of mmWave communications are indoor hotspots and outdoor hotspots, these two scenarios have been measured in Huawei Chengdu branch.

The outdoor scenario is selected in dense office buildings, see Fig. 2. Both transmitter and receiver are set at the height of 2–4 m. In non line-of-sight (NLOS) channel, there are building and trees for blocking but there are reflected signals from the opposite building where glass window and concrete wall provide reflections. Occasionally there are a few people going across the road which causes body penetration loss. A dining room is selected as the indoor hotspot scenario, see Fig. 3. The dining room is 70 m long and 25 m wide. Food lines are located at both ends. Most walls are made of glass and the rest are made of concrete.

For the line-of-sight (LOS) TX-RX location, the TX and RX antennas were aligned by means of optical. Then the TX antenna fixed pointing at the aligned angle. The RX antenna adjusts elevation angle of the aligned line to −10°, 0° and 10°, respectively, with the antenna scanning around the entire 360° azimuth plane in 10° (i.e. antenna half power beam width) step.

For the NLOS scenario, there are two steps to complete the measurement procedure. First, searching the strongest path in each TX-RX location. In this step, both the TX and the RX antennas were automatically rotated to find the strongest

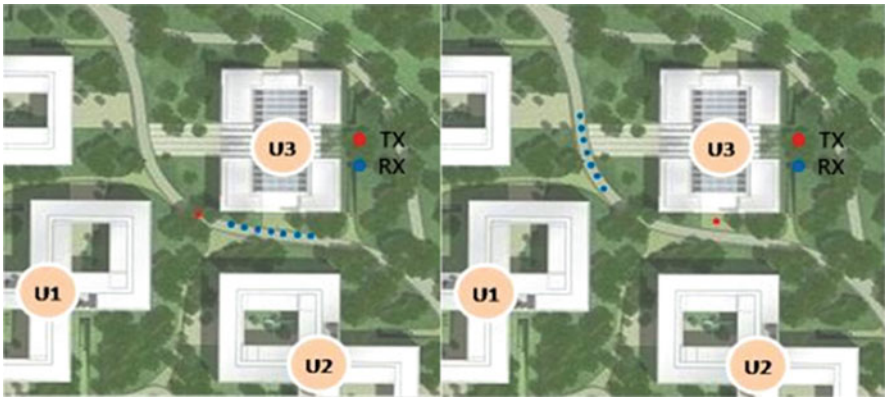


**Fig. 2** UMi outdoor scenario, LOS (*left*) and NLOS (*right*)


**Fig. 3** Indoor hotspot scenario, LOS (*left*) and NLOS (*right*)

received power. Second, Setting the pointing angle of the TX and RX antenna for maximum received power as the 0° azimuth and elevation angles for the TX and the RX antenna, respectively. The RX antenna adjusts elevation angle of the aligned line to −10°, 0° and 10°, three statuses with the antenna scanning around the entire 360° azimuth plane in 10° step.

The measurement results are shown in Fig. 4. We take close-in reference model with reference distance 1 m for path loss since all the parameters have physical meanings [1]. It is reasonable because the measurements cannot cover all distances and sites. Table 1 lists the parameters of close-in reference model based on the measurements. The NLOS path loss exponents (PLEs) of 72 GHz is 4.08 for indoor and 3.67 for outdoor which are comparable to that of 2 GHz in 3GPP model. However, a more accurate path loss channel model and more measurement data are required before making any conclusion.

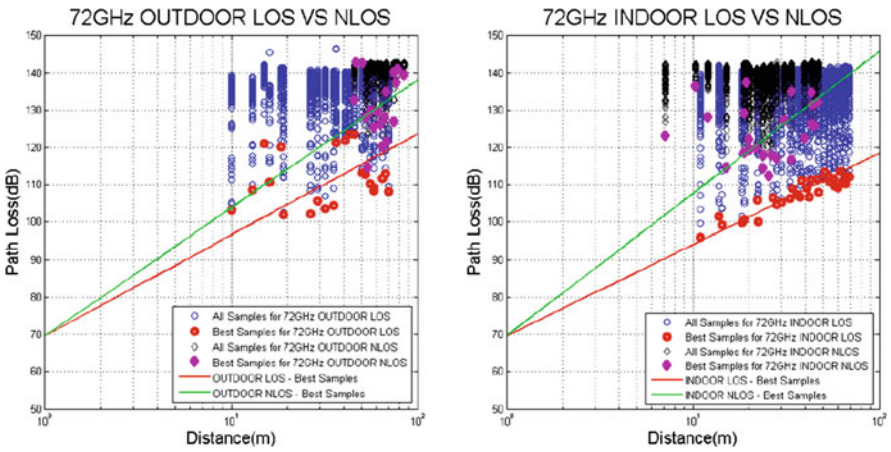


**Fig. 4** E-band path loss in UMi outdoor and indoor

**Table 1** The path loss parameters in 72 GHz, close-in reference model with reference distance 1 m is applied

| Scenario | LOS/NLOS | PLE($\beta$) | $\alpha$ | $\sigma$ |
|---|---|---|---|---|
| Indoor | LOS | 2.58 | 69.47 | 2.38 |
| | NLOS | 4.08 | 69.47 | 10.63 |
| Outdoor | LOS | 2.86 | 69.47 | 9.76 |
| | NLOS | 3.67 | 69.47 | 8.34 |

## *2.3 Channel Modeling*

There are a number of organizations and researchers working on channel modeling in mmWave bands (ref. the survey in [1, 6]). Mobile and Wireless Communications Enablers for the Twenty-twenty Information Society (METIS) project in Europe proposed a map-based channel model which is essentially a simplified ray-tracing method to model channels and fully depends on the defined map. Institute for Electrical and Electronic Engineers (IEEE) 802.11ay standard is making efforts to extend the Saleh-Valenzuela (S-V) model in 60 GHz to larger distance up to 100 m for both outdoor and indoor scenarios. In June 2015, 3GPP Radio Access Network (RAN) agreed to initiate a study item to study channel model for above-6 GHz frequency bands. Meanwhile, International Telecommunication Union Radiocommunication Section (ITU-R) Working Party 5D (WP5D) announced to standardize channel model in above-6 GHz, which is expected to reach a "standard" channel model in mid-2017. A fact is that the channel models in ITU and 3GPP are spatial channel models (SCM) for sub-6 GHz bands, which mainly come from WINNER project researches. 3GPP further extended SCM to three-dimension (3D) cases. Both 3G/4G cellular systems are based on the common SCM model to do system-level evaluation. It is promising that the channel model for 5G systems will be an extension of the SCM model.

A new feature for mmWave communications is that high-gain beams are required in base-station (BS) and potentially in user terminal (UE) in order to compensate the large propagation loss of mmWave signals. Large-scale antenna array is widely used to form high-gain beams. For example, a size of 66 mm by 66 mm may accommodate 1024 antenna elements at E-band, and can form HPBW as narrow as 3°. Thus mmWave transmission link will be sensitive to angle changes in departure and arrival. It is different from the system in sub-6 GHz bands where wide beams are used and the angle variation can be ignored in a drop. Therefore, it is necessary to introduce variant angles for mmWave channel models.

In the framework of 3GPP SCM, WINNER project studied the variant angles and variant delay spreads in a drop. An accurate formula for variant angles is derived in LOS channel, and random angles are approximated for NLOS channel in the local coordination systems (LCS). Here, an independent study on variant angles are investigated and the contributions include: (1) derivation of variant angles in the global coordinate system (GCS); (2) approximation of variant angles with linear method; (3) extension of variant angles to 3D SCM model. Please find more detailed derivations and ray-tracing based simulations in [10, 11].

The idea of SCM model with variant angles (VA-SCM) is to update the angles at each time when the receiver is moving within the duration of one drop. A drop duration denoted by $T_{max}$ is around 1000 transmission time intervals (TTIs) in 3GPP SCM where one TTI corresponds to 1 s. Following 3GPP SCM framework, large-scale parameters (e.g. path loss, delay spread, number of clusters, angular spread etc.) are fixed in a drop. The channel impulse response consists of multiple clusters, and each cluster has different time delays and average receive power which are randomly generated in terms of a given probability distribution function. Each cluster is the superposition of a number of rays. For NLOS cluster, the corresponding scatters are randomly generated for each drop, and the scatters for every ray in a cluster are distributed in a small range depending on angular spread in configuration.

Variant angles are introduced for each ray including azimuth angle of departure and arrival (AoD, AoA) and zenith angle of departure and arrival (ZoD, ZoA). Receiver location at each time can be accurately calculated based on the moving speed of UE and the moving direction. Accordingly the angles can be updated with transmitter and receiver information in the global coordination system (GCS). However, an accurate method to calculate variant angles has high computational complexity, particularly in link-level simulation where channel impulse responses are calculated for each sample or each data symbol. Linear approximation is an efficient way to reduce complexity with acceptable errors. A linear model for variant angles of the $p$th ray of the $q$th cluster is given by

$$\begin{aligned}
\theta_{p,q,ZoA}(t) &= \theta_{p,q,ZoA}(t_0) + S_{p,q,ZoA} \cdot (t - t_0) , \quad t \in [t_0, t_0 + T_{max}] \\
\theta_{p,q,ZoD}(t) &= \theta_{p,q,ZoD}(t_0) + S_{p,q,ZoD} \cdot (t - t_0) , \quad t \in [t_0, t_0 + T_{max}] \\
\phi_{p,q,AoA}(t) &= \phi_{p,q,AoA}(t_0) + S_{p,q,AoA} \cdot (t - t_0) , \quad t \in [t_0, t_0 + T_{max}] \\
\phi_{p,q,AoD}(t) &= \phi_{p,q,AoD}(t_0) + S_{p,q,AoD} \cdot (t - t_0) , \quad t \in [t_0, t_0 + T_{max}]
\end{aligned} \tag{1}$$

where $S_{p,q,ZoA}$ and $S_{p,q,ZoD}$ are the slopes of variant angles in vertical direction and $S_{p,q,AoA}$ and $S_{p,q,AoD}$ are the slopes of variant angles in horizontal direction. Notice the four slopes are fixed in a drop period $T_{max}$ although they can be extended to time-varying version but at the cost of higher computational complexity.

For easy of description, $p$ and $q$ is omitted in the rest of the section. Assuming receiver is moving at velocity $v$ with angle $\phi_v$. $h_{BS}$ and $h_{UE}$ are the height of BS and UE. For LOS cluster, the expression of AoD and ZoD slopes are given by

$$\begin{aligned} S_{ZoD} &= -S_{ZoA} = \frac{v\cos(\phi_v - \phi_{AoD}(t_0))}{(h_{BS} - h_{UE})/\cos(\theta_{ZoD}(t_0))}, \\ S_{AoD} &= S_{AoA} = -\frac{v\sin(\phi_v - \phi_{AoD}(t_0))}{(h_{BS} - h_{UE})\tan(\theta_{ZoD}(t_0))}. \end{aligned} \tag{2}$$

For NLOS cluster with one reflection ray, the model can be simplified by introducing a virtual UE which is the mirror image of UE based on the reflection surface, as is shown in Fig. 5. The simplified slopes in NLOS channel are given by

$$\begin{aligned} S_{ZoD} &= -S_{ZoA} = -\frac{v\cos(\phi_v + \phi_{AoD}(t_0) - \phi_{RS})}{(h_{BS} - h_{UE})/\cos(\theta_{ZoD}(t_0))}, \\ S_{AoD} &= S_{AoA} = -\frac{v\sin(\phi_v + \phi_{AoD}(t_0) - \phi_{RS})}{(h_{BS} - h_{UE})\tan(\theta_{ZoD}(t_0))}, \end{aligned} \tag{3}$$

where $\phi_{RS}$ is the angle of the reflection surface and it can be deduced from the initial $\phi_{AoD}$ and $\phi_{AoA}$.

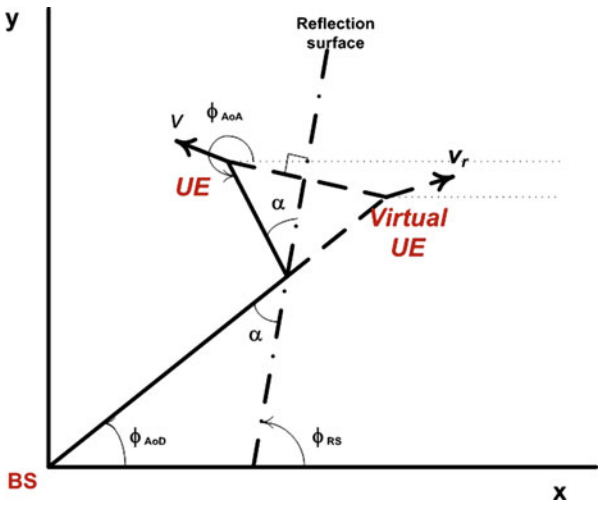


**Fig. 5** Global coordination system of an NLOS ray

For NLOS clusters with multiple reflections, it is hard to obtain the accurate location of reflectors. A solution is to take multiple reflections as one-order reflection and randomly set the reflection surface. The channel coefficients of each cluster can be generated following the steps of SCM by considering time variant angles.

## 3 Beamforming and Beam Tracking in mmWave

The air interface of mmWave communications is featured with antenna-array based beamforming and tracking. Both mmWave base station (mB) and user equipment (UE) use antenna array to compensate the large path loss of mmWave propagation. Such scheme with high-gain narrow beams brings challenges for air interface design. In this section, the concept of uniform air interface and multi-mode beamforming are presented.

### *3.1 Uniform Air Interface*

In all spectrum access system of 5G, the mmWave communication is not an independent air interface but subscribes to several constraints from lower frequency air interface and unified access and backhaul (UAB) architecture. Hence the concept of uniform air interface is expected to be realized with parameter configuration and to flexibly meet different requirements.

Firstly, there might be multiple mmWave bands allocated to 5G, and air interface should be uniform among different carrier frequencies, such as frame and physical numerologies. It is recommended that mmWave could reuse the 10 ms frame structure and 1 ms subframe. Slot can be redesigned to be 125 $\mu$s for all mmWave bands. The continuous spectrum bands may have similar channel propagation properties, for example, 6–20 GHz, 20–50 GHz, and 50–90 GHz. Different physical numerologies can be used for different frequency band groups.

Secondly, UAB architecture requests a uniform air interface design between radio access and backhaul. It is better that backhaul works like a special UE of mB where backhaul and UEs share the same radio resource. The difference is that backhaul will use special link configuration, e.g. beams, coding and modulation, to meet its quality of service (QoS) requirements. This will be further discussed in multi-mode beamforming.

Thirdly, mmWave communication allows efficient waveform design to meet different QoS requirements in different scenarios, such as power efficiency, latency etc. Both orthogonal frequency division multiplexing (OFDM) and single carrier OFDM (SC-OFDM) waveforms used in long term evolution (LTE) can be used for mmWave communications. Other candidate waveforms might be filtered OFDM (F-OFDM) and sparse code multiple access (SCMA) currently proposed for 5G

low frequency systems. A good combination of them is to use software defined adaptive air interface to select different waveforms for different scenario usages. However, a common constraint is that the waveform is robust to phase noise which may dominate other distortions in front-end. Peak to average power (PAPR) is also an important issue to waveform selection. It is expected to deploy efficient PAPR reduction algorithms to improve power amplifier efficiency particularly when using multiple carrier waveforms.

## *3.2 Multi-Mode Beamforming*

Multi-mode beamforming (MM-BF) scheme is proposed in this section. It allocates antenna elements, intermedium and radio frequency (IRF) channels and baseband units to the beams of backhaul links and radio access adaptively. Antenna arrays are divided into sub-arrays, and the beamforming processing is carried out in not only digital baseband but also IRF part. User data streams and backhaul streams are first put into adaptive multiple-input multiple-output (MIMO )mode selection (AMMS), and AMMS will do MIMO processing of each stream based on the current channel environment, where MIMO mode can be spatial multiplexing (SM), space time coding (STC), space time beamforming, etc. After AMMS, the data streams are precoded in digital domain based on beam requirements and downlink channel estimation results, then the signal streams passes through digital-to-analogue converter, infrared (IR) and radio frequency (RF) processing. Before the signals are fed to the antenna elements, the phase of signals should be shifted for each antenna element based beams' requirements and channel estimation results. Such structure can flexibly realize multiple antennas techniques including MIMO, beamforming, diversity and their hybrid forms.

In a mobile scenario with multiple users accessing, the challenge is to design algorithms to align the narrow beams between mB and UEs. Overhead cost, complexity and tracking ability are the key merits of figure to evaluate performance. Two beam phases are proposed to finish the beam alignment: beam training and beam tracking. Beam training performs a rough beam alignment where both quasi-omni-directional beams and wide beams can be used for training. Since exhaustive beam search might bring a high cost in design pilots, there are potential methods to shorten beam training period and overhead. Hierarchical beam training method [12] is an efficient way which firstly uses sector-level beam to do training and then uses narrow beams for searching. Frame design for beam training, for example, centralized training can also shorten training time.

Beam tracking performs channel information update during the time when there is no beam training. The updated channel information is typically AoA and AoD for transmitter and receiver to do beamforming. There are two methods to solve the problem. One is to use reference signal and old channel information to predict the AoD and AoA for next data transmission. The authors in [13] propose a new channel tracking technique based on sequentially updating the beamforming at transmitter

and receiver. Numerical results demonstrate the superior channel tracking ability of the proposed technique over various baselines in the literature. The second method is to estimate the AoA and AoD based on reference signal. The challenge is that the reference signal cost is increasing with the antenna size in BS and UE which is prohibitive in case of antenna size larger than 16. An efficient method to overcome the problem is to use compression sensing (CS) technique which well exploit the sparse property of mmWave channel, and can significantly reduce the overhead. The authors in [14] propose a joint channel estimation and beamforming method where CS is used to estimate AoA and AoD jointly. Results show that the overhead can be saved up to 75 % compared to traditional non-CS estimation method, e.g., least square algorithm, under practical scenarios. Furthermore, the proposed method in [14] has only 2–3 dB loss compared to the method with perfect channel information.

## 4 Unified Access and Backhaul in mmWave

In the design of 5G cellular networks, the mmWave band is expected to be served as a complementary to the low frequency band. The basic structure is a hybrid network where the low frequency and high frequency communication coexist. The conventional LTE is responsible for seamless coverage due to its longer range. In the meanwhile, the mmWave communications are served as multiple hotspots scattered inside the range of the network. Based on these rules, a prototype of unified access and backhaul network has been introduced. The UAB network utilizes a C/U split configuration. The control-plane (C-plane) is managed by macro base stations (MBs) through low frequency bands while the user plane (U-plane) is processed via mmWave base stations (mBs) through high frequency bands. A large number of UEs associated with mBs will bring significant challenges to the backhaul link from the mBs to the MBs. The cost of fiber link to the small cells will be prohibitively expensive. Hence, the wireless backhaul can be applied to reduce the cost. The conventional LTE adopt out-of-band backhaul since the frequency band in radio access is already quite limited. In mmWave networks, an in-band backhaul is suitable thanks to the abundant available spectrum of high frequency bands. In this section, three methods are proposed to determine the resource allocation in the in-band backhaul scheme.

### *4.1 Network Topology*

The UAB topology is illustrated in Fig. 6. As the figure shows, there are mainly three kinds of entities in this network: MBs, mBs and UEs. MBs are responsible for transmitting control information operating in 2 GHz low frequency bands. In Fig. 6, each MB will cover three macro cells. Thus, the whole network is covered

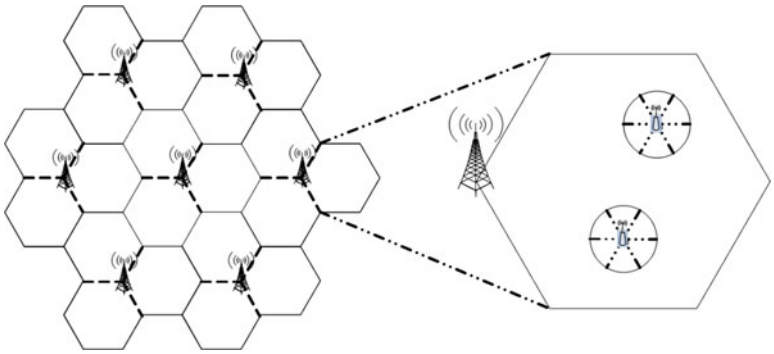

**Fig. 6** UAB network topology with seven MBs and twenty-one macro cells

by the macro cells. Inside each macro cell, mBs are deployed. The mBs will cover a relatively smaller range, functioning as small cells. Each mB can support six pico cells with 60° each. Inside each pico cell, there will be a large number of UEs connecting to the mBs. These UEs communicate with their associated mBs via high frequency bands. For those UEs that are unable to connect with mBs, they can choose to communicate with the MBs via low frequency bands.

In mmWave UAB networks, both the radio access and the backhaul share a continuous bandwidth in mmWave frequency band. In order to mitigate interference between radio access and backhaul, it is necessary to separate them either in time or in frequency resource blocks. All mBs share a same resource partition ratio so as not to interfere with each other. The ratios can also be dynamically adapted over time.

## *4.2 Dynamic Resource Allocation*

The resource can be divided either in time or in frequency. Frequency division is taken as an example here. The extension to time division is straightforward. Assuming that the overall bandwidth is $B$, where backhaul (BH) is allocated bandwidth $B^{BH}$ and radio access (RA) is allocated bandwidth $B^{RA}$. Since the overall bandwidth is a fixed value, $B = B^{BH} + B^{RA}$ is always fulfilled.

Consider the $n$th mB, the backhaul throughput can be written as a function of $B^{BH}$ by

$$T_n^{BH} = R_n^{BH} B^{BH} \tag{4}$$

and for radio access the function is

$$T_n^{RA} = R_n^{RA} B^{RA} = R_n^{RA} (B - B^{BH}) \,, \tag{5}$$

where $R_n^{BH}$ and $R_n^{RA}$ are the corresponding data rates in bit/s/Hz. $R_n^{BH}$ is determined by the link quality between the mB and MB. It can be treated as the backhaul bandwidth efficiency. $R_n^{RA}$ is determined by UEs inside the mB's range. Hence, this value is computed by summing up all the active UEs' bandwidth efficiency at a certain time. Both $R_n^{BH}$ and $R_n^{RA}$ can be pre-calculated by measuring the throughputs during a certain time period and then dividing the current bandwidth.

For any mB, its throughput $T_n$ should be limited by the minimum value of the two throughput values:

$$T_n = \min(T_n^{BH}, T_n^{RA}) \ . \tag{6}$$

By taking (4) and (5) into (6), and after some derivations, $T_n$ can be written as a function of $B^{BH}$, provided by

$$T_n(B^{BH}) = \begin{cases} R_n^{BH} B^{BH} \ \text{if} B^{BH} < B_{n,max}^{BH} \\ \frac{R_n^{BH} R_n^{RA}}{R_n^{BH} + R_n^{RA}} B \ \text{if} B^{BH} = B_{n,max}^{BH} \\ R_n^{RA}(B - B^{BH}) \ \text{if} B^{BH} > B_{n,max}^{BH} \end{cases} . \tag{7}$$

This function is a triangular function. $T_n(B^{BH})$ is first linearly increasing with a slope of $R_n^{BH}$. In this range, the backhaul throughput is lower than the radio access throughput. The overall throughput is "BH-limited". The value reaches its maximum when

$$B_{n,max}^{BH} = \frac{R_n^{RA}}{R_n^{BH} + R_n^{RA}} B \ . \tag{8}$$

and the peak value is

$$T_{n,max} = \frac{R_n^{BH} R_n^{RA}}{R_n^{BH} + R_n^{RA}} B. \tag{9}$$

Afterwards, it is reducing linearly with a slope of $R_n^{RA}$. At this range, the radio access throughput is lower than the backhaul throughput, so the overall throughput is "RA-limited".

For multiple mB, each mB will have different peak values and corresponding backhaul bandwidth. Under the assumption that the backhaul bandwidth is the same for every mB. In order to satisfy different requirements, three dynamic resource allocation methods are introduced and listed as follows:

- Max-Min: Maximization of the minimum throughput
- Max-Sum: Maximization of the sum throughput
- Quasi-PF: Maximization of the satisfactory factor

### 4.2.1 Max-Min

According to the definition, the minimum throughput of all mBs can be expressed by

$$T_{min}(B^{BH}) = \min(T_1, T_2, \cdots, T_N) \ . \tag{10}$$

The goal is to find a suitable backhaul bandwidth, where $T_{min}(B^{BH})$ is maximum:

$$B_m^{BH} = \arg\max\{T_{min}(B^{BH})\} \ . \tag{11}$$

In fact, the minimum throughput is determined by the minimum $R_{min}^{RA}$ and the minimum $R_{min}^{BH}$ of all mBs, with

$$R_{min}^{BH} = \min(R_1^{BH}, R_2^{BH}, \ldots, R_N^{BH}) \ ,$$

and

$$R_{min}^{RA} = \min(R_1^{RA}, R_2^{RA}, \ldots, R_N^{RA}) \ .$$

Thus, the minimum throughput is again a triangular function which is of the form

$$T_{min}(B^{BH}) = \begin{cases} R_{min}^{BH} B^{BH} \ \text{if} B^{BH} < B_m^{BH} \\ \frac{R_{min}^{BH} R_{min}^{RA}}{R_{min}^{BH} + R_{min}^{RA}} B \ \text{if} B^{BH} = B_m^{BH} \\ R_{min}^{RA}(B - B^{BH}) \ \text{if} B^{BH} > B_m^{BH} \end{cases} \ . \tag{12}$$

To achieve the maximum value for $T_{min}(B^{BH})$, the chosen backhaul bandwidth for all mBs should be

$$B_m^{BH} = \frac{R_{min}^{RA}}{R_{min}^{BH} + R_{min}^{RA}} B \ . \tag{13}$$

and the radio access bandwidth is

$$B_m^{RA} = B - B_m^{BH} \ . \tag{14}$$

### 4.2.2 Max-Sum

The purpose of this algorithm is to maximize the summation of all mBs' throughputs $\sum T_n$. For different mBs, the triangular functions are different in the ascending slope, the descending slope and the peak point. When the backhaul bandwidth starts from zero, all mBs are BH limited. With all the throughputs increasing, the sum throughput is increasing at the maximum speed. When the backhaul bandwidth reaches the first peak of a certain mB, this mB's throughput begins to decrease

**Algorithm 1** Max-Sum Algorithm

1: Computing the chosen BH bandwidth where $T_n$ is maximum for each mB, reordering them from minimum to maximum, with the resulting sequence $[B'^{BH}_1, B'^{BH}_2, \cdots, B'^{BH}_i, \cdots B'^{BH}_N]$;
2: Based on the above mentioned sequence, listing the corresponding BH and RA rates sequence as $[R'^{BH}_1, R'^{BH}_2, \cdots, R'^{BH}_i, \cdots R'^{BH}_N]$ and $[R'^{RA}_1, R'^{RA}_2, \cdots, R'^{RA}_i, \cdots R'^{RA}_N]$;
3: Writing a general form of $\sum T_n = \sum_{i=l+1}^{N} R'^{BH}_i B^{BH} + \sum_{i=1}^{l} R'^{RA}_i (B - B^{BH})$ , $B^{BH} \in [B'^{BH}_l, B'^{BH}_{l+1}]$;
4: Finding $[B'^{BH}_l, B'^{BH}_{l+1}]$ where $\frac{\partial(\sum T_n)}{\partial B^{BH}}$ is minimum positive;
5: **return** $B^{BH}_m = B'^{BH}_{l+1}$ as the uniform backhaul bandwidth for all mBs;

while the other mBs' throughputs continue increasing, but the ascending rate of the sum throughput is lower than before. The ascending rate continues decreasing after each mB's peak reaches until the sum throughput begins to decrease. At that point, the sum throughput is maximum. Afterwards, the descending rate speeds up until all mBs are RA limited. The function of $\sum T_n$ is a piecewise function with multiple sections.

Basically, the main goal of this method is to find the peak point of the sum throughput given the BH and RA rates of all mBs:

$$B^{BH}_m = \arg\max\{\sum T_n(B^{BH})\} . \tag{15}$$

The detailed derivations of this procedure can be found in [15], where the major steps are summarized in Algorithm 1.

### 4.2.3 Quasi-PF

The previous two algorithms only focus on the instantaneous bandwidth allocation where the history of the allocation is not considered. Both Max-Min and Max-Sum allocate resources based on the current radio access and backhaul rates. However, it may cause unfairness among mBs. For instance, if one mB has better channel conditions, the Max-Sum resource allocation will always provide it with the maximum data rates while other mBs' will always transmit at lower rates. In the long run, the mBs with high throughput will become extremely higher, and mBs with low throughputs will become extremely lower and UEs of those low throughput mBs will not be satisfied. To satisfy as many UEs as possible, the current resource allocation should also take the accumulated historical throughputs into account. Hence, the idea of the well-known proportional fairness (PF) algorithm [19] can be borrowed herein.

To evaluate among mBs, a satisfactory factor is introduced for each mB and each time index. The mB's satisfactory factor at the $t$th TTI $\eta_n(t)$ is defined as

$$\eta_n(t) = \frac{T_{n,max}(t)}{\overline{T}_{n,pre}(t)} , \tag{16}$$

where $T_{n,max}(t)$ is the maximum achievable throughput of the $n$th mB at current time index. This value is determined by the current $R_n^{RA}$ and $R_n^{BH}$ using (9). $\overline{T}_{n,pre}(t)$ is the average throughput consumed by this mB during the previous $t-1$ TTIs. The larger the nominator, the higher maximum throughput can achieve currently. The smaller the denominator, the less average throughput has been transmitted previously. Therefore, to compromise a balance between the achievable maximum throughput and the satisfactory of UEs, the mB with the largest satisfactory factor should be chosen:

$$n = \arg\max(\eta_n(t)) \ . \tag{17}$$

In case the mB with the largest $\eta_n(t)$ has been found, the chosen overall system's BH bandwidth is the peak point of this mB's triangular function.

To evaluate the proposed algorithms, simulations are carried out in the UAB networks. The network topology follows Fig. 6, operating in E-band. Radio access and backhaul share a total bandwidth of 5 GHz. To compare the algorithms in various kinds of scenarios, three different UE densities are considered, which are

- Case 1: 20 % mBs have high UE loads, 80 % mBs have low UE loads;
- Case 2: 50 % mBs have high UE loads, 50 % mBs have low UE loads;
- Case 3: all mBs have the same UE loads.

Figure 7 compares the system's performance of three cases. The mean throughput is computed by summing up the throughputs of all mB and then dividing the mB

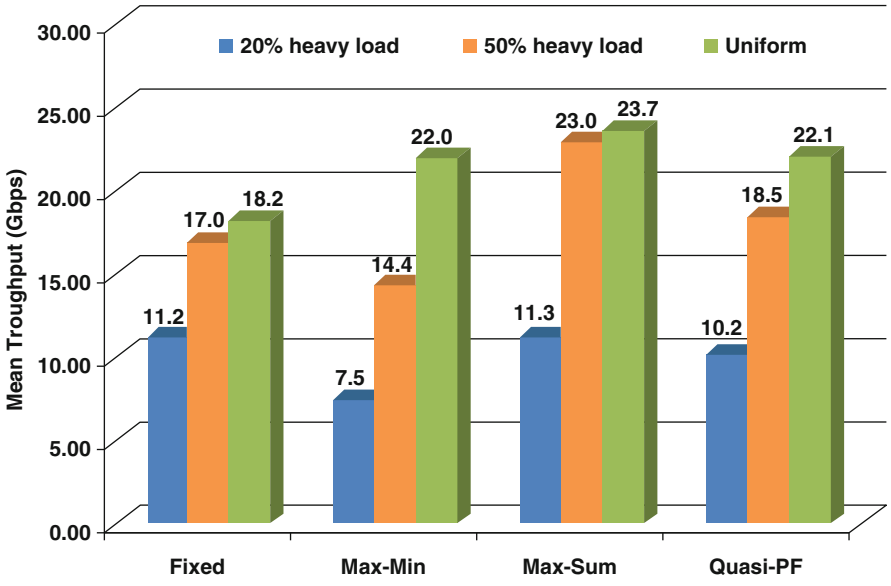


**Fig. 7** Mean throughputs for different cases

number and the TTI number. It can be found that the Max-Min algorithm provides the lowest mean throughput for all three cases. This is reasonable as this way of allocation only satisfies the worst case. The Max-Sum algorithm in all three cases delivers the best performance. The throughput of Quasi-PF algorithm is in between the fixed case and the Max-Sum algorithm in case 2 and case 3. There is 35.3 and 30.2 % performance improvement compared to the fixed one in case 2 and case 3, respectively. Hence, it can be concluded that in terms of the whole system throughput, the Max-Sum algorithm provides the best solution. The Quasi-PF algorithm aims at providing a better trade off between overall capacity and mB fairness.

## 5 Prototype Verification

Several companies have announced their prototype verification for mmWave communications. Samsung realized a peak data rate of 7.5 Gbit/s at 28 GHz frequency; DoCoMo realized a peak data rate of 10 Gbit/s at 11 GHz frequency [20]. Ericsson realized a peak data rate of 5 Gbit/s at 15 GHz. Huawei and Nokia demonstrated a peak data rate of 115 Gbit/s [4] and 10 Gbit/s [21] in 72 GHz bands, respectively. For the waveform of prototypes, Nokia applies single carrier waveform (called NCP-SC) which is essentially discrete fourier transform (DFT)-precoded OFDM waveform but cyclic prefix is null. Samsung, DoCoMo, Ericsson and Huawei are using OFDM waveforms in prototype.

In this section, a 115 Gbit/s prototype working in E-band (72 GHz) is presented and a mmWave platform supporting multiple bands. The 115 Gbit/s prototype, see Fig. 8, was designed to demonstrate the peak data rate by using a total available bandwidth of 10 GHz in E-band. Point-to-point transmission is demonstrated in indoor LOS scenario. MIMO is configured with 2 polarized antennas in mB and 2 polarized antennas in UE. With LOS MIMO, two data streams are transmitted together. An OFDM waveform and low-density parity-check (LDPC) channel coding scheme is used for link transmission. Considering 20 % overhead, two streams, each with 64 Quadrature Amplitude Modulation (QAM), have been successfully transmitted which deliver a data rate of 115 Gbit/s.

To reduce the severe phase noise at E-band, a low complexity two stage estimation and compensation scheme was demonstrated for OFDM-MIMO systems. This is a joint channel estimation and phase noise compensation method. Results show that the proposed algorithm is robust to phase noise even when the detailed noise model is uncertain.


**Fig. 8** 115 Gbps prototype E-band, transmitter and receiver (*left*), real-time receiver performance (*right*)

## 6 Chapter Summary

The key technologies of mmWave mobile communications are investigated in this chapter. Channel measurements show that mmWave signals suffer from much larger propagation loss and are suitable for small cell coverage. A hybrid network is presented where mmWave is used for capacity enhancement in hotspots and low frequency network is applied for seamless coverage. Uniform air interface is a consequence to simplify the design between mmWave bands and low frequency bands. Unified access and backhaul technique not only reduces the cost of backhaul, but also can meet the requirement of 1000 times capacity enhancement than LTE system.

## References

1. T. Rappaport, R.W. Heath, R.C. Daniels, J.N. Murdock, *Millimeter Wave Wireless Communications* (Prentice Hall, Englewood Cliffs, NJ, 2014)
2. Y. Wang, J. Li, L. Huang, Y. Jing, A. Georgakopoulos, P. Demestichas, Spectrum broadening to higher-frequency bands to support high data rates. IEEE Veh. Technol. Mag. **9**, 39–46 (2014)
3. A.I. Sulyman, A.T. Nassar, M. Samimi, G.R. MacCartney, T.S. Rappaport, A. Alsanie, Radio propagation path loss models for 5G cellular networks in the 28 GHz and 38 GHz millimeter-wave bands. IEEE Commun. Mag. **52**, 78–86 (2014)

4. Y. Wang, L. Huang, Z. Shi, H. Huang, D. Steer, J. Li, G. Wang, W. Tong, An introduction to 5G mmWave communications, in *Proceedings of IEEE Globecom 2015 Workshop*, San Diego, 2015
5. Channel models for 60 GHz WLAN systems. IEEE Std. 802.11-09/00 334r8 (2010)
6. P. Almers, E. Bonek, A. Burr et al., Survey of channel and radio propagation models for wireless MIMO systems. EURASIP J. Wirel. Commun. Netw. **2007**, 19pp (2007). Article ID 19070
7. W. Roh, J.Y. Seol, J.H. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results. IEEE Commun. Mag. **52**, 106–113 (2014)
8. Preliminary draft new recommendation ITU-R M.[IMT.VISION], ITU-R WP 5D std., Document 5D/TEMP/548(Rev.3) (2015)
9. Getting ready for 5G. 3GPP std. TSG RAN 67, Shanghai, 9–12 March 2015
10. Y. Wang, L. Huang, Z. Shi, K. Liu, X. Zou, A millimeter wave channel model with variant angles under 3GPP SCM framework, in *Proceedings of IEEE PIMRC 2015 Workshop*, Hong Kong, 2015
11. Y. Wang, Z. Shi, M. Du, W. Tong, A millimeter wave spatial channel model with variant angles and variant path loss, in *Proceedings of IEEE WCNC 2016*, Doha, 2016
12. Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. IEEE std 802.11ad. (2012)
13. J. He, T. Kim, H. Ghauch, K. Liu, G. Wang, Millimeter wave MIMO channel tracking systems, in *Proceedings of IEEE Globecom '14 Workshop*, Austin (2014), pp. 414–419
14. H. Huang, K. Liu, R. Wen, Y. Wang, G. Wang, Joint channel estimation and beamforming for millimeter wave cellular system, in *Proceedings of IEEE Globecom '15*, San Diego, 2015
15. Z. Shi, Y. Wang, L. Huang, T. Wang, Dynamic resource allocation in mmWave unified access and backhaul network, in *Proceedings of IEEE PIMRC '15 Workshop*, Hongkong, 2015
16. H. Li, L. Huang, Y. Wang, Scheduling schemes for interference suppression in millimeter-wave cellular network, in *Proceedings of IEEE PIMRC '15 Workshop*, Hongkong, 2015
17. G. MacCartney, J. Zhang, S. Nie, T.S. Rappaport, Path loss models for 5G millimeter wave propagation channels in urban microcells, in *Proceedings of IEEE Globecom '13* (2013), pp. 3948–3953
18. J. Zhang, A. Beletchi, Y. Yi, H. Zhuang, Capacity performance of millimeter wave heterogeneous networks at 28 GHz/73 GHz, in *Proceedings of IEEE Globecom '14 Workshop*, Austin, 2014
19. P. Viswanath, D. Tse, R. Laroia, Opportunistic beamforming using dumb antennas. IEEE Trans. Inf. Theory **48**(6), 1277–1294 (2002)
20. S. Suyama, J. Shen, Y. Oda, H. Suzuki, K. Fukawa, DoCoMo and Tokyo Institute of Technology achieve world's first 10 Gbps packet transmission in outdoor experiment, in *Proceedings of IEEE PIMRC '13*, 2013
21. Nokia, Nokia networks showcases 5G speed of 10Gbps with NI at the Brooklyn 5G summit, 2015. http://networks.nokia.com/news-events/press-room/press-releases/nokia-networks-showcases-5g-speed-of-10gbps-with-ni-at-the-brooklyn-5g-summit
22. 3GPP TSG RAN E-UTRA, Further advancements for E-UTRA physical layer aspects (Release 9). 3GPP std. TR 36.814 v9.0.0 (2010)

# Non-Orthogonal Multiple Access (NOMA) for Future Radio Access

**Razieh Razavi, Mehrdad Dianati, and Muhammad Ali Imran**

**Abstract** Radio access technologies for cellular mobile communications are typically characterized by multiple access schemes, e.g., frequency division multiple access (FDMA), time division multiple access (TDMA), code division multiple access (CDMA), and OFDMA. In the 4th generation (4G) mobile communication systems such as Long-Term Evolution (LTE) (Au et al., Uplink contention based SCMA for 5G radio access. Globecom Workshops (GC Wkshps), 2014. doi:10.1109/GLOCOMW.2014.7063547) and LTE-Advanced (Baracca et al., IEEE Trans. Commun., 2011. doi:10.1109/TCOMM.2011.121410.090252; Barry et al., Digital Communication, Kluwer, Dordrecht, 2004), standardized by the 3rd Generation Partnership Project (3GPP), orthogonal multiple access based on OFDMA or single carrier (SC)-FDMA is adopted. Orthogonal multiple access was a reasonable choice for achieving good system-level throughput performance with simple single-user detection. However, considering the trend in 5G, achieving significant gains in capacity and system throughput performance is a high priority requirement in view of the recent exponential increase in the volume of mobile traffic. In addition the proposed system should be able to support enhanced delay-sensitive high-volume services such as video streaming and cloud computing. Another high-level target of 5G is reduced cost, higher energy efficiency and robustness against emergencies.

## 1 Introduction

Radio access technologies for cellular mobile communications are typically characterized by multiple access schemes, e.g., frequency division multiple access (FDMA), time division multiple access (TDMA), code division multiple

R. Razavi (✉)
Group Research and Development, Vodafone Group Services Limited, Emerald House, Newbury Business Park, London Road, Newbury, Berks., RG14 2PZ, UK
e-mail: Razieh.Razavi@vodafone.com

M. Dianati • M.A. Imran
Institute for Communication Systems (ICS), Home of the 5G Innovation Centre, University of Surrey, Surrey GU2 7XH, UK
e-mail: m.dianati@surrey.ac.uk; m.Imran@surrey.ac.uk



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_6

access (CDMA), and OFDMA. In the 4th generation (4G) mobile communication systems such as Long-Term Evolution (LTE) and LTE-Advanced, orthogonal multiple access based on OFDMA or single carrier (SC)-FDMA is adopted. Orthogonal multiple access was a reasonable choice for achieving good system-level throughput performance with simple single-user detection. However, considering the trend in 5G, achieving significant gains in capacity and system throughput performance is a high priority requirement in view of the recent exponential increase in the volume of mobile traffic. In addition the proposed system should be able to support enhanced delay-sensitive high-volume services such as video streaming and cloud computing. Another high-level target of 5G is reduced cost, higher energy efficiency and robustness against emergencies. 5G has to provide increased capacity per unit network cost and be energy efficient and resilient to natural disasters. Figure 1 shows how NOMA fits in the system architecture of 5G.

To accommodate such demands, novel radio access technologies are needed. In this chapter, we introduce non-orthogonal multiple access techniques and evaluate the low density spreading based system which is a strong candidate for the next generation of mobile networks. we investigate the achievable rate region of the low density signature (LDS) schemes for flat and frequency selective channels. The main objective is to define an effective channel matrix taking into account sparse allocation of chips to the users that forces certain elements of the effective channel matrix be zero. In other words, the effective channel matrix becomes sparse since each user is allowed to use a certain set of chips. We take into account practical considerations such as finite size modulation and convolutional coding to improve the accuracy of the proposed analysis.

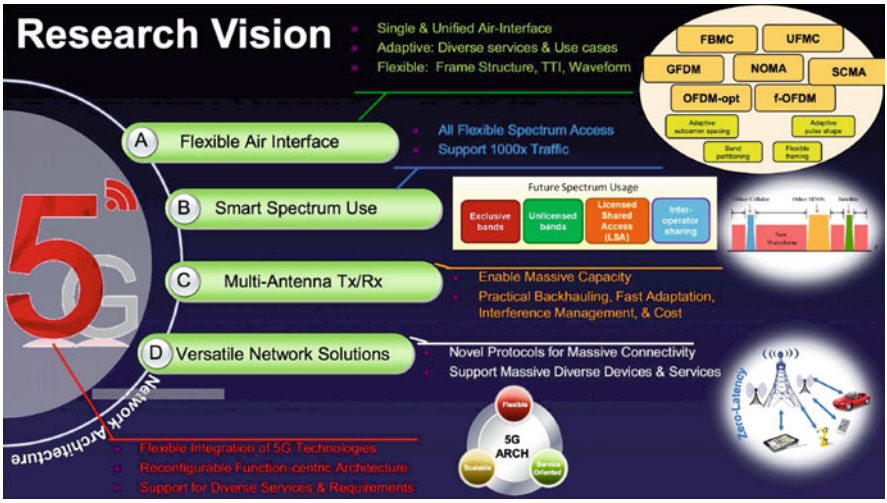


**Fig. 1** 5G system architecture [33]

## 2 State of the Art Non-Orthogonal Multiple Access Techniques

Multiple access techniques enable users to share the available resources, in other words, using these techniques multiple users can effectively share radio resources. These techniques provide each user: access to system resources with conflict avoidance; fair resource allocation or different quality of service (QoS) satisfaction; as well as system capacity enhancement. Although there are many different multiple access techniques in the literature, this section focuses only on recently proposed state of the art techniques and those which are adopted in 3G and LTE-Advanced.

### *2.1 Code Division Multiple Access (CDMA)*

CDMA is a single carrier technique that enables multiple users to share the same spectrum using their unique code sequence (spreading code). The code sequence, applied to an information-bearing, narrow-band signal, spreads the signal across the operating frequency band. Every additional user increases interference level (background noise) in the channel because all users use the entire bandwidth. Therefore, the interference level depends on the number of users and their transmission power. Power control is mandatory in order to overcome the near-far effect and to increase system capacity. Code sequences have to be chosen very carefully because well designed codes can provide orthogonality between users. Good cross-correlation features can decrease multiple access interference (MAI) whereas good auto-correlation features can decrease inter symbol interference (ISI).

### *2.2 Large Area Synchronised CDMA*

Considerable research efforts have been invested in designing spreading sequences for CDMA systems, which benefit from zero correlation values, when the relative delay-induced code offset is in the so-called zero correlation zone (ZCZ) or interference free window (IFW) of the spreading code [12]. Combining the large area codes introduced in [22] with loosely synchronised codes [35] resulted in large area synchronised (LAS) CDMA spreading sequences. The resultant LAS codes exhibit an IFW, where both the off-peak aperiodic auto-correlation values and the aperiodic cross-correlation values become zero. Therefore this system has zero ISI and zero MAI, as long as the time-offset of the codes is within the IFW. One disadvantage of LAS codes is that the number of codes having an IFW is limited. Furthermore, the auto-correlation and cross-correlation function of LAS codes have a higher value outside the IFW compared to that of traditional random codes. Therefor, LAS-CDMA faces more serious MAI and multipath interference compared to conventional CDMA when it operates in an asynchronous manner [43].

### 2.3 Multi-Carrier CDMA (MC-CDMA)

In MC-CDMA systems a data symbol is sent on multiple subcarriers by using a spreading code, which is different for the multiple access users [26]. Multiple user signals overlap in the time and frequency domain but they can be detected at the receiver by using the knowledge of spreading codes. Therefore, MC-CDMA can be considered as a combination of OFDM and CDMA schemes, combining the benefits of both schemes. Thus, this system is considered to be a suitable approach to cope with challenging service demands due to its ability of exploiting both time and frequency resources [8, 13, 28, 45]. In the uplink channel using MC-CDMA, non-orthogonality of received effective signatures causes multi-user interference (MUI). Multiuser detection can be used to moderate the detrimental effects of MUI similar to CDMA systems.

One of the advantages of MC-CDMA systems is that since this system is a combination of OFDM and CDMA, it can decrease the symbol rate in each subcarrier so that a longer symbol duration makes it easier to quasi-synchronise the transmission [18]. However, implementation of optimum multiuser detector (MUD) is computationally complex for MC-CDMA systems.

In order to reduce the complexity of MUD for MC-CDMA systems, Group Orthogonal MC-CDMA (GO-MC-CDMA) [44] has been proposed. The basic idea of this scheme is to partition the available subcarriers into groups and distribute users among the groups. The users that are assigned subcarriers of the same group are separated using spreading codes. The users in each group are immune to interference from other groups. The main advantage of GO-MC-CDMA is that each group behaves as an independent MC-CDMA system with a smaller number of users, making the optimum MUD computationally feasible within each group. It has been shown that by dividing users into subgroups, GO-MC-CDMA is able to achieve a performance very close to the single-user bound [44], with a reasonable computational complexity.

### 2.4 Interleave Division Multiple Access (IDMA)

An interleaver-based multiple access scheme has been studied in [23, 29] for high spectral efficiency, improved performance and low receiver complexity. This scheme relies on interleaving as the only means to distinguish the signals from different users, and hence it has been called interleave-division multiple-access. IDMA inherits many properties of CDMA based systems, in particular, diversity against fading and mitigation of the worst-case other-cell user interference problem. Furthermore, it allows a very simple chip-by-chip iterative MUD strategy [23, 29]. In this system, the normalized MUD cost (per user) is independent of the number of users. Furthermore, since user separation can be done by using interleaving, the whole bandwidth can be devoted for coding and, thus, paving the way for the

information theoretic view regarding the coding-spreading saga, IDMA, in essence, may achieve high spectral efficiency. However, in order to achieve optimality, IDMA requires a large number of observations. The observation may come from a large number of spreading chips or a large frame size. It has been shown in [24] that there exists a trade-off between performance and the length of the interleaver.

## 2.5 Low Density Signature CDMA (LDS-CDMA)

In order to reduce the complexity of MUD for CDMA systems, a novel spreading signature structure was introduced. By introducing the new structure, termed as LDS structure, higher loading became possible for those systems [20]. Such CDMA systems with LDS structure are called LDS-CDMA. In this system, the interference is managed intelligently at the transmitters side. The main idea of this new technique is to switch off a large number of chips for each symbol so the signature matrix will be a sparse matrix. In other words, each user will spread its data over a small number of chips (or effective processing gain $\nu_k \ll N$). Consequently at each received chip, a user will have only maximum of $d_c - 1$ interferers, where $d_c \ll K$ is the number of users that are allowed to interfere to each other at each chip. The complexity order of a maximum a posteriori (MAP) receiver turns out to be $\mathcal{O}(|\mathbb{X}|^{d_c})$ instead of $\mathcal{O}(|\mathbb{X}|^{K})$ of an optimum receiver for a conventional CDMA system.

Further reduction in complexity of multiuser detection for LDS-CDMA system is possible using grouped-based technique introduced in [42]. This is done by arranging the interfering users of a chip into two groups and approximating the log-likelihood ratio (LLR) of the interference coming from other group as a single symmetric Gaussian distributed variable.

The LDS-CDMA has been reported to approach near single-user performance even under a load of 200 % [20]. However one drawback of LDS-CDMA is that in wideband channels, its performance degrades due to multipath fading. This happens because the ISI introduced by multipath fading causes different chips interfere to each other. Therefore, more number of users will interfere in one chip; this increases the density of LDS's graph by adding edges to the graph representation of LDS. To overcome this issue LDS-OFDM was proposed [21].

To improve the performance in high order modulations, sparse code multiple access (SCMA) is introduced in [1, 27, 37]. This scheme is a multi-dimensional codebook-based non-orthogonal spreading technique. In SCMA, the procedure of bit to QAM symbol mapping and spreading are combined together and incoming bits are directly mapped to multi-dimensional codewords of SCMA codebook sets. Each layer has its dedicated codebook. Shaping gain of a multi-dimensional constellation is one of the main sources of the performance improvement in comparison to the simple repetition of QAM symbols in LDS.

### 2.6 *Space-Division Multiple-Access (SDMA)*

Antennas arrays, sometimes called smart multiple access antennas, have been introduced in mobile radio systems to exploit the spatial domain. In these systems, signals are received and sent only from and into a limited angular range, following the directional nature of multipath. This improves coverage or link quality in noise-limited situations and enhances capacity in interference-limited situations.

SDMA systems enable two or more users, associated to the same base station, to use the same time and frequency and code resources on the grounds of their physical location or spatial separation [36]. Therefore, we can discriminate amongst the users by exploiting the fact that different users impinge different spatial signatures on the receive antenna array. This is generally done with directional antennas. SDMA techniques can be used for increasing cell coverage areas. Also, they are useful in reducing the interference for improved service quality [25]. We showed a summary of the discussed NOMA schemes in Table 1.

## 3 Performance Evaluation of LDS Based Systems

In this section, we will evaluate LDS based systems since it is a strong candidate for 5th generation (5G) of mobile networks. LDS was initially proposed for CDMA systems, known as LDS-CDMA in [20]. The main idea is to spread information of each user on a limited number of chips, in contrast to conventional CDMA which spreads information of each user on all chips. Consequently, for a system with $K$ users, only a limited number of users are allowed to interfere at each chip.

**Table 1** Summary of different NOMA schemes

| | |
|---|---|
| CDMA | Several transmitters can send information simultaneously over a single communication channel. CDMA employs spread-spectrum technology and a special coding scheme. |
| LAS-CDMA | Applies a new set of spreading codes that exhibit interference free window to separate users in the wireless channel. |
| MC-CDMA | Spreads each user symbol in the frequency domain. That is, each user symbol is carried over multiple orthogonal subcarriers. |
| IDMA | Interleavers are used as the only means for user separation. IDMA inherits many advantages from CDMA, such as diversity against fading and mitigation of the worst-case other-cell user interference problem. |
| LDS-CDMA | Uses spreading sequences of low density for spreading the data symbols in time domain. This technique benefits from a less complex MUD compared to conventional CDMA with optimum MUD. |
| SCMA | Enables two or more users, associated to the same base station, to use the same time and frequency and code resources on the grounds of their physical location or spatial separation. |

The effective number of interfering users at chip $n$ will be reduced to $\eta_n \ll K$. Thus, the complexity order of the MUD for LDS-CDMA receiver will reduce to $\mathcal{O}\left(|\mathbb{X}|^{\eta_n}\right)$ compared to $\mathcal{O}\left(|\mathbb{X}|^{K}\right)$ of CDMA.

Recently, most of the modern communication systems adopt orthogonal frequency division multiplexing (OFDM) transmission techniques due to its resilience against inter symbol interference (ISI) for wideband channels. To this end, LDS for MC-CDMA (LDS-OFDM) has been proposed in [21, 31]. In this scheme, the information bits of each user are spread over a subset of sub-carriers. Therefore, transmission of each user will only be interfered by a small number of users. Similar to LDS-CDMA, detection of LDS-OFDM can be based on the message passing algorithm (MPA). An LDS-OFDM system can be considered as a system that uses LDS as multiple access technique and OFDM for multi-carrier transmission. In [21], we have shown that LDS-OFDM outperforms OFDMA in terms of bit error rate (BER). Figure 2 shows the BER results of 200 % loaded BPSK LDS-OFDMA and QPSK OFDMA when both are using the same error correcting convolutional code with rate 1/2. As we can see from this Figure, even for system with channel coding still LDS-OFDMA maintains its superiority to OFDMA systems by approximately 3.4 dB and 5.54 dB at a BER of $10^{-3}$ and $10^{-5}$, respectively.

One of the non-trivial and open issues for LDS schemes is the analysis of its achievable rate region. Analysis of the conventional MAC for fading channels has already been studied in the literature in [5, 39], where two notions of capacity for the fading channel are considered. The first is the classic notion of Shannon capacity with a constraint on BER, which is directly applied to fading channels. The second analysis defines an alternative notion of capacity by setting a constraint on coding delay for fading channels.

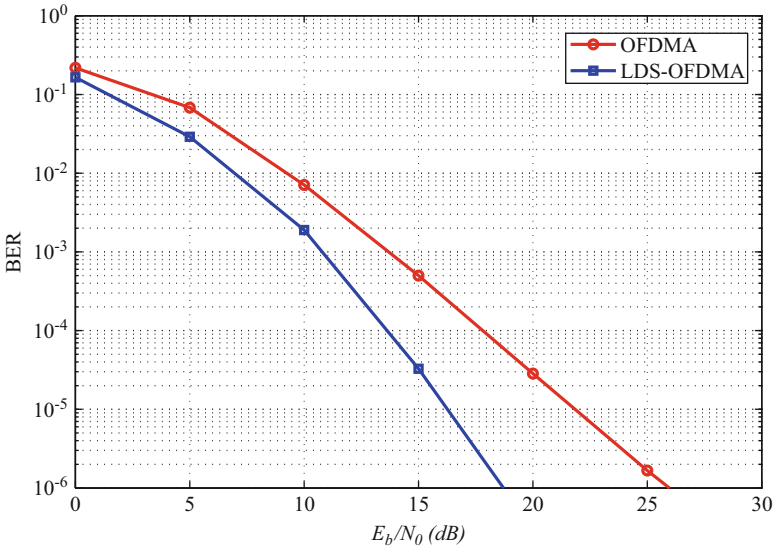


**Fig. 2** QPSK OFDMA vs. 200 % overloaded LDS-OFDMA BER performance for coded case

## 3.1 System Model

We consider a multiple access channel (MAC) which typically corresponds to the uplink communications from multiple users to a single base station in practical system. The following sub-sections explain the important aspects of our system model, including the block diagram of LDS schemes, the corresponding LDS design, and the signal model for LDS schemes.

### 3.1.1 LDS Schemes

The block diagram of a general LDS based system is shown in Fig. 3. In this scheme, unique low density spreading sequences (signatures) are assigned to each user, and the receiver has full knowledge of these assignments. Low density spreading sequences are designed in a way that allow each user to spread its data only over a small number of chips. Then, zero padding is performed such that the processing gain remains the same as that of a conventional CDMA system. The spreading sequences are then uniquely permuted for each user such that the resultant signature matrix become highly sparse. As shown in Fig. 3, for both LDS-CDMA and LDS-OFDM systems the encoded users' data will be first spread using low density signature. Then, for LDS-CDMA, the sub-carrier mapper allocate one wideband carrier to all users, while for LDS-OFDM, the sub-carrier mapper assign orthogonal sub-carriers to different users and generates an OFDM signal. In other words, for LDS-OFDM the spreading is followed by OFDM modulation, which means that each chip is mapped to a single sub-carrier. Thus, the combination of the spread users' symbols plus the AWGN form the received OFDM symbol. As shown in Fig. 4, an LDS system with $K$ users and $N$ chips can be represented by a factor graph $\mathcal{G}(\mathcal{U}, \mathcal{C})$, where the symbols of users are represented by variable nodes $u \in \mathcal{U}$ and chips are represented by function nodes $c \in \mathcal{C}$. In this graph edges

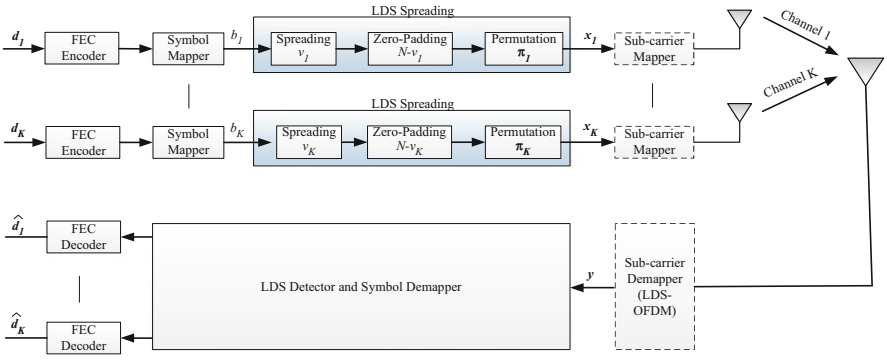


**Fig. 3** Block diagram of LDS scheme

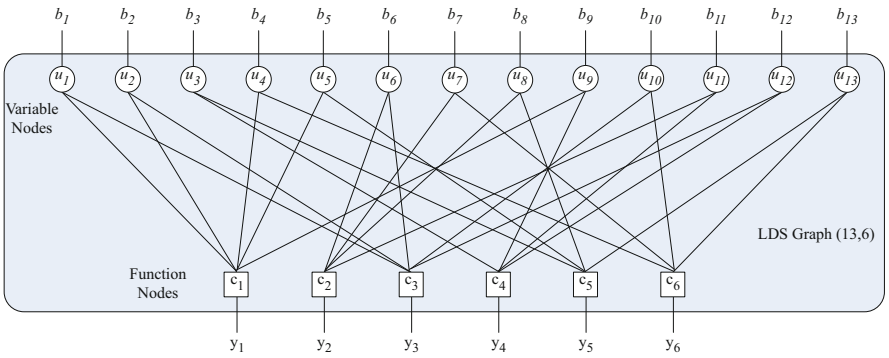


**Fig. 4** Factor graph representation of LDS scheme

represent the connection between the received chip and the corresponding users. At the receiver, chip level iterated (CLi) MUD is used to detect different users' symbols [20]. In this iterative MUD, using MPA in an iterative manner, messages containing the reliability values of the corresponding variable nodes are exchanged between the variable nodes and function nodes. At the end, after appropriate number of iterations the soft output which is the calculated log likelihood ratio at each variable node will be sent to the channel decoder.

### 3.1.2 Low Density Signature Design

In this section, we explain how low density signatures, i.e., the sequences which are mainly populated with zeros, are created. This structure helps switch off a large number of chips in each user's signature. We consider an LDS system with $K$ users with user indices $k = 1, \ldots, K$. LDS with density factor $0 < \zeta_k \leq 1$, which represents the ratio of the number of chips used by $k$th user to the total number of chips $N$, is generated in a process with the following steps:

1. A random signature (vector) with length $\nu_k = \zeta_k N$ for user $k$ is generated. Each element of this vector is a random variable with a Gaussian distribution. This means that the effective processing gain for user $k$ is equal to $\nu_k$.
2. Zero padding is carried out by adding $N - \nu_k$ zeros at the end of the signature for user $k$ such that the processing gain of the system becomes $N$.
3. The permutation pattern $\boldsymbol{\pi}_k = \{\pi_k(1), \pi_k(2), \ldots, \pi_k(N)\}$ is chosen randomly and applied to the zero-padded signatures of user $k$. Thus, the none-zero components of the permuted signature for user $k$ will appear in $\mathscr{C}_k = \{\pi_k(1), \pi_k(2), \ldots, \pi_k(\nu_k)\}$ positions. Furthermore, we define $\mathscr{P} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_K\}$ as the set of all users' permutation patterns.

Let $\boldsymbol{s}_k = [s_{k,1}, \ldots, s_{k,N}]^T$ be the generated signature for user $k$. By arranging the generated signature for all users the LDS matrix $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K]$ will be produced, where column $k$ represents the permuted signature of user $k$.

### 3.1.3 LDS Signal Model

As mentioned above, we consider the uplink scenario where a set of $K$ users communicate to a single receiver. The conceptual system block diagram for this scenario is shown Fig. 3. In this model, $\boldsymbol{x}_k$ is the sequence of spread symbols for any single modulated symbol of user $k$, namely, $b_k$ at the transmitter side:

$$\boldsymbol{x}_k = b_k \boldsymbol{s}_k, \tag{1}$$

where $\boldsymbol{s}_k$ is the signature for user $k$. Denote $\boldsymbol{y} = [y_1, \ldots, y_N]^T$ as the corresponding received sequence at the receiver side after match filtering and equalization. Considering a perfect symbol-synchronous conventional CDMA system under flat fading channel at chip level, the received signal sampled at $i$th time instance (chip) can be written as follows [41]:

$$y_i = \sum_{k=1}^{K} b_k s_{k,i} h_{k,i} + v_i, \tag{2}$$

where $v_i$ is the complex white Gaussian noise with variance per real symbol $N_0/2$ and $h_{k,i}$ is the fading channel gain of the $k$th user at time instance $i$. User $k$ is also subject to an average transmitter power constraint of $\bar{P}_k$. This system model is provided for a time varying channel, therefore, according to [30] the coherence time of channel for a vehicle traveling 60 mph using a 900 MHz carrier is around 2.22 ms while according to Interim Standard 95 (IS-95), the symbol length is around 10 ms which is more than the coherence time. Therefore, in our system model a unique fading coefficient is not multiplied by all chips.

In LDS schemes, for a given time instance $i = 1, \ldots, N$; let $\mathscr{K}_i$ denote the set of users that are active at time instance $i$, i.e., $\mathscr{K}_i = \{k | i \in \mathscr{C}_k\}$. Considering the LDS scheme explained earlier, for LDS-CDMA, (2) can be given by

$$y_i = \sum_{k \in \mathscr{K}_i} b_k s_{k,i} h_{k,i} + v_i. \tag{3}$$

This equation demonstrates that the number of interfering users on time instance $i$ is significantly reduced in an LDS system. Note that in this model, we consider fading effects which are frequency nonselective. For LDS-OFDM, for both selective and nonselective frequency fading, the received signal is the same as (3) if $i$ is equal to the frequency index.

### 3.2 Achievable Rate Region of LDS-CDMA

Capacity of conventional MAC is analysed in [7, 10, 11, 19, 32]. In multi-user wireless communication, the channel capacity is characterized by achievable rate region, where each point in the region represents a vector of simultaneously achievable rates for the users, constrained by probabilities of bit or symbol error rate for each user [10]. The set of all such achievable rate vectors is known as the capacity region, and dubbed as the achievable rate region for LDS based multiple access in this chapter.

We shall now focus on the scenario of interest in this section, where the channels are time-varying. It is assumed that the receiver has full channel state information (CSI). The capacity region of time varying fading MAC is well known [16, 34]; it is the set of all rate vectors $\boldsymbol{R}$ calculated as follows:

$$\mathscr{C}_{MAC}(\boldsymbol{h},\boldsymbol{P}) = \left\{ \boldsymbol{R} : \boldsymbol{R}(\mathscr{S}) \leq W\mathbb{E}_{\boldsymbol{h}}\left[ \log\left(1 + \frac{\sum_{k\in\mathscr{S}} |h_k|^2 P_k}{WN_0}\right)\right], \forall \mathscr{S} \subseteq \{1,\ldots,K\} \right\}, \tag{4}$$

where $\boldsymbol{P} = [P_1,\ldots,P_K]$ and $\boldsymbol{h} = [h_1,\ldots,h_K]$ is a random vector with elements which have the stationary distribution of the joint fading processes and $W$ is the bandwidth of the channel. In this chapter, for any vector $\boldsymbol{v}$, the notation $\boldsymbol{v}(\mathscr{S})$ is used to denote $\sum_{i\in\mathscr{S}} \boldsymbol{v}(i)$. A rigorous proof of this result can be found in [34]. An intuitive understanding of this result can be obtained by viewing capacities in terms of time averages of mutual information [16].

The achievability of the upper bound on the rate region is subject to the fact that received messages are decoded to a message in the codebook which is uniquely jointly typical [10]. For a fixed density factor $\zeta_k$, as the overall signature length approaches infinity, the length of the full density part will also approach to infinity. We can extend the joint typicality definition to low density signatures by applying the original definition. A detailed proof is given in the Appendix.

As mentioned earlier, the spreading sequence for user $k$ contains many zeros and some non-zero elements. Thus, the achievable rate region of LDS-CDMA under time varying fading channel can be derived by defining an equivalent LDS channel: $\tilde{h}_{k,i} = a_{k,i}h_{k,i}$ where $a_{i,k}$ denotes the activity of user $k$ at time $i$, specifically $a_{k,i}$ will be equal to one if $k \in \mathscr{K}_i$ otherwise it will be zero. For notation brevity, all users are assumed to have the same density factor, i.e., $\zeta = \zeta_k, \forall k = 1,\ldots,K$. Thus, in order to derive the achievable rate region of fading LDS-CDMA we must replace the original fading MAC channel with its LDS equivalent channel in (4) as shown bellow

$$\mathscr{C}_{LDS}(\tilde{\boldsymbol{h}},\tilde{\boldsymbol{P}}) = \left\{ \boldsymbol{R} : \boldsymbol{R}(S) \leq W\mathbb{E}_{\tilde{\boldsymbol{h}}}\left[ \log\left(1 + \frac{\sum_{k\in\mathscr{S}} \left|\tilde{h}_k\right|^2 \tilde{P}_k}{WN_0}\right)\right], \forall \mathscr{S} \subseteq \{1,\ldots,K\} \right\}, \tag{5}$$

where $\tilde{P}_k : P_k/\zeta$, as for a proportion of time the users don't send data, thus, their transmit power denoted by $\tilde{P}_k$ will be larger than their original transmit power.

More specifically, for a fixed permutation pattern and thus a fixed low density structure, the derived achievable rate region will be

$$\mathscr{C}_{LDS}(\boldsymbol{h}, \tilde{\boldsymbol{P}}) = \left\{ \boldsymbol{R} : \boldsymbol{R}(\mathscr{S}) \leq \frac{1}{N} \sum_{i=1}^{N} W \mathbb{E}_{\boldsymbol{h}} \left[ \log \left( 1 + \frac{\sum_{k \in \mathscr{S}} a_{k,i} \left| h_k \right|^2 \tilde{P}_k}{W N_0} \right) \right], \forall \mathscr{S} \subseteq \{1, \ldots, K\} \right\}. \tag{6}$$

Now if we take into account the ensemble of LDS scheme over all possible permutation patterns, the activity parameters will be independent and follow Bernoulli distribution with probability of being 1, given as $\zeta$. Therefore, the achievable rate region will be

$$\mathscr{C}_{LDS}(\boldsymbol{a}, \boldsymbol{h}, \boldsymbol{P}) = \left\{ \boldsymbol{R} : \boldsymbol{R}(\mathscr{S}) \leq W \mathbb{E}_{\boldsymbol{a},\boldsymbol{h}} \left[ \log \left( 1 + \frac{\sum_{k \in \mathscr{S}} a_k \left| h_k \right|^2 P_k / \zeta}{W N_0} \right) \right], \forall \mathscr{S} \subseteq \{1, \ldots, K\} \right\}, \tag{7}$$

where $\boldsymbol{a} = [a_1, \ldots, a_K] \in \{0, 1\}^K$ is a vector that contains Bernoulli components with probability $\zeta$. It can be noted that the derived model allows any number of users to be active in a given symbol because of the Bernoulli distribution of the activity of the users. In this regard, we call the ensemble of the produced low density signatures as $\mathrm{LDS}_K(\zeta, K)$. If the number of active users is limited, it becomes less than a specific number $\xi \leq K$, the generated LDS will be called $\mathrm{LDS}_K(\zeta, \xi)$. In this case, the achievable rate region still will have the same form as (15), but the distribution for the elements of activity vector $\boldsymbol{a}$ must be modified. The original independent and identically distributed Bernoulli distribution of the element of the activity vector imply

$$p(\boldsymbol{a}; \zeta, K) = \prod_{k=1}^{K} \zeta^{a_k} (1 - \zeta)^{1 - a_k}. \tag{8}$$

On the other hand, the distribution for $\mathrm{LDS_K}(\zeta, \xi)$ can be expressed in the following form:

$$p(\boldsymbol{a}; \zeta, \xi) = \begin{cases} \beta \prod_{k=1}^{K} \zeta^{a_k} (1 - \zeta)^{1 - a_k}, & \sum_{k=1}^{K} a_k \leq \xi \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where $\beta$ is a normalizing factor and is expressed as follows:

$$\beta = \left( \sum_{d=0}^{\xi} \binom{K}{d} \zeta^d (1-\zeta)^{K-d} \right)^{-1}. \tag{10}$$

It should be noted that the channel resources will be unused if there exist a chip with no active user in our signature. Therefore, it is necessary to further restrict the LDS ensemble to those that have at least one user active at a given time instance. Inspired by the schemes used for expurgating the codes and the union bound for error probability [4, 6], we expurgated the condition where no user is active at a chip or time instance. This expurgated scheme will be called $\mathrm{LDS}_K(\zeta, \xi^+)$ with the activity vector distribution

$$p(\boldsymbol{a}; \zeta, \xi^+) = \begin{cases} \beta^+ \prod_{k=1}^{K} \zeta^{a_k} (1-\zeta)^{1-a_k}, & 0 < \sum_{k=1}^{K} a_k \leq \xi \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where

$$\beta^+ = \left( \sum_{d=1}^{\xi} \binom{K}{d} \zeta^d (1-\zeta)^{K-d} \right)^{-1}. \tag{12}$$

By introducing the constraint of maximum number of active users at a specific time instance, the probability of having value 1 in each element of activity vector, $\hat{\zeta}$, will diverge from its initial value, $\zeta$, according to number of active users $d$. $\hat{\zeta}$ can be calculated as follows:

$$\hat{\zeta} = \sum_{d=1}^{\xi} p(d)\, \hat{\zeta}(d), \tag{13}$$

$$p(d) = \binom{K}{d} (1-\zeta)^{K-d} \zeta^d \beta^+ \quad \& \quad \hat{\zeta}(d) = \frac{d}{K}, \tag{14}$$

$$\Rightarrow \quad \hat{\zeta} = \sum_{d=1}^{\xi} \binom{K}{d} \zeta^d (1-\zeta)^{K-d} \beta^+ \frac{d}{K}. \tag{15}$$

By plugging $\hat{\zeta}$ from (15) into (7) we obtain the achievable rate region of LDS-CDMA. Monte Carlo simulations can be used to compute and evaluate achievable rate of LDS-CDMA given by (7). Here, we present the simulation results for a representative scenario with 20 users. The channel bandwidth is considered to be 5 MHz. For comparison, achievable sum rate (SR) of LDS-CDMA is compared

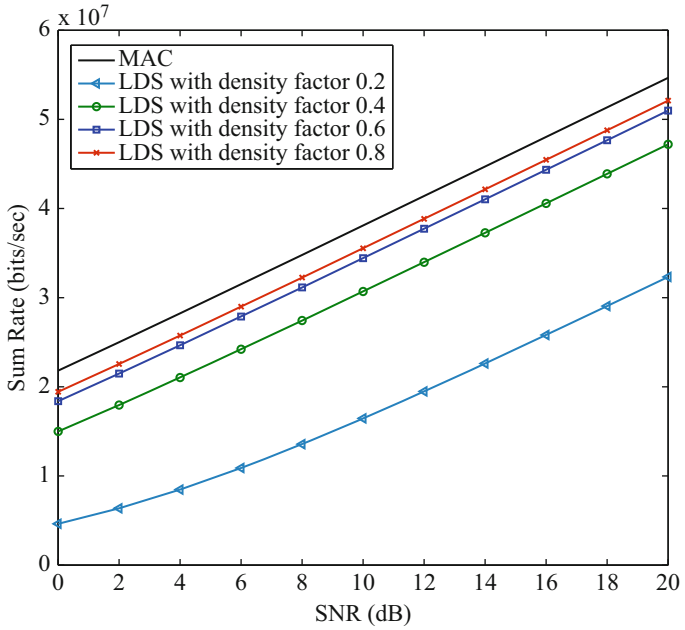


**Fig. 5** Sum rate for different density factors with maximum number of allowed users per chip equal to 8

to that of conventional MAC in a similar scenario. Figures 5 and 6 show the comparison between the SR of LDS-CDMA and conventional MAC for various density factors and maximum number of users allowed per chip, respectively. As it can be seen from Fig. 5, conventional MAC outperforms LDS-CDMA in terms of achieved sum rate. This performance loss is traded for significant reduction of the complexity as it will be discussed later. These figures show that for LDS systems, increasing $\zeta$ will increase the SR and reducing the maximum number of allowed users at each time instance will also result in higher SR. Furthermore, Fig. 5 implies that there is a slightly large gap between the SR curves for density factor 0.4 and 0.2, while the gap decrease dramatically for density factors larger than 0.4. Thus, density factor 0.4 would be an optimal choice as we would loos less capacity while reducing the complexity enormously.

As shown in Fig. 7, by applying LDS structure to MAC, the complexity order of the MUD will reduce from total number of users, $\mathcal{O}\left(|\mathbb{X}|^K\right)$, to the maximum number of allowed users per chip, $\mathcal{O}\left(|\mathbb{X}|^{\eta_n}\right)$. Figure 7 shows that by decreasing the density factor, the computational complexity of MUD for LDS systems reduces significantly.

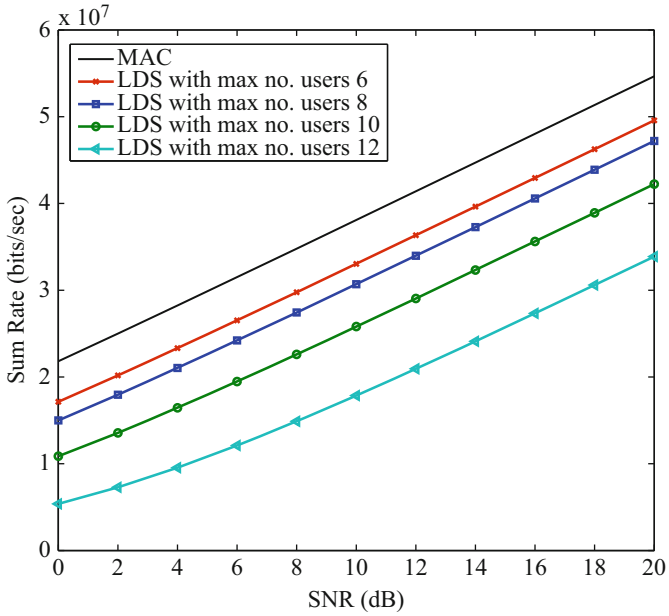


**Fig. 6** Sum rate for different maximum number of allowed users per chip and density factor equal to 0.4

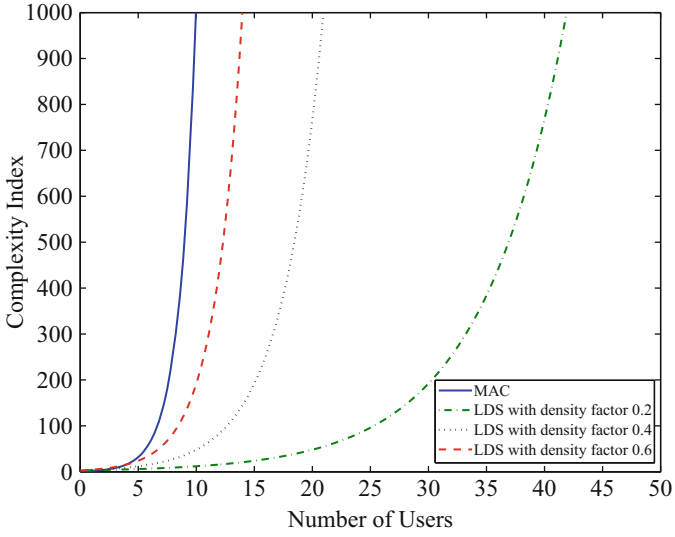


**Fig. 7** Complexity of MUD for MAC in comparison with LDS with different density factors

### 3.3 Achievable Rate Region for LDS-OFDM

OFDM is widely used in current wireless communications systems and is a core technique for the fourth generation (4G) cellular telephony. OFDM provides a high spectral efficiency over a frequency selective channel [2]. In this regard under wideband channels applying LDS scheme to OFDM would help us avoid multi-path fading effect on the density of graph and have a better performance. In LDS-OFDM, each chip will play the role of an OFDM sub-carrier to make the system more robust in frequency selective environment.

In this section, we aim to find the achievable rate region for LDS-OFDM. To this end, we will firstly explain how LDS-OFDM differs from LDS-CDMA conceptually.

In LDS-OFDM, instead of sending data on different chips or time instances, the data is spread on orthogonal sub-carriers, thus, in LDS-OFDM we switch from time domain to frequency domain. Figure 8 shows how the two schemes are different in variables in our analysis and different textures show different users that share the same chip or frequency. This figure also shows that power allocation is not necessarily uniform; thus, a data symbol might be transmitted over its assigned sub-carriers with different power levels and with different complex phases. The analysis in previous section can be extended to LDS-OFDM by replacing time instance parameters with frequency, thus, we will have

$$\tilde{h}_k(f) = a_{k,f} h_k(f), \tag{16}$$

where $a_{k,f}$ denotes the activity of user $k$ at sub-carrier $f$, in other words, $a_{k,f}$ is equal to one if user $k$ spread its data over sub-carrier $f$ otherwise it will be zero. Thus, the achievable rate region for LDS-OFDM is calculated by finding the expected value of the channel on frequency domain

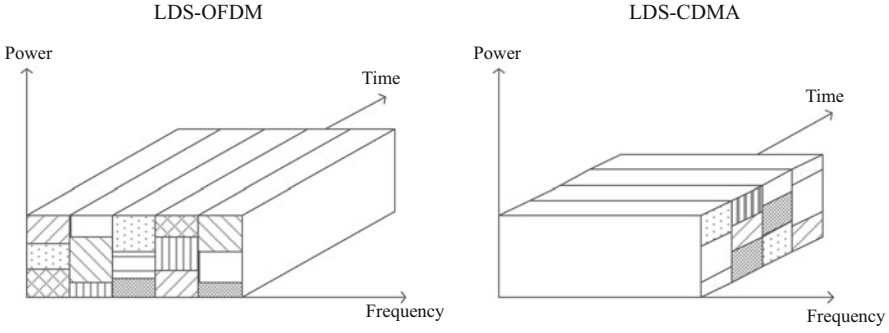


**Fig. 8** LDS-OFDM vs. LDS-CDMA

$$\mathscr{C}_{LDS-OFDM}(\boldsymbol{a},\boldsymbol{h},\boldsymbol{P}) =$$
$$\left\{\boldsymbol{R} : \boldsymbol{R}(\mathscr{S}) \leq W\mathbb{E}_{\boldsymbol{a},\boldsymbol{h}}\left[\log\left(1+\frac{\sum_{k\in\mathscr{S}} a_k\left|h_k\right|^2 P_k/\zeta}{WN_0}\right)\right], \forall \mathscr{S} \subseteq \{1,\ldots,K\}\right\}. \tag{17}$$

This shows the mentioned analysis is valid for LDS-OFDM as well.

### 3.4 *Effect of Irregularity on LDS Schemes*

The uplink BER performance of a multi-user system such as LDS-CDMA can be improved significantly by careful design of low density signature. As discussed in Sect. II, an LDS system can be represented by a factor graph, where the symbols of users and chips are presented by variable nodes and function nodes respectively. The graph edges represent the connection between the received chip and the corresponding user. Therefore, this graph can be called regular if all of its variable nodes and function nodes have the same number of edges. A main factor in design of low density signature is its regularity. This analysis will help us evaluate the effect of irregular LDS structure on performance improvement and its cost in terms of complexity. Throughout this section, for ease of analysis, the LDS structures are classified as:

- Regular: When $v_k = v,\ \forall k = 1,\ldots,K$, and $\eta_n = \eta,\ \forall n = 1,\ldots,N$
- Irregular: Otherwise

In order to analyse the effect of irregularity, initially the effect of number of active users on sum rate is shown in Fig. 9. The evaluation is done for a system with 20 users with SNR=20 dB and density factor equal to 0.4. The difference between the SR for different number of active users is the penalty which is paid for reducing the complexity in LDS. For irregular LDS, different number of users are active at different chips. We define

$$\alpha = W\log\left(1+\frac{\sum_{k\in\mathscr{S}} a_k\left|h_k\right|^2 P_k/\zeta}{WN_0}\right). \tag{18}$$

Considering that SR is related to $\mathbb{E}_{\boldsymbol{a},\boldsymbol{h}}[\alpha]$, we can say that the expectation is taken over different combinations of active users. Having $\boldsymbol{a} \in \mathscr{A}$, where $\mathscr{A}$ is the set of possible combinations of active users,

$$\mathscr{A} : \left\{\boldsymbol{a} \middle| \sum_{k=1}^{K} a_k = d_1 \ \text{ or } \sum_{k=1}^{K} a_k = d_2 \ \ldots \ \text{ or } \sum_{k=1}^{K} a_k = d_\xi\right\}. \tag{19}$$

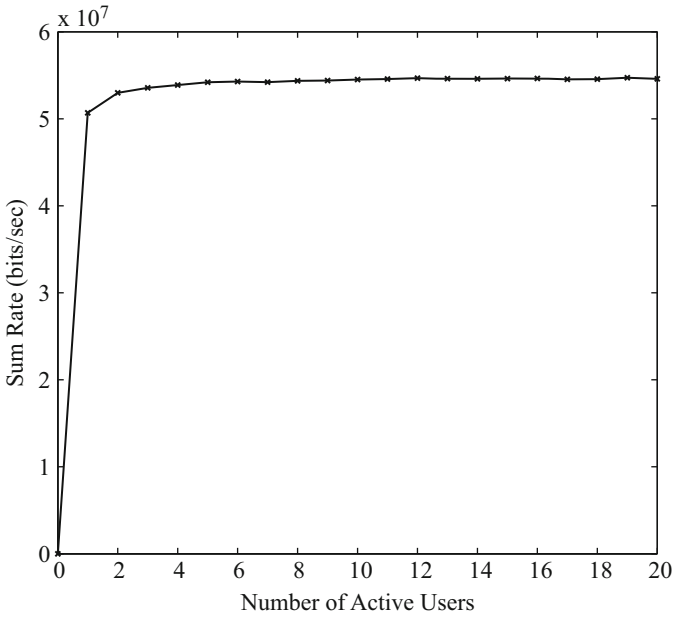


**Fig. 9** Effect of number of active users on sum rate

By dividing the set $\mathscr{A}$ in $\xi$ subsets $\mathscr{A}_1, \ldots, \mathscr{A}_\xi$ we will have

$$\mathbb{E}_{\boldsymbol{a},\boldsymbol{h}}[\alpha] = \mathbb{E}_{\boldsymbol{a}|\mathscr{A}_1,\boldsymbol{h}}[\alpha]\, p(\mathscr{A}_1) + \mathbb{E}_{\boldsymbol{a}|\mathscr{A}_2,\boldsymbol{h}}[\alpha]\, p(\mathscr{A}_2) + \ldots + \mathbb{E}_{\boldsymbol{a}|\mathscr{A}_\xi,\boldsymbol{h}}[\alpha]\, p\left(\mathscr{A}_\xi\right), \quad (20)$$

where

$$\mathscr{A}_j : \left\{ \boldsymbol{a} | \sum_{k=1}^{K} a_k = d_j \right\}. \quad (21)$$

This shows that $\mathbb{E}_{\boldsymbol{a},\boldsymbol{h}}[\alpha]$ is linear with respect to $\mathbb{E}_{\boldsymbol{a}|A_j,\boldsymbol{h}}[\alpha]$. Thus, the rate of LDS scheme with different number of active users is linear with respect to the rate of LDS scheme with fixed number of active users at all chips. On the other hand, the complexity increases exponentially by the number of active users. Therefore, by combining various number of active users (irregular LDS) we will have more complexity for a fixed SR, because the SR changes linearly while complexity changes exponentially. For instance if we consider a system with number of active users 4 and 6, the SR will be similar to a regular system with 5 active users (due to linearity) while the complexity is more as it is shown in Fig. 10. Therefore, regular LDS design introduces less complexity while keeping the SR the same.

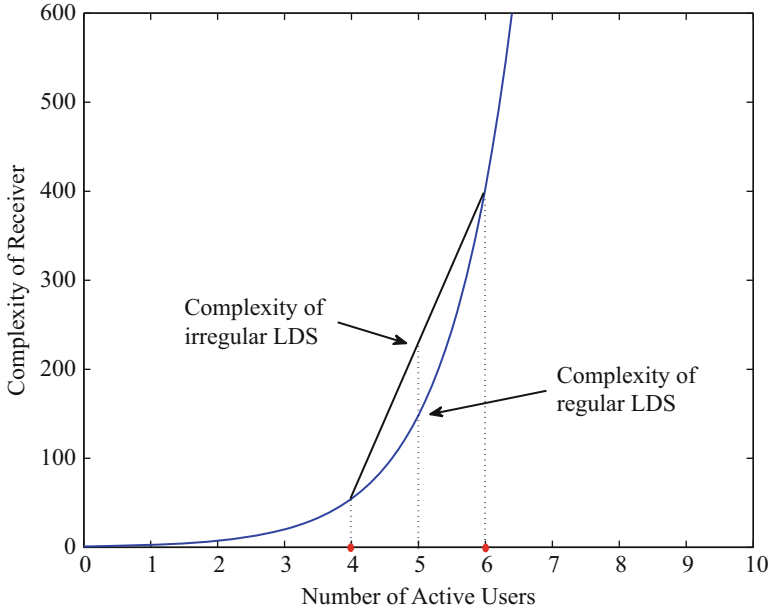


**Fig. 10** Complexity of irregular LDS in comparison with similar regular LDS

## *3.5 Analysis of Achievable Region of LDS System with Practical Modulation and Coding Schemes*

We analysed a system with infinite constellation space size. However, practical systems employ a finite constellation size. To analyse a practical system with finite modulation size, in this section we extend our analysis to a system with Quadrature Amplitude Modulation (QAM) modulation and convolutional channel coding.

### 3.5.1 Achievable Rate Region of QAM LDS System

The SNR gap approximation, first introduces in [9], is an effective tool to model the achievable rates when a suboptimal transmission scheme is used for a given probability of symbol error, $P_e$. Thus, using the gap approximation the rate achieved by a discrete constellation over an AWGN channel can be expressed in a form similar to Shannon's capacity formula. In this chapter, for finding the effect of modulation on the achievable rate of LDS schemes we introduce the SNR gap concept. To achieve a certain data rate for a given probability of error with modulation, a value of SNR greater than that suggested by Shannon's capacity is required. This extra amount of power is called the SNR gap, denoted in this chapter by $\Gamma_{\text{mod}}$. The SNR gap is widely studied for single user case under AWGN

channel for different modulation types by considering the error probability relation with modulation order and SNR. Since LDS scheme is a multi-user scenario, we are interested in the analysis of the SNR gap for multi-user case, provided in [15]. On the other hand for the case of fading channel with non-adaptive transmission (constant transmit power and rate), the average BER is obtained by integrating the BER in AWGN over the fading distribution [17]. Considering that in previous section, for calculating the achievable rate the average was taken over fading process, the calculation of achievable rate of modulated LDS system can be performed for fading channel straightforwardly. The sum rate for the LDS system where users apply the same modulation constellation, $\mathscr{C}_{M-LDS}$, can be expressed as follows:

$$\mathscr{C}_{M-LDS}(\boldsymbol{a},\boldsymbol{h},\boldsymbol{P}) = \left\{\boldsymbol{R} : \boldsymbol{R}(S) \leq W\mathbb{E}_{\boldsymbol{a},\boldsymbol{h}}\left[\log\left(1+\frac{\sum_{k\in S} a_k |h_k|^2 P_k/\zeta}{\Gamma_{\text{mod}}\, WN_0}\right)\right], \forall S \subseteq \{1,\ldots,K\}\right\}. \tag{22}$$

For a system with M-QAM modulation $\Gamma_{\text{mod}}$ is calculated as follows [38]:

$$\Gamma_{\text{mod}}(P_e, M) = \left[Q^{-1}\left(\frac{MP_e}{2(M-1)}\right)\right]^2/3. \tag{23}$$

In order to have a fair comparison for different modulation orders we need to consider the $E_b/N_0$ gap, $\Gamma'_{\text{mod}}$, which is calculated as follows [3]:

$$\Gamma'_{\text{mod}}(P_b, M) = \left[Q^{-1}\left(\frac{P_b \log_2 M}{4\left(1-2^{\frac{\log_2 M}{2}}\right)}\right)\right]^2/3. \tag{24}$$

By plugging $\Gamma'$ from (32) into (30) we obtain the achievable rate region of modulated LDS MAC. The relation between SR and $E_b/N_0$ is presented in Fig. 11 for an LDS system with $\zeta$ equal to 0.8 and maximum of 8 users allowed per chip. Furthermore, this figure shows that the comparison of different modulation orders; 64-QAM, 16-QAM, and QPSK (quadrature phase shift keying) for $P_b = 10^{-6}$. Figure 11 shows the SR is reduced 10 Mbits/sec at $E_b/N_0 = 14$ for QPSK modulation, thus, even at reasonably good $E_b/N_0$, the modulated scheme is about 6 dB, short of achievable rate. Also it is shown higher order modulations leads to better SR which matches the results given in [14].

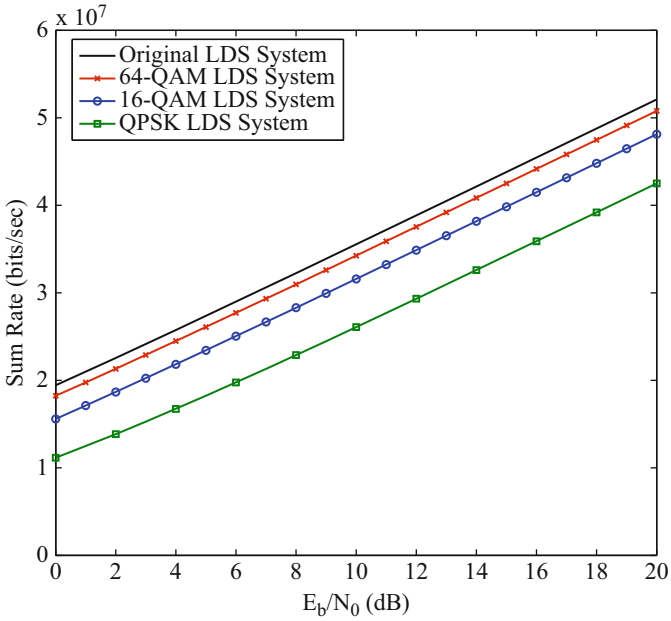


**Fig. 11** Effect of modulation order on sum rate of LDS system

### 3.5.2 Achievable Rate Region of Convolutional Coded LDS System

Having introduced commonly used signal constellations for data transmission of LDS scheme, several performance measures are used to compare the coded systems based on these constellations. In order to take into account the effect of coding in the achievable rate region, coding gain is defined [9]. Coding gain, $\gamma$, specifies the improvement of one constellation over another when used to transmit the same information. Once the gain with respect to un-coded transmission for a given coding method is known, the gap becomes

$$\Gamma_{\text{cod}} = \Gamma_{\text{mod}} - 10 \log_2 \gamma_{\mathit{eff}}, \tag{25}$$

where $\gamma_{\mathit{eff}}$ is the effective coding gain. For a rate $k/n$ binary linear convolutional code $C$ with free distance $d$, the nominal coding gain can be calculated as follows [14]:

$$\gamma(C) = \left(\frac{k}{n}\right) d. \tag{26}$$

To estimate from $\gamma(C)$ the effective coding gain $\gamma_{\mathit{eff}}(C)$, we use the normalized error coefficient $K(C) = K_d/k$, and apply the rule introduced in [14]. Here, $K_d$ is the number of code sequences with minimum weight $d$ in a given $k$-input

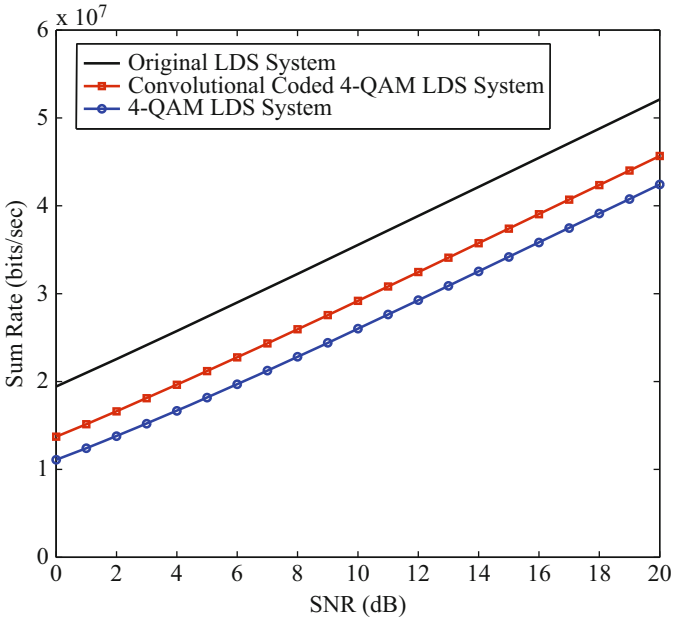


**Fig. 12** Effect of coding on sum rate of LDS system

$n$-output block. We should emphasis that estimated effective coding gain assume soft decisions, maximum likelihood decoding.

Figure 12 shows the effect of coding on the sum rate of LDS with density factor 0.8 and maximum number of allowed user equal to 8. The code type is a half rate convolutional code with $d = 5$ and $K_d$ equal to 1, thus, the nominal coding gain is $\gamma(C) = 5/2\,(4\,\text{dB})$, and the effective coding gain is also 4 dB. This figure illustrates that this simple convolutional code was able to increase the SR curve 4 Mbits/sec at SNR equal to 16 dB.

## *3.6 Frequency-Selective Fading Channels*

In the previous sections we have analysed a flat fading model which is appropriate if the Nyquist sampling period is large compared to the delay spread of the multi-paths in the received signal, so that the individual paths are not resolvable in the sampled system. For wideband applications, the multi-paths can be resolved, and hence channel has memory. The appropriate model in this case is the time-varying frequency-selective fading channel. In this section, we will extend our previous results for LDS-CDMA to this model. For LDS-OFDM, frequency-selective fading channel is converted to parallel flat-fading channels in frequency domain, thus, the model remains the same for frequency selective fading channels.

First the signal model for LDS-CDMA under frequency selective fading channel must be driven. The baseband signal model for conventional MAC is well known [17]. Calling the notations given in Sect. 3.1.3, the signal model for LDS-CDMA is shown for a channel with $l$ paths (which results in $l$ taps in the impulse respond for fading multi-path channel),

$$y_i = \sum_{k \in \mathscr{K}_i} \sum_{l} h_{k,i,l} x_{k,i-l} + v_i. \tag{27}$$

We denote $h_{k,i,l}$ as the $l$th channel filter tap for user $k$ at time instance $i$ and $x_{k,i} = b_k s_{k,i}$ similar to the expression given in Sect. 3.1.3. In order to analyse the achievable rate region of this channel, when both the transmitters and the receiver can track the channel, we will first focus on the case where the channel is time-invariant. Therefore, we will have

$$y_i = \sum_{k \in \mathscr{K}_i} \sum_{l} h_{k,l} x_{k,i-l} + v_i. \tag{28}$$

Considering that this channel has the characteristics of a Gaussian multiple access channel with ISI, the achievable rate region of this channel for conventional MAC is given in [7]. For LDS-CDMA as discussed earlier the channel is multiplied by the activity parameters which are independent and follow Bernoulli distribution. Considering that the distribution of activity parameters will remain Bernoulli after the Fourier transform, for LDS-CDMA the achievable rate region is

$$\mathscr{C}_{LDS}(\boldsymbol{a}, \boldsymbol{P}) = \Bigg\{ \boldsymbol{R} : \boldsymbol{R}(\mathscr{S}) \leqslant \mathbb{E}_{\boldsymbol{a}} \left[ \int_{-\frac{W}{2}}^{\frac{W}{2}} \log \left( 1 + \frac{\sum_{k \in \mathscr{S}} a_k \left| \hat{h}_k(f) \right|^2 P_k / \zeta}{W N_0} \right) df \right] \quad \forall \mathscr{S} \subset \{1, \ldots, K\} \Bigg\}. \tag{29}$$

where $\hat{h}_k(f)$ is the Fourier transform of the channel. Next the general case where the channel is time-varying is analysed. To do so we assume that the time variations of the channel are random and ergodic and the channel varies very slowly compared to the multi-path delay spread. The achievable rate region related to such channel for a convectional MAC is given in [16]. For LDS-CDMA after applying the concept of activity parameters the achievable rate region will be

$$\mathscr{C}_{LDS}(\boldsymbol{a}, \hat{\boldsymbol{h}}, \boldsymbol{P}) = \Bigg\{ \boldsymbol{R} : \boldsymbol{R}(\mathscr{S}) \leqslant \mathbb{E}_{\boldsymbol{a}, \hat{\boldsymbol{h}}} \left[ \int_{-\frac{W}{2}}^{\frac{W}{2}} \log \left( 1 + \frac{\sum_{k \in \mathscr{S}} a_k \left| \hat{h}_k(f) \right|^2 P_k / \zeta}{W N_0} \right) df \right] \quad \forall \mathscr{S} \subset \{1, \ldots, K\} \Bigg\}, \tag{30}$$

where $\hat{\boldsymbol{h}} = [\hat{h}_1(f), \ldots, \hat{h}_K(f)]$, thus the expectation is taken over fading states and the activity elements. For each time instance $t$, $\hat{h}_k(f, t)$ is the frequency response of user $k$ at fading state $t$.

We use a numerical example to evaluate the validity of our assumption of slow variations of the channel in time domain compared to multi-path delay spread in practice. Consider a typical mobile scenario where the vehicle is moving at 60 km/h and the carrier frequency is 1 GHz. Time variation of the channel becomes uncorrelated every half carrier wavelength over distance [17]. Thus, the time constant associated with the fading effect is in the order of time taken for the mobile to travel one wavelength at the transmitter frequency which is 0.0185 s for this example. In indoor channels delay spread typically ranges from 10 to 1000 ns, in suburbs it ranges from 200 to 2000 ns, and in urban areas it ranges from 1 to 30 μs [17]. Hence, even for the urban areas, the time variation of the channel is significantly slower than the delay spread which indicate that the assumption is quite reasonable for typical wireless systems.

Figure 13 shows the SR for both LDS-OFDM and LDS-CDMA under frequency-selective fading channel. For simulations, ITU Vehicular B channel model is considered and all the 20 users share a total bandwidth of 60 MHz. As expected LDS-OFDM showed to have a better sum rate compared to LDS-CDMA, this is because LDS-OFDM is able to combat the negative effect of frequency-selective fading channel by dividing it into a number of flat fading sub-channels.

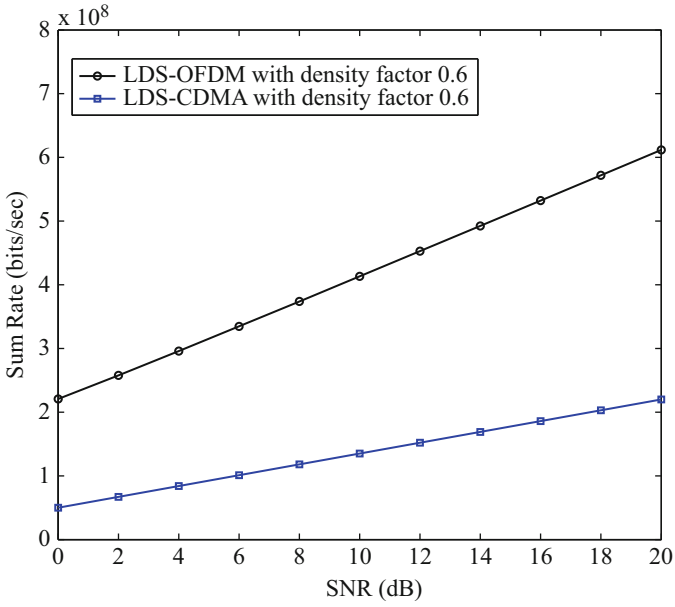


**Fig. 13** Sum rate for LDS-CDMA and LDS-OFDM with maximum number of allowed users per chip/sub-carrier equal to 8

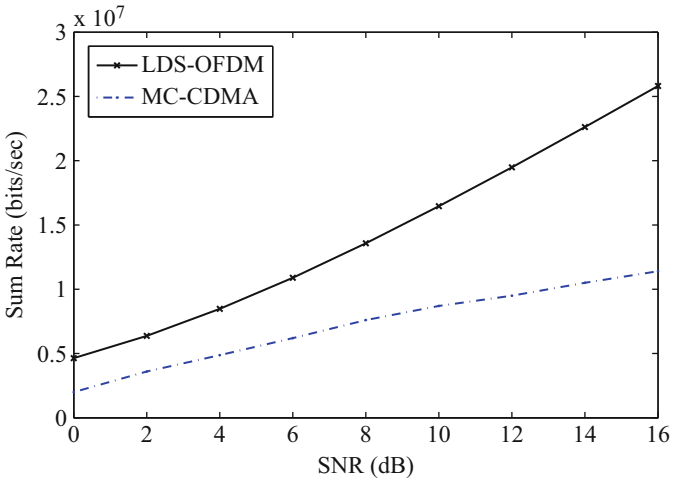


**Fig. 14** Sum rate for LDS-OFDM system in comparison with MC-CDMA system with MMSE MUD

Finally, in order to provide more motivation for using an LDS-OFDM system in practice we compare the sum-rate of the LDS-OFDM system with a low-complexity multiuser detection MAC system. Therefore, Fig. 14 shows the sum rate for an LDS-OFDM system with density factor equal to 0.2 and an MC-CDMA system with minimum mean-square error (MMSE) multiuser detector. The simulation results for the MC-CDMA system with load 2.5 are produced using the derivations carried out in [40]. As can be seen LDS-OFDMA is able to achieve higher sum rate under frequency selective fading channel, and this superiority is even more at higher SNRs.

## 4 Summary

In this chapter, we first reviewed different state of the art non orthogonal multiple access techniques. We then rigorously analysed the performance of LDS based non-orthogonal multiple access techniques that are considered to be among promising solutions for the next generation of cellular access networks due to their well-known advantages compared to state of the art techniques based on OFDMA. In this regard the achievable rate region of LDS schemes is calculated using information theoretic analysis by introducing an effective channel gain, based on the activity of users. It is shown that LDS significantly reduces complexity of multi-user detection at the price of moderate loss of the system performance in terms of achievable sum rate. Furthermore, the effect of LDS parameters such as density factor and maximum number of users at each time instance on the sum rate is evaluated. The results show that the achievable rate approaches to that of MAC when increasing the

density factor towards one. The effect of irregularity on the complexity has also been evaluated, which leads to the conclusion that regular signatures are better in terms of complexity. Furthermore, we have extended our model by taking into account practical system with a finite constellation size; the effect of finite size modulation was introduced into the our model using SNR gap. Furthermore, it was shown that the loss of achievable rate which had been caused by modulation can be compensated by using a suitable channel coding scheme. Finally, comparing with the MC-CDMA system with an MMSE MUD we showed that LDS-OFDM is able to achieve higher sum rate under frequency selective fading channel, and this superiority is even more at higher SNRs.

## Appendix: Joint Typicality for Low Density Signatures

Here we explain how the joint typicality definition is extended to low density signatures (codewords). The signatures can be consider as codewords because in theory coding is a scheme that maps a block of symbols to another block of variables which can be generated randomly. Therefore, considering that LDS scheme also maps a symbol to a block of variables which are generated randomly we can say that the signature matrix in LDS scheme plays the role of a codebook (please note that term coding here doesn't mean channel coding). First we assume that low density codebook is generated randomly using a probability distribution with high probability of random variable being equal to zero. Therefore, $(X_1, X_2, \ldots, X_k)$ denote a finite collection of discrete random variables with some fixed joint distribution. Let $\mathscr{B}$ denote an ordered subset of these random variables and consider $N$ independent copies of $\mathscr{B}$ denoted by $\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_N$. As before, let $\mathscr{C}_k$ denote the set of time instances over which user $k$ spreads its data, i.e., $\mathscr{C}_k = \{i | k \in \mathscr{K}_i\}$. We further define $\mathscr{D}(\mathscr{B})$ as the ordered set of time indices over which users related to the members of $\mathscr{B}$ are active, i.e., $\mathscr{D}(\mathscr{B}) = \bigcup_{k \in \hat{\mathscr{B}}} \mathscr{C}_k$, where $\hat{\mathscr{B}}$ is the ordered set of user indices that their random variables are a member of $\mathscr{B}$. Considering that each of the copies is independent of the other subsets, the joint probability mass function of the subsets will be

$$p(\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_N) = \prod_{i \in \mathscr{D}(\mathscr{B})} p(\mathscr{B}_i), \tag{31}$$

where $\mathscr{D}(\mathscr{B})$ has at least $\nu_k$ members, let $N'$ be the size of $\mathscr{D}(\mathscr{B})$. Considering that $\zeta_k = \frac{\nu_k}{N}$, for a fixed density factor, $\zeta_k$, as the overall signature length, $N$, approaches to infinity, the length of full density portion, $\nu_k$, will also approach to infinity. Thus, for sufficiently large $N$, $N'$ will approach infinity as well. Therefore, by the law of large numbers for sufficiently large $N$, for any subset $\mathscr{B}$ of random variables,

$$-\frac{1}{N'} \log p(\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_N) = -\frac{1}{N'} \sum_{i \in \mathscr{D}(\mathscr{B})} \log p(\mathscr{B}_i) \to H(\mathscr{B}), \tag{32}$$

where the convergence takes place with probability 1 for all $2^k$ subsets $\mathscr{B} \subseteq \{X^{(1)}, \ldots, X^{(k)}\}$.

**Definition.** The set $A_\epsilon^{(N)}$ of jointly typical $N$-sequences $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ with empirical entropies $\epsilon$-close to the true entropies is defined by [10]

$$A_\epsilon^{(N)}(X^{(1)}, \ldots, X^{(k)}) = A_\epsilon^{(N)} = \left\{ (\mathbf{x}_1, \ldots, \mathbf{x}_k) : \left| -\frac{1}{N} \log p(\boldsymbol{b}) - H(\mathscr{B}) \right| < \epsilon, \forall \mathscr{B} \subseteq \{X^{(1)}, \ldots, X^{(k)}\} \right\}.$$

Furthermore, let $A_\epsilon^{(N)}(\mathscr{B})$ denote the restriction of $A_\epsilon^{(N)}$ to the coordinates of $\mathscr{B}$. For low density codes we can write

$$A_\epsilon^{(N)} = \left\{ (\mathbf{x}_1, \ldots, \mathbf{x}_k) : \left| -\frac{1}{N'} \log p(\boldsymbol{b}) - H(\mathscr{B}) \right| < \epsilon, \forall \mathscr{B} \subseteq \{X^{(1)}, \ldots, X^{(k)}\} \right\}. \tag{33}$$

***Lemma.*** *For any $\epsilon > 0$, for sufficiently large $N$, the probability of having jointly typical sequences is*

$$\Pr(A_\epsilon^{(N)}(\mathscr{B})) \geqslant 1 - \epsilon, \qquad \forall \mathscr{B} \subseteq \{X^{(1)}, \ldots, X^{(k)}\} \tag{34}$$

*Proof.* This follows from the law of large numbers for random variables in the definition of $A_\epsilon^{(N)}(\mathscr{B})$. Calling (6), given $\epsilon > 0$, there exist $N_1$ such that for all $N > N_1$,

$$\Pr\left( \left| -\frac{1}{N'} \log p(\mathscr{B}_1, \mathscr{B}_2, \ldots, \mathscr{B}_N) - H(\mathscr{B}) \right| \geqslant \epsilon \right) < \epsilon, \qquad \forall \mathscr{B} \subseteq \{X^{(1)}, \ldots, X^{(k)}\} \tag{35}$$

$$\Rightarrow \Pr\left( \left| -\frac{1}{N'} \log p(\boldsymbol{b}) - H(\mathscr{B}) \right| \geqslant \epsilon \right) < \epsilon, \qquad \forall \mathscr{B} \subseteq \{X^{(1)}, \ldots, X^{(k)}\} \tag{36}$$

Thus, for sufficiently large $N$ for low density codes we will have,

$$\Pr(A_\epsilon^{(N)}(\mathscr{B})) \geqslant 1 - \epsilon, \qquad \forall \mathscr{B} \subseteq \{X^{(1)}, \ldots, X^{(k)}\} \tag{37}$$

Therefore, we could validate the joint typicality concept for low density signatures.

## References


1. K. Au, Zh. Liqing, H. Nikopour et al., Uplink contention based SCMA for 5G radio access, in *Globecom Workshops (GC Wkshps)*, 2014. doi:10.1109/GLOCOMW.2014.7063547
2. P. Baracca, S. Tomasin, L. Vangelista et al., Per sub-block equalization of very long OFDM blocks in mobile communications. IEEE Trans. Commun. (2011). doi:10.1109/TCOMM.2011.121410.090252
3. J.R. Barry, E.A. Lee, D.E. Messerschmit, *Digital Communication* (Kluwer, Dordrecht, 2004)
4. E. Biglieri, G. Caire, G. Taricco, Expurgating the union bound to error probability: a generalization of the Verdu-Shields theorem, in *Proceedings of 1997 IEEE International Symposium on Information Theory*, 1997. doi:10.1109/ISIT.1997.613310
5. E. Biglieri, J. Proakis, S. Shamai, Fading channels: information-theoretic and communications aspects. IEEE Trans. Inf. Theory (1998). doi:10.1109/18.720551
6. G. Caire, G. Taricco, E. Biglieri, Bit-interleaved coded modulation. IEEE Trans. Inf. Theory (1998). doi:10.1109/18.669123
7. R.S. Cheng, S. Verdu, Gaussian multi-access channels with ISI: capacity region and multi-user water-filling. IEEE Trans. Inf. Theory (1993). doi:10.1109/18.256487
8. A. Chouly, A. Brajal, S. Jourdan, Orthogonal multicarrier techniques applied to direct sequence spread spectrum CDMA systems, in *Proceedings of IEEE Global Telecommunications Conference, including a Communications Theory Mini-Conference. IEEE Technical Program Conference Record, Houston. GLOBECOM '93*, 1993. doi:10.1109/GLOCOM.1993.318364
9. J. Cioffi, *Course notes for Digital Communication: Signal Processing* (Stanford Bookstore Custom Publishing, Stanford, CA, 2007)
10. T.M. Cover, *Elements of Information Theory* (Wiley, New York, 2006)
11. T. Cover, R. McEliece, E. Posner, Asynchronous multiple-access channel capacity. IEEE Trans. Inf. Theory (1981). doi:10.1109/TIT.1981.1056382
12. P. Fan, L. Hao, Generalized orthogonal sequences and their applications in synchronous CDMA systems. IEICE Trans. Fundam. **E83–A**, 2054–2069 (2000)
13. K. Fazel, Performance of CDMA/OFDM for mobile communication system, in *Proceedings of 2nd International Conference on Universal Personal Communications: Gateway to the 21st Century*, 1993. doi:10.1109/ICUPC.1993.528524
14. J. Forney, G. Ungerboeck, Modulation and coding for linear Gaussian channels. IEEE Trans. Inf. Theory (1998). doi:10.1109/18.720542
15. C.H.F. Fung, W. Yu, T.J. Lim, Precoding for the multiantenna downlink: multiuser SNR gap and optimal user ordering. IEEE Trans. Commun. (2007). doi:10.1109/TCOMM.2006.885095
16. R. Gallager, An inequality on the capacity region of multi-access fading, in *Channels Communications and Cryptography—Two Sides of One Tapestry* (1994), pp. 129–139
17. A. Goldsmith, *Wireless Communication* (Cambridge University Press, Cambridge, 2005)
18. S. Hara, R. Prasad, Overview of multicarrier CDMA. IEEE Commun. Mag. (1997). doi:10.1109/ISSSTA.1996.563752
19. Ch. Hon-Fah, M. Motani, Capacity region of the asynchronous gaussian vector multiple-access channel. IEEE Trans. Inf. Theory (2013). doi:10.1109/TIT.2013.2272012
20. R. Hoshyar, F.P. Wathan, R. Tafazolli, Novel low-density signature for synchronous CDMA systems over AWGN channel. IEEE Trans. Signal Process. (2008). doi:10.1109/TSP.2007.909320
21. R. Hoshyar, R. Razavi, M. Al-Imari, LDS-OFDM an efficient multiple access technique, in *Proceedings of IEEE 71st Vehicular Technology Conference (VTC 2010-Spring)*, 2010. doi:10.1109/VETECS.2010.5493941
22. D. Li, A high spectrum efficient multiple access code. Communications (1999). APCC/OECC '99. doi:10.1109/APCC.1999.824954
23. L. Liu, L.K. Leung, L. Ping, Simple iterative chip-by-chip multiuser detection for CDMA systems, in *Proceedings of VTC 2003-Spring Vehicular Technology Conference. The 57th IEEE Semiannual*, 2003. doi:10.1109/VETECS.2003.1207209

24. L. Liu, J. Tong, L. Ping, Analysis and optimization of CDMA systems with chip-level interleavers. IEEE J. Sel. Areas Commun. (2006). doi:10.1109/JSAC.2005.858896
25. M.P. Lotter, P. Van Rooyen, An overview of space division multiple access techniques in cellular systems, in *Proceedings of South African Symposium Communications and Signal Processing COMSIG '98*, 1998. doi:10.1109/COMSIG.1998.736941
26. R.V. Nee, R. Prasad, *OFDM for Wireless Multimedia Communications* (Artech House, Boston, 2000)
27. H. Nikopour, E. Yi, A. Bayesteh et al., SCMA for downlink multiple access of 5G wireless networks, in *IEEE Global Communications Conference (GLOBECOM)* (2014). doi:10.1109/GLOCOM.2014.7037423
28. A. Persson, T. Ottosson, E. Strom, Time-frequency localized CDMA for downlink multi-carrier systems, in *Proceedings of IEEE Seventh International Spread Spectrum Techniques and Applications Symposium*, 2002. doi:10.1109/ISSSTA.2002.1049298
29. L. Ping, L. Liu, K. Wu, W.K. Leung, Interleave division multiple-access. IEEE Trans. Wireless Commun. (2006). doi:10.1109/TWC.2006.1618943
30. T.S. Rappaport, *Wireless Communications: Principles and Practice* (Prentice Hall, Englewood Cliffs, NJ, 2007)
31. R. Razavi, M. Al-Imari, M.A. Imran, On receiver design for uplink low density signature OFDM (LDS-OFDM). IEEE Trans. Commun. (2012). doi:10.1109/TCOMM.2012.082812.110284
32. Z. Rezki, M.S. Alouini, On the capacity of multiple access and broadcast fading channels with full channel state information at low SNR. IEEE Trans. Wireless Commun. (2014). doi:10.1109/TWC.2013.120113.130895
33. E. Schulz, Forward 2020 5G (2015). http://cwbackoffice.co.uk/Presentation/RTSS
34. S. Shamai, A.D. Wyner, Information-theoretic considerations for symmetric, cellular, multiple-access fading channels. ii. IEEE Trans. Inf. Theory (1997). doi:10.1109/18.641554
35. S. Stanczak, H. Boche, M. Haardt, Are LAS-codes a miracle? in *Proceedings of IEEE Global Telecommunications Conference GLOBECOM '01*, 2001. doi:10.1109/GLOCOM.2001.965185
36. B. Suard, G. Xu, H. Liu, T. Kailath, Uplink channel capacity of space-division-multiple-access schemes. IEEE Trans. Inf. Theory (1998). doi:10.1109/18.681322
37. M. Taherzadeh, H. Nikopour, A. Bayesteh, H. Baligh, SCMA codebook design, in *IEEE 80th Vehicular Technology Conference (VTC Fall)*, 2014. doi:10.1109/VTCFall.2014.6966170
38. D. Toumpakaris, J. Lee, The gap approximation for Gaussian multiple access channels, in *Proceedings of IEEE Global Telecommunications Conference GLOBECOM*, 2009. doi:10.1109/GLOCOM.2009.5425889
39. D.N.C. Tse, S.V. Hanly, Multiaccess fading channels. i. Polymatroid structure, optimal resource allocation and throughput capacities. IEEE Trans. Inf. Theory (1998). doi:10.1109/18.737513
40. A.M. Tulino, L. Li, S. Verdu, Spectral efficiency of multicarrier CDMA. IEEE Trans. Inf. Theory (2005). doi:10.1109/TIT.2004.840875
41. S. Verdu, *Multiuser Detection* (Cambridge University Press, Cambridge, 1998)
42. F. Wathan, R. Hoshyar, R. Tafazolli, Dynamic grouped chip-level iterated multiuser detection based on Gaussian forcing technique. IEEE Commun. Lett. (2008). doi:10.1109/LCOMM.2008.071931
43. H. Wei, L. Hanzo, On the uplink performance of LAS-CDMA. IEEE Trans. Wireless Commun. (2006). doi:10.1109/TWC.2006.1633372
44. C. Xiaodong, Z. Shengli, G.B. Giannakis, Group-orthogonal multicarrier CDMA. IEEE Trans. Commun. (2004). doi:10.1109/TCOMM.2003.822174
45. N. Yee, J.P. Linnartz, BER of multi-carrier CDMA in an indoor Rician fading channel, in *Proceedings of Conference Signals, Systems and Computers Record of The Twenty-Seventh Asilomar Conference*, 1993. doi:10.1109/ACSSC.1993.342548

# New Multicarrier Modulations for 5G

**Davide Mattera, Mario Tanda, and Maurice Bellanger**

**Abstract** Recent advances in filter bank multicarrier (FBMC) techniques are presented and compared with the conventional cyclic prefix (CP)-OFDM approach, in the context of 5G. After a brief description of some adaptations of CP-OFDM, FBMC combined with offset-QAM is considered, pointing out the crucial issue of subchannel equalization to compensate for the absence of cyclic prefix. Implementation has to be carried out in the frequency domain, in order to reach high performance without further increasing latency and allow for per user equalization and carrier frequency offset compensation. However, due to the delay objectives and the computational complexity, the prototype filter length, and overlapping factor, must be small, which limits the performance. Then, an alternative approach is proposed, FBMC combined with pulse amplitude modulation (PAM). Like OFDM, FBMC-PAM possesses the perfect reconstruction property and, equipped with the sine prototype filter, it yields good performance with the overlapping factor $K = 2$, which makes it an attractive option whenever asynchronous access and high level of out-of-band rejection are required. Finally, the case of nonoverlapping emitted symbols is considered and a CP-less OFDM scheme with frequency domain equalizer in the receiver is included in the performance comparison.

## 1 Introduction

Many specific requirements for 5G are linked to the multi-user context, such as asynchronous access, coexistence issues and per-user channel equalization and carrier frequency offset compensation. In addition, latency has to be kept to a minimum. At the system level, it is necessary to maintain coherence between the

D. Mattera • M. Tanda (✉)
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università degli Studi di Napoli Federico II, Via Claudio, 21, 80125 Napoli, Italy
e-mail: mattera@unina.it; tanda@unina.it

M. Bellanger
CNAM-Electronique, 292 rue Saint-Martin, 75141, Paris Cedex 03, France
e-mail: maurice.bellanger@cnam.fr



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_7

performance of the signal processing techniques and the existing or foreseeable capabilities of radiofrequency interfaces. Obviously, these requirements have a critical impact on the selection of the physical layer techniques.

Multicarrier modulations have proved successful in cable and wireless communication systems because of their spectral efficiency and their operational flexibility and they are strong candidates for the physical layer in 5G and beyond. So far, the most widely deployed technique is cyclic-prefix orthogonal frequency division multiplexing (CP-OFDM), which is simple in concept and easy to implement. Basically, it consists of an inverse fast Fourier transform (IFFT) in the transmitter and an FFT in the receiver, whose size $M$ defines the number of subcarriers in the system. When the blocks of $M$ time samples generated by the IFFT are fed to the FFT, the subcarriers are orthogonal and the input data samples are perfectly reconstructed. In the presence of channel distortion, a cyclic prefix has to be introduced at the transmitter side so that orthogonality can be restored at the receiver side, through an appropriate time shift, under the condition that the length of the cyclic prefix exceeds the duration of the channel impulse response. Then, the demodulation process is completed by per subcarrier frequency domain equalization, which is the key to the spectral efficiency of the scheme.

The CP-OFDM approach is well suited to the transmission of single user over a bandwidth interval much larger than the channel coherence bandwidth. In the presence of several simultaneous users exploiting the same OFDM frame, it is necessary to make sure that time and frequency alignment is obtained with sufficient accuracy, so that the orthogonality of the signals transmitted by all the users is secured. The cost of such an alignment depends on the number of users involved in the process. In case of independent adjacent users, the global spectral efficiency is impacted by the filtering effect of the multicarrier modulations involved. The basic OFDM scheme offers limited spectral separation and a number of requirements defined by the use case families considered for the definition of the 5G standards [6] tend to make the approach unfit to the considered scenarios. Specifically, the limited filtering effect of OFDM leads to losses in spectral efficiency that can be severe in case of asynchronous access and coexistence issues, as well as in the presence of fragmented spectrum. Then, in order to keep spectral efficiency in such conditions, the filter bank multicarrier (FBMC) modulation has been proposed as an alternative to OFDM. In a filter bank, a low-pass prototype filter with $KM$ coefficients is shifted on the frequency axis by integer multiples of the subcarrier spacing, leading to a set of $M$ subchannels. A key difference with OFDM is that $K$ emitted multicarrier symbols overlap, and $K$, the so-called overlapping factor, determines the filter attenuation characteristics and, thus, the spectral separation between adjacent users. With this approach, the cyclic-prefix concept can be abandoned and the channel distortion must be mitigated by subchannel equalization. In fact, FBMC without multitap or frequency-domain subchannel equalization can be considered similar to OFDM with a very short cyclic prefix.

Regarding implementation, two options exist, as shown in Fig. 1, for the receiver. In the time domain scheme, a set of $M$ digital filters, having $K$ coefficients each and called polyphase network (PPN), is cascaded with the FFT. In the frequency domain

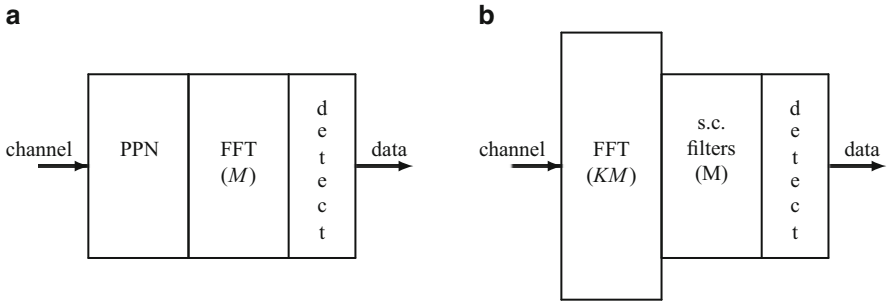


**Fig. 1** Structures of filter banks in FBMC receivers: (**a**) PPN-FFT scheme, (**b**) Frequency domain scheme

scheme, an FFT of size $KM$ is followed by a set of per-subchannel decimation filters whose coefficients are derived from the time domain coefficients through discrete Fourier transform. Of course, the frequency domain scheme requires more computations and it is practical only for small overlapping factors, typically 2 or 3. However, it has two key benefits

- channel equalization is performed per subchannel at the FFT output, which leads to high performance with no additional delay, contrary to the PPN approach which requires time domain equalization with the corresponding delay. The same applies to per-user carrier frequency offset (CFO) compensation.
- although it is performed per subchannel, the equalization at the FFT output before filtering is equivalent to the global equalization of the received signal, which means that perfect reconstruction techniques can be used and orthogonality is possible, as in OFDM.

Now, to complete the system description, data processing is considered. In OFDM systems, the data sequence is mapped into quadrature amplitude modulation (QAM) samples at the input of the transmitter IFFT. This mapping cannot be applied to FBMC, because of the filter overlapping.

The signals transmitted on the subcarriers $k-1$ and $k+1$ generate interference in the subcarrier $k$ and there are two approaches to achieve full rate transmission in spite of this interference

- Offset-QAM modulation leading to FBMC-OQAM systems. The principle relies on the impulse responses of interference filters that cross the zero axis at multiples of the symbol period $M$, but with a shift of $M/2$ between the real part and the imaginary part. Then, full rate is obtained if the symbol rate is doubled and real and imaginary data samples alternate in the time domain and in the frequency domain;
- Pulse amplitude modulation (PAM) leading to FBMC-PAM systems. The phase shift $\pi/2$ is applied to the subcarriers of adjacent sub-channels and the transmitter input is a real sequence of data samples.

The poor spectral containment and the synchronization requirements in multiple access are the weak points of the classical OFDM system that make it unfit to the next generation communication systems; its simpler evolutions improve on such issues but introduce a further loss in terms of spectral efficiency. The FBMC-OQAM system achieves significant improvements over classical OFDM system in terms of spectral containment and robustness to the imperfections of the timing and CFO estimation procedures with the same spectral efficiency; such improvements are more significant for larger $K$. Its main drawbacks are the increase of the computational complexity and the inefficiency in short-burst transmissions, which is also an important requirement for next generation systems; both the complexity and the inefficiency increase with $K$. The FBMC-PAM, instead, is superior to the classical OFDM system in terms of spectral containment and synchronization requirement but it also exhibits smaller complexity than FBMC- OQAM; moreover, the inefficiency of FBMC-PAM in short-burst transmissions is marginal. Therefore, in many use cases, FBMC-PAM represents a good compromise; only in extreme scenarios, characterized by severe requirements in terms of spectral containment and spectrum fragmentation, FBMC-OQAM with $K = 3$ (or also with $K = 4$) may become necessary.

Following this brief overview of the multicarrier transmission field, the rest of the chapter is organized as follows. In Sect. 2, the classical OFDM system is presented and some of the evolutions which have been proposed to improve filtering are discussed. Section 3 is dedicated to the FBMC-OQAM system, with emphasis on the frequency domain implementation option and the performance of channel equalization and CFO compensation. Then, in Sect. 4, the FBMC-PAM system is considered in combination with perfect reconstruction filter design, so that the orthogonality property is obtained, as in OFDM. Finally, a CP-less OFDM scheme with frequency domain equalization in the receiver is discussed in Sect. 5 and, in Sect. 6, the considered systems are compared by simulation experiments.

## 2 The Classical OFDM and Its Evolutions

For the sake of clarity, we shortly resume the characteristics of the standard CP-OFDM; the classical rectangular pulse $g^{(\mathrm{O})}[n]$ is used to transmit a complex-valued symbol:

$$g^{(\mathrm{O})}[n] = u_{M+M_p}[n] \triangleq \begin{cases} 1 \; n \in \{0, 1, \ldots, M + M_p - 1\} \\ 0 \; n \notin \{0, 1, \ldots, M + M_p - 1\} \end{cases}. \tag{1}$$

The pulse transmitted in the first symbol period $\{0, 1, \ldots, M + M_p - 1\}$ on the $k$th subcarrier is written as

$$g_k^{(\mathrm{O})}[n] \triangleq g^{(\mathrm{O})}[n] e^{j2\pi \frac{k}{M}(n - M_p)}. \tag{2}$$

The problem of intersymbol interference is solved by using a symbol period $M+M_p$ larger than the pulse duration $M$; the length of the cyclic prefix $M_p$ has to be larger than the channel delay spread $M_h$ in order to eliminate the interference effects from previous symbol intervals on the current one; for such a reason, the first samples of the received signal, which are affected by pulse transmitting the previous symbols, are not utilized at the receiver. The interferences among the signals transmitted on the different subcarriers are determined by the coefficients:

$$I_k^{(\mathrm{O})}[n_k,\varepsilon_k] \triangleq \left|\sum_{n=0}^{M-1} g^{(\mathrm{O})}[n+M_p-n_k]\left(g^{(\mathrm{O})}[n+M_p]\right)^* e^{j2\pi\frac{k+\varepsilon_k}{M}n}\right|$$
$$= \left|\int_{-1/2}^{1/2} U_M(F)U_M^*\left(F+\frac{k+\varepsilon_k}{M}\right)e^{-j2\pi\overline{n}_k F}\mathrm{d}F\right| \tag{3}$$

where $n_k=\tilde{n}_k+n_0$ and $\varepsilon_k=\tilde{\varepsilon}_k+\epsilon_0$ with $\tilde{n}_k$ and $\tilde{\varepsilon}_k$ denoting the relative timing offset and the relative normalized CFO, respectively, of the user transmitting at a distance of $k$ subcarriers from that of interest and with $n_0$ and $\epsilon_0$ denoting the estimation error of the timing offset and of the normalized CFO, respectively, of the user of interest. Moreover, for $|n_k|\le M/2$, we have defined

$$\overline{n}_k \triangleq \begin{cases} 0 & 0\le n_k\le M_p \\ n_k-M_p & n_k>M_p \\ n_k & n_k<0 \end{cases} \tag{4}$$

and

$$U_M(F) \triangleq \mathscr{F}\{u_M[n]\} = e^{-j\pi F(M-1)}\frac{\sin(\pi FM)}{\sin(\pi F)}. \tag{5}$$

We here consider the case where the subcarrier signal, which is transmitted on $k$th subcarrier adjacent to that of interest, is possibly used by a different user (in such a case $\tilde{n}_k\neq 0$ unless a proper timing alignment procedure is employed). The well-known property $I_k^{(\mathrm{O})}[0,0]=0\ \forall k\neq 0$ guarantees that there is no interference between two subcarriers employed by the same user (i.e., $\tilde{n}_k=\tilde{\varepsilon}_k=0$) provided that $M_h\le M_p$ and the timing offset and CFO of the user of interest can be perfectly estimated (i.e., $n_0=\varepsilon_0=0$). This property, which holds for $0\le n_k\le M_p$ (since $\overline{n}_k=0$ for such values of $n_k$), has allowed the great diffusion of CP-OFDM in classical downlink scenarios. At the same time, the fact that the same property does not hold when $\overline{n}_k\neq 0$ or $\varepsilon_k\neq 0$ represents one of the main drawbacks that complicated the use of CP-OFDM in the uplink scenarios.

In order to reduce $I_k^{(\mathrm{O})}[n_k,\varepsilon_k]$, the spectral behavior of the chosen pulse has to be modified. In fact, (3) clearly shows that, for $|k|\ge 2$ (i.e., for all subcarriers except the two adjacent ones), the first lobes of the two pulses are separated. Such a separation would imply that $I_k^{(\mathrm{O})}[n_k,\varepsilon_k]$ is negligible also when $(\overline{n}_k,\varepsilon_k)\neq(0,0)$

provided that the secondary lobes of the chosen pulse $U_M(F)$ were sufficiently small. In such a case, the relative timing offset $n_k$ and the relative normalized CFO $\varepsilon_k$ could be also very large without generating significant interferences from non-adjacent subcarriers. This would imply that the devices of the radio access network would not need to be precisely synchronized in time and frequency with each other as it is currently required in the classical OFDM systems. Such a requirement will tend to become more cumbersome (and, in perspective, also unfeasible) as the support of multiple traffic types, which is going to be required by the wireless radio interface of the next 5G standard, will imply the need to maintain synchronized a massive number of low-rate machine-type communication devices, along with the usual high-rate hand-held devices [19, 20].

The main disadvantage of the classical OFDM system lies, therefore, in the poor spectral containment of $U_M(F)$ outside its first lobe, i.e., outside the bandwidth $(-\frac{1}{M}, \frac{1}{M})$. The same spectral behavior of $U_M(F)$ is responsible for the other limit of OFDM: the difficulty to efficiently transmit over a small bandwidth fragment. In fact, the power spectral density (PSD) $P_{s_k}^{(\mathrm{O})}(F)$ of the signal $s_k^{(\mathrm{O})}[n]$ transmitted on the $k$th subcarrier

$$s_k^{(\mathrm{O})}[n] \triangleq \sqrt{\frac{E_s}{M+M_p}} \sum_{i=-\infty}^{+\infty} c_k[i] g_k^{(\mathrm{O})}[n - i\left(M + M_p\right)] \tag{6}$$

for a complex-valued symbol sequence $c_k[i]$ satisfying the condition

$$E\left[c_k[i] c_k^*[i-m]\right] = \delta[m], \tag{7}$$

where $\delta[\cdot]$ is the Kronecker delta, can be written as

$$P_{s_k}^{(\mathrm{O})}(F) = \frac{E_s}{\left(M+M_p\right)^2} \left| U_{M+M_p}\left(F - \frac{k}{M}\right)\right|^2 = E_s \frac{\sin^2\left(\pi\left(F - \frac{k}{M}\right)\left(M+M_p\right)\right)}{\left(M+M_p\right)^2 \sin^2\left(\pi\left(F - \frac{k}{M}\right)\right)} \tag{8}$$

and, assuming the independence of the symbol sequences used on the different subcarriers, the PSD of the signal transmitted by each user can be written as $\sum_{k\in\mathscr{A}} P_{s_k}^{(\mathrm{O})}(F)$ where $\mathscr{A}$ is the set of subcarriers exploited by the considered user. Here we have denoted with $E_s$ the energy transmitted for each complex-valued symbol.

Let us show in detail with an example the problem created by the poor spectral containment of the pulse used in CP-OFDM in the transmission over a bandwidth fragment. Assume that the bandwidth available for transmission is the numerical bandwidth $[0.2, 0.4]$ and let us choose the following parameters $M = 256$ and $M_p = M/8$; if there were no constraint on the PSD outside the available bandwidth, the set $\{53, 54, \ldots, 101\}$ of subcarriers could be used by the transmitter and the resulting PSD is depicted in Fig. 2. However, if we introduce the constraint that the PSD outside the chosen bandwidth has to be $-30$ dB below the maximum value of the PSD, which is a realistic constraint when we resort to a fragmented usage of the spectrum, we have to note that the standard CP-OFDM cannot be employed.

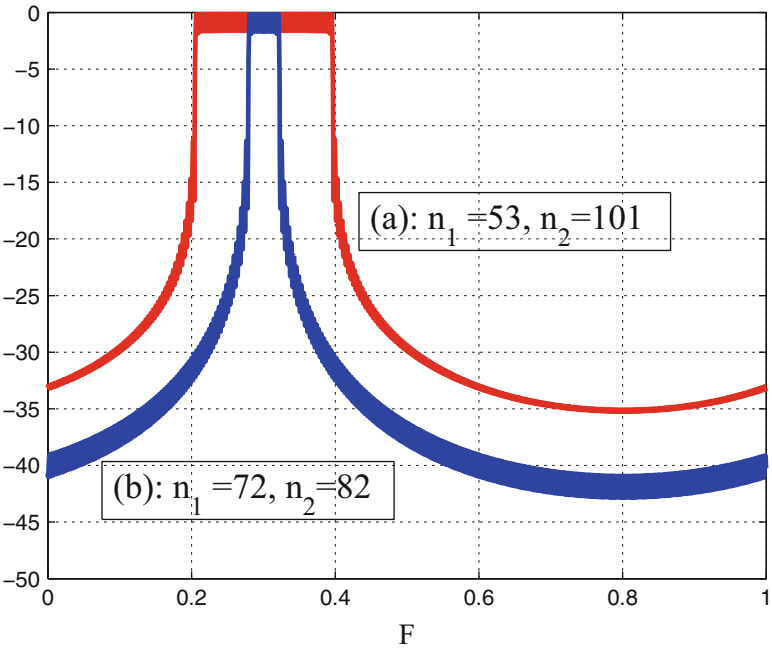


**Fig. 2** The PSD of the transmitted signal, defined as $\frac{1}{E_s}\sum_{k=n_1}^{n_2} P_{s_k}^{(O)}(F)$, when *(a)* all the subcarriers inside the bandwidth fragment $(0.2, 0.4)$ are employed and *(b)* when the minimum number of guard subcarriers needed to satisfy the constraint on PSD outside the considered fragment at $-30$ dB are used

The first basic line to circumvent the problem of spectral containment is the well-known practice of using the null guard subcarriers. With reference to the previous example, we can transmit only on the subcarriers in the following set $\{72, 73, \ldots, 82\}$; the resulting PSD is also reported in Fig. 2. If we had specified the slightly stronger constraint of $-35$ dB outside the chosen bandwidth, the set of the subcarriers that could be used would be further reduced to the following set: $\{78, 79\}$; more selective constraints cannot be satisfied by CP-OFDM by resorting to the null subcarriers. It is clear that the strong reduction of the number of active subcarriers is caused by the slow decrease of the pulse in (1); we pass from 49 to 11 (or possibly to 2) subcarriers. Therefore, using the null subcarriers can solve the problem of spectral containment, but only if the constraint introduced on the out-of-band PSD is not more severe than 35–40 dB; in any case, also if the problem can be solved, this implies a significant loss in the achieved bit-rate, also when the CP-OFDM has to be employed in the downlink scenario on a large fragment such as the considered numerical bandwidth $(0.2, 0.4)$ including in this example $N_f = 49$ subcarriers. In some scenarios, also smaller values of $N_f$ could be available, making the spectral inefficiency of the classical OFDM system using guard subcarriers also more severe.

## 2.1 The Time Separation Approach for Evolving Classical OFDM

With regard to the theoretical proposals of a multicarrier system for data transmission, the classical OFDM technique is not the oldest one [7]. The main difficulties that delayed the adoption of the multicarrier techniques mainly lie in their computational complexity; the first version commonly adopted in various standards, the classical OFDM system, was therefore that minimizing the computational complexity. The overall network constraints were designed in order to circumvent the main drawback of the OFDM system, the poor spectral containment. However, many researchers have tried to propose different variants of the classical OFDM with smaller out-of-band emissions. In the last years, many contributions appeared in the literature (e.g., see [2, 3] and references therein) are proposing evolved multicarrier systems since the overall network scenario is making appear much probable that, for the next 5G standard, the classical OFDM system may be abandoned, at least in some use cases.

Solving the problem of the spectral containment requires to adopt a basic pulse whose spectral secondary lobes can be neglected. As it is well-known, this is however associated with the increase of the width of the main lobe. However, the subcarrier bandwidth is usually imposed by external constraints; therefore, adopting the same subcarrier bandwidth, the new basic pulse needs to become longer in time; let us denote with $(M + M_p)/\rho$ the length of the new pulse with the amplitude of the secondary lobes sufficiently small, i.e., let us denote with $\rho$ the loss of spectral efficiency due to the increased pulse length.

This imposes a basic choice whether to maintain or to abandon the approach followed in classical OFDM about the issue of intersymbol interference. A first line of evolution, named here *time separation*, motivated by the greater compatibility with the current technology, proposes to limit the changes to the implementation platform, and in particular to the equalization stage. Therefore, two time-shifted versions of the basic pulse are still maintained separated in the time domain; such a separation is imposed at the transmitter output and is maintained also at the receiver input by resorting to an adequate cyclic prefix. In other terms, the improvement of the spectral containment is mainly paid in terms of the bit-rate achieved on the given bandwidth, i.e., in terms of spectral efficiency, while the computational complexity of the classic OFDM transceiver is maintained.

The line of *time-separation* includes the so-called *pulse-shaping* OFDM where the rectangular pulse in (1) is replaced by a properly chosen pulse shape; the choice of the new pulse has to take into account both the issue of spectral containment and that of the intercarrier interference (ICI). In particular, one of the first proposals considered in [8] for pulse-shaping OFDM is the well-known Hanning pulse. An important choice in the category of pulse-shaping is given by the time-window approach where the central rectangular shape of the pulse in (1) is maintained but a tail is appended on the two sides of the pulse in order to improve the spectral containment; typically, the tail is given a cosine shape, so obtaining that the pulse is

shaped as a raised-cosine (RC) in the time domain:

$$g^{(RC)}[n] \triangleq \begin{cases} \frac{1-\cos\left(\pi \frac{n+1}{M_r+1}\right)}{2} & 0 \le n \le M_r - 1 \\ 1 & M_r \le n \le M_r + M + M_p - 1 \\ \frac{1+\cos\left(\pi \frac{n+1-M-M_p-M_r}{M_r+1}\right)}{2} & M + M_p + M_r \le n \le M + M_p + 2M_r - 1 \\ 0 & n \notin \{0, 1, \ldots, M + M_p + 2M_r - 1\} \end{cases} \tag{9}$$

where $M_r = (M + M_p)(\frac{1}{\rho} - 1)$. Such a choice provides three important advantages: the first advantage is that the receiver adaptation to the new pulse-shape can be avoided by selecting at the receiver only the central rectangular part of the transmitted raised-cosine pulse. The second advantage, connected to the first one, is that the tail on the right side of the previous pulse can be time overlapped to that on the left side of the next pulse, consequently reducing the loss in spectral efficiency. The third advantage is given by the fact that using only the central rectangular part of the pulse also guarantees the absence of ICI.

The principal analysis about such a line of evolution for OFDM concerns the quantitative values of the loss of spectral efficiency due to the need to improve the spectral containment. It can be written as follows:

$$\rho \left(1 - \frac{N_g}{N_f}\right) \left(\frac{M}{M + M_p}\right) \tag{10}$$

where we have denoted with $\rho$ the loss of spectral efficiency due to the pulse length increase, with $N_g$ the number of guard subcarriers and with $N_f$ the maximum number of subcarriers that could be employed in the available spectral fragment. Note that the last two terms are already present in the classical OFDM while the loss factor $\rho$, which is unit in classical OFDM, is peculiar of the considered line of the *time separation*.

In order to quantify such a loss, in Figs. 3 and 4 the PSD of the transmitted signal at the edge of the spectral bandwidth is reported; the various curves correspond to different values of $\rho$, which determines the pulse length; they are plotted by assuming that the PSD of the signal transmitted on each subcarrier is unit at the subcarrier center. The figures refer to the two usually considered pulses in pulse shaping [18] (the Hanning pulse and the RC one) for $M+M_p = 256(1+1/8) = 288$.

With reference to the Hanning pulse, we obtain that the achieved out-of-band attenuation depends on the number of guard subcarriers and on the loss of spectral efficiency. Since also the use of guard subcarriers causes a loss of spectral efficiency, the choice of the pulse length should take into account the two factors $\rho$ and $(1 - N_g/N_f)$. For example, if the out-of-band attenuation is set to $-40$ dB, we cannot

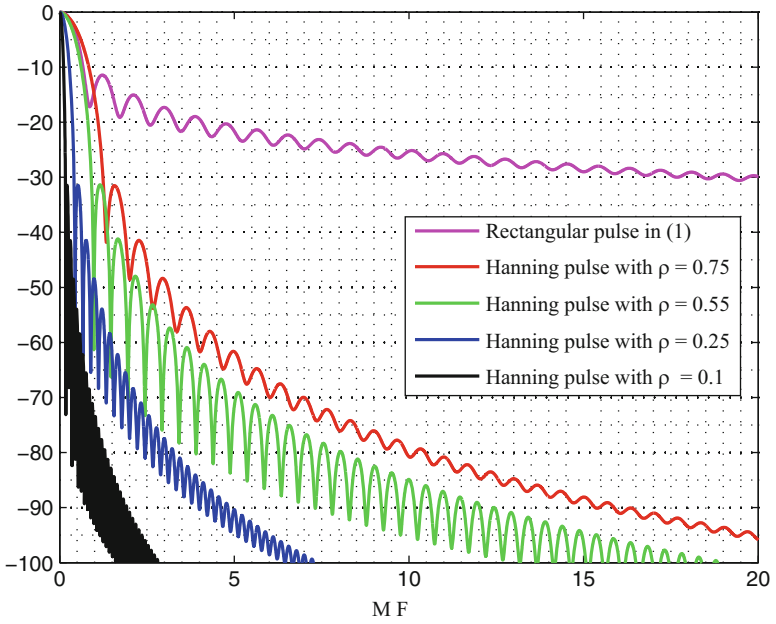


**Fig. 3** PSD of the transmitted signal at the edge of the spectral bandwidth with the Hanning pulse for $\rho \in \{0.75, 0.55, 0.25, 0.1\}$ and with the rectangular pulse in (1)

achieve it with $\rho = 1$, but we can achieve it with $\rho = 0.75$ and 2 guard subcarriers, or with $\rho = 0.55$ and $N_g = 1$ or with $\rho = 0.1$ and $N_g = 0$. It is evident that often the first choice is the most convenient.

When we have to achieve important out-of-band attenuation, it can be useful to resort to longer pulses obtained with smaller values of $\rho$. For example, if we have to achieve an out-of-band attenuation of −80 dB, 3 guard subcarriers are sufficient with $\rho = 0.25$ while 10 subcarriers are needed when $\rho = 0.75$. The overall efficiencies in the two cases are

$$(1 - 6/N_f)0.25 \qquad (1 - 20/N_f)0.75 \tag{11}$$

where $N_f$ is the number of subcarriers in the assigned bandwidth fragment. Therefore, the best choice between the two considered alternatives depends on the condition $N_f > 27$, and, obviously for $7 \leq N_f \leq 27$, only a single choice would be possible. Similar reasoning applies also with the RC pulse, whose spectral behavior is reported in Fig. 4 again for $M + M_p = 256(1 + 1/8) = 288$, and the results can be also worst. To achieve an out-of-band attenuation of −80 dB, 2 guard subcarriers are needed for $\rho = 0.25$ while 16 subcarriers are needed for $\rho = 0.75$ and the new condition becomes $N_f > 46$. To achieve, instead, an out-of-band attenuation of −40 dB, no guard subcarrier is needed for $\rho = 0.25$ and 3 guard subcarriers are needed for $\rho = 0.75$, the relative condition becomes $N_f > 9$.

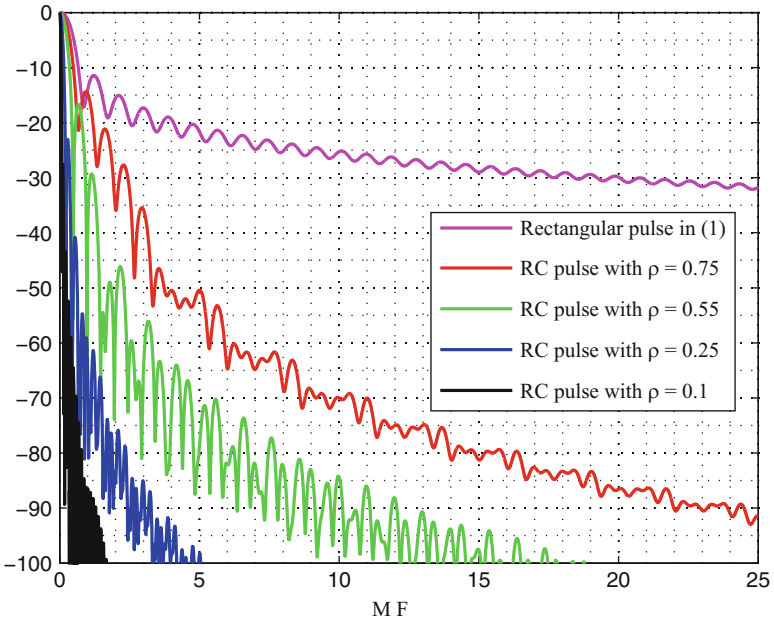


**Fig. 4** PSD of the transmitted signal at the edge of the spectral bandwidth with the RC pulse in (9) for $\rho \in \{1, 0.75, 0.55, 0.25, 0.1\}$ and with the rectangular pulse in (1)

We can deduce from the above considered example the following general result about the analysis of the pulse shaping approach. When the fragment is composed by a small number of subcarriers and a strong constraint is imposed on the out-of-band attenuation, the best choice is obtained with small values of $\rho$, and therefore the pulse-shaped OFDM achieves the spectral containment by accepting a significant loss in terms of overall spectral efficiency. Otherwise, when the required out-of-band attenuation is moderate or low, and the fragment under consideration is composed of a large number of subcarriers, the pulse-shaping OFDM can use values of $\rho$ around 0.75 and, therefore, pulse-shaped OFDM suffers from a moderate loss in terms of spectral efficiency for achieving its target of spectral containment. Both factors affect the loss of spectral efficiency: the required out-of-band attenuation and the size of the bandwidth fragment (in terms of number of subcarriers). It is important to note that the imperfections of the analog amplifiers can make appear not interesting the capability to achieve severe values of out-of-band attenuation; on the other hand, the possibility to allocate small bandwidth fragments to the machine-type communications can make appear interesting the case where $N_f$ is particularly small.

The network scenarios considered in 5*G* standardization include the extreme cases. There are use cases where the loss of spectral efficiency related to the use, with minor changes, of the classical OFDM transceiver can be considered contained and therefore still acceptable, in the meantime that the technology improvements

become available. Note however that the time will render less and less attractive solutions that have their strong advantage in the limited computational complexity; this is especially true as the available bandwidth is going to become more important. Such use cases concern the availability of bandwidth fragments assigned to the transceiver whose size is relatively large and the spectral containment required is moderate. On the other hand, there are use cases where such conditions are not satisfied and to maintain the line of *time separation* requires a significant reduction of the bit-rate, specified by the parameter $\rho$.

The same reasoning applies to the other approaches that fall within the line of *time separation* to simplify the standard approach to equalization. In particular, filtered OFDM [1] follows the same general reasoning already developed for pulse-shaping OFDM and, therefore, accepts similar losses in spectral efficiency in order to simply improve the standard OFDM system.

In such an approach, in fact, the final filter with $M_f + 1$ taps operates on the sum of the signals transmitted on the different subcarriers by the same user, which is equivalent to use filtered versions of the rectangular pulse in (1). There are two main differences:

- the first one lies in the fact that each subcarrier signal is constructed with a different equivalent pulse;
- the second difference lies in the fact that the transmitted signal is constructed by properly filtering the sum of the signals generated using the same standard pulse on each subcarrier; this implies a complexity cost associated with such a final filtering.

The length of the cyclic prefix introduced before the final filtering has to be increased from $M_p$ to $M_f + M_p$, i.e., the cyclic prefix must be increased by $M_f$ in order to maintain the line of the time separation. The choice $M_f = (M + M_p)\left(\frac{1}{\rho} - 1\right)$ implies that the resulting spectral efficiency can be still expressed as in (10) provided that the first $M_f$ samples of a time interval overlaps with the final $M_f$ samples of the previous interval. By selecting (for the FFT processing) the central block of $M$ samples, not affected by the filtering effects of the transmission filter and of the multipath channel, the orthogonality of the different pulses is preserved and, therefore, the classical OFDM receiver can be employed also in filtered OFDM. The fraction of wasted energy is equal to $\frac{M_f + M_p}{M + M_f + M_p}$; the effects of violating the usual condition $M_p \geq M_h$ can be alleviated by the fact that the final $M_f$ samples of the pulse transmitted in the previous time interval (overlapped on the first $M_f$ samples of the pulse transmitted in the present time interval) exhibits smaller power in comparison with that of the useful samples; following such a perspective, schemes resorting to the extreme choice $M_p = 0$ have been proposed (e.g., [17]) in the literature in order to reduce the loss in spectral efficiency; this implies the abandon of the line of *time separation* and is discussed in the next section. As in pulse-shaping OFDM, also in filtered OFDM the choice of $\rho$ has to trade-off (achieving similar results) the reduction of the spectral efficiency in (10) with the increase of the filter length $M_f$, and therefore with the consequent advantage in terms of spectral containment.

To complete the scenario followed in the line of *time separation*, we have to finally discuss the approach to achieving the spectral containment that makes use of the coefficients transmitted on the guard subcarriers (consequently, renamed cancellation subcarriers) in each symbol period as a function of the data symbol transmitted on the active subcarriers. Such coefficients are usually set to zero; different algorithms have been proposed to achieve the spectral containment. The target of large spectral efficiency, as well as the power lost on the guard subcarriers, should be limited by the small percentage of such subcarriers but this also limits the degrees of freedom required to achieve spectral containment on a large bandwidth.

The use of cancellation subcarriers does not appear useful to obtain significant increase of the value of $\rho$ achieved through pulse-shaping with *time separation* because the obtained spectral containment is often important only in limited bandwidths outside the useful fragment.

The use of cancellation subcarriers can be interesting when we have already accepted to operate along the line of *time separation* with reasonable values of $\rho$; in such a case, we can reduce the number of guard and cancellation subcarriers, further slightly improving the overall spectral efficiency, by paying a limited price in terms of computational complexity and wasted transmitted power (see, for example, [11]).

## 3 Frequency Domain FBMC-OQAM

The choice of a basic pulse longer than $M$ is necessary for solving the problem of the spectral containment, as already discussed in Sect. 2.1. The requirement to not pay such a solution in terms of spectral efficiency imposes to abandon the line of the *time separation* and, therefore, to admit the time overlapping of the signals associated with the successive symbols transmitted on the same subcarrier (as well as with those transmitted on the other subcarriers). The resulting transceiver design has therefore to consider not only the problem of ICI, as already done in the classical OFDM and in its variants introduced in Sect. 2, but also the problem of ISI, both on the flat and on the dispersive channel.

On a flat channel the two problems are solved by a careful transceiver design; the transceiver has also to be adapted in order to operate on a dispersive channel and in presence of timing and carrier offsets of the signal received by different transmitters. We first discuss how to design the transceiver on a flat channel, then we discuss its adaptation to the channel imperfections through proper procedures for equalization and synchronization.

### *3.1 Transceiver Design on a Flat Channel*

The length of the basic pulse, usually called the prototype filter, is denoted with $KM$ where $K$ is the overlap parameter; we do not use $\rho$ to denote it, as done in Sect. 2.1, because the increase of pulse length does not affect the spectral efficiency.

Once abandoned the line of the *time separation*, also the cyclic prefix can be avoided; however, the pulse length $KM$ should be roughly dimensioned as the length $(M + M_p)/\rho$ of the pulses considered in Sect. 2.1 in order to obtain the same level of spectral containment.

On time-dispersive channels, we aim to solve the problem of the spectral containment without any loss in terms of spectral efficiency; we need to use the designed prototype filter with an intercarrier spacing of $1/M$, which generates interferences between adjacent subcarriers. The presence of non-negligible secondary lobes in the spectral shape $U_M(F)$ of the rectangular pulse (1) used in the classical OFDM system is crucial to guarantee the orthogonality of adjacent subcarriers. Figure 5 shows how the secondary lobes are crucial to make null the real part of the integral in (3) for $\overline{n}_k = \varepsilon_k = 0$.

When the secondary lobes can be neglected, the orthogonality does not hold any more. Let us call $g[n]$ the prototype filter designed in the interval $I_K \triangleq \{-(KM/2-1), \ldots, -1, 0, 1, \ldots, KM/2-1\}$ (i.e., $g[n] = 0$ for $n \notin I_K$) in order to achieve the required spectral containment and assume that the design has imposed that the filter be real valued and symmetric (i.e, with $g[n] = g[-n]$). Consequently, its Fourier transform $G(F)$ is real and even. The orthogonality between adjacent subcarriers requires to consider the product $G^{\text{prod}}(F) \triangleq G(F)G(F - 1/M)$. If the

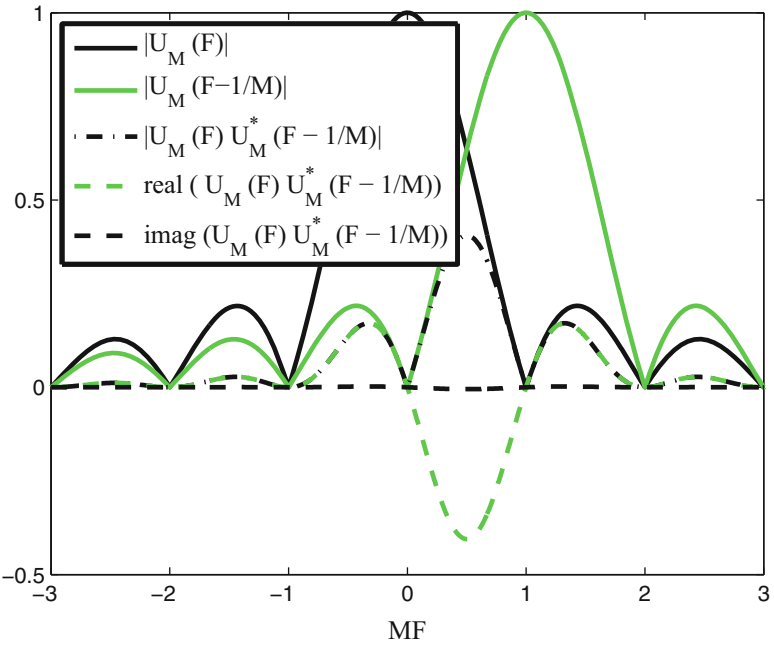


**Fig. 5** A plot of $|U_M(F)|$, $|U_M(F-1/M)|$, $|U_M(F)U_M^*(F-1/M)|$, $\Re\{U_M(F)U_M^*(F-1/M)\}$, and $\Im\{U_M(F)U_M^*(F-1/M)\}$ versus $MF$ for $M = 256$

secondary lobes are negligible, the energy of $G^{\text{prod}}(F)$ is mainly concentrated in the interval $[0, 1/M]$; moreover, in such interval, the symmetry of $G(F)$ implies that $G^{\text{prod}}(F)$ is symmetric around the point $F = 1/(2M)$. Consequently, $G^{\text{prod}}(F)$ can be reasonably approximated, as far as the secondary lobes are negligible, with the function $G^{\text{symm}}(F - \frac{1}{2M})$ where:

$$G^{\text{symm}}(F) \triangleq \begin{cases} G^{\text{prod}}(F + \frac{1}{2M}) & |F| \leq \frac{1}{2M} \\ 0 & \text{otherwise} \end{cases}. \tag{12}$$

Consequently, $G^{\text{symm}}(F)$ is real and even and, also its inverse Fourier transform $g^{\text{symm}}[n]$ is real (and also symmetric). Then, the inverse Fourier of $G^{\text{prod}}(F)$ can be written as $g^{\text{symm}}[n]e^{j2\pi\frac{1}{2M}n}$. The symmetry property imposed on the designed filter and the spectral containment of the pulse energy within its main lobe implies, therefore, that the scalar product $r_i$ between $g[n - iM/2]$ and $g[n]e^{j2\pi\frac{1}{M}n}$ can be written as follows:

$$\begin{aligned} r_i &\triangleq \sum_{n=-\infty}^{+\infty} g\left[n - i\frac{M}{2}\right]\left(g[n]e^{j2\pi\frac{1}{M}n}\right)^* = \int_{-\frac{1}{2}}^{\frac{1}{2}} G(F)G(F - 1/M)e^{-j2\pi i\frac{M}{2}F}\mathrm{d}F \\ &\simeq \left[g^{\text{symm}}[n]e^{j2\pi\frac{1}{2M}n}\right]_{|n=-iM/2} = j^{-i}g^{\text{symm}}\left[i\frac{M}{2}\right]. \end{aligned} \tag{13}$$

The relation (13) implies that the scalar product between adjacent subcarriers is purely real for even $i$ and purely imaginary for odd $i$ because we have already shown that $g^{\text{symm}}[n]$ is real valued. Therefore, we can still recover the transmitted symbol provided that a purely imaginary symbol is transmitted on the adjacent interfering subcarrier for even $i$ and a purely real one for odd $i$ when a purely real symbol is transmitted on the useful one and, *vice versa*, a purely real one for even $i$ and a purely imaginary one for odd $i$ when a purely imaginary symbol is transmitted on the useful one; here $iM/2$ represents the misalignment between the pulses transmitted on the adjacent subcarriers, as shown in the left-hand side of (13). More specifically, the signal $s_k[n]$ transmitted on the $k$ subcarrier is written as follows

$$s_k[n] \triangleq \sum_{i=-\infty}^{+\infty} j^i a_k[i] g_k\left[n - i\frac{M}{2}\right] \tag{14}$$

where the pulse $g_k[n]$ used for transmissions on the $k$th subcarrier is defined as follows:

$$g_k[n] \triangleq j^k g[n]e^{j2\pi\frac{k}{M}n} \tag{15}$$

and $a_k[i]$ denotes the real-valued symbol transmitted on the $k$th subcarrier with the $i$th time-shifted pulse. Relation (13) implies that the scalar product between the interfering pulse $j^{i+m}g_{k+1}[n-(i+m)M/2]$ on the adjacent subcarrier and the useful

one $j^i g_k[n - iM/2]$ is purely imaginary. Therefore, the adjacent subcarrier does not interfere with the detection of the useful symbol $a_k[i]$ provided that it is real-valued. Moreover, the PSD of the signal $s_k[n]$ can be written as $\frac{P_a}{M}\left|G\left(F - \frac{k}{M}\right)\right|^2$ where $Pa \triangleq E[|a_k[i]|^2]$; furthermore, the PSDs of the signal transmitted on each subcarrier are summed when the symbols transmitted on the different subcarriers are assumed to be uncorrelated.

Let us also note that the basic property used here (i.e., $j^i r_i$ real-valued) does not hold for relative delays different from $iM/2$ (see relation (13)); therefore, a relative timing offset between the signals received on adjacent subcarriers generates intercarrier interference. This represents a problem when a timing estimation error is present while, also in an uplink channel that does not use a procedure for timing alignment among users, no interference is present in practice when the timing offset of the useful signal is perfectly estimated, provided that a single guard subcarrier among the different users is introduced.

No ICI problem from non-adjacent subcarriers is present because the secondary lobes of the designed prototype filter $g[n]$ are negligible. The spectral separation of the signals transmitted on non-adjacent subcarriers also implies that the presence of a relative timing offset and of a relative CFO among different users does not affect the receiver performance. It is only necessary that the receiver correctly estimates the timing offset and the CFO of each user; no timing and CFO alignment procedure among the different users is necessary, which is a crucial advantage when the number of different users is large.

The signal design in (14) and (15) is also able to solve the ISI problem in the considered transceiver. In fact, the scalar product between the interfering term $j^{i+m} g_k[n - (i+m)M/2]$ and the useful pulse $j^i g_k[n - iM/2]$ is purely imaginary for odd $m$; this property straightforwardly follows from the fact that

$$\begin{aligned} g_k[n-(i+m)M/2]g_k^*[n-iM/2] &= \left[j^k g[n-(i+m)M/2]e^{j2\pi\frac{k}{M}(n-(i+m)M/2)}\right] \\ &\quad \times \left[j^k g[n-iM/2]e^{j2\pi\frac{k}{M}(n-iM/2)}\right]^* \\ &= g[n-(i+m)M/2]e^{-j2\pi\frac{k}{M}mM/2}g[n-iM/2] \end{aligned} \quad (16)$$

is real-valued. Such a scalar product, for even $m \neq 0$, is equivalent to the scalar product between the interfering pulse $g_k[n - iM/2 - mM]$ and the useful one $g_k[n - iM/2]$, or also between the pulses $g_k[n - mM]$ and $g_k[n]$. It is well-known that the interference between the last two pulses is null provided that $g_k[n]$ satisfies the Nyquist condition, i.e., provided that the following quantity

$$\sum_{\ell=0}^{M-1}\left|G_k\left(F-\frac{\ell}{M}\right)\right|^2 = \sum_{\ell=0}^{M-1}\left|G\left(F-\frac{\ell}{M}\right)\right|^2 \quad (17)$$

**Table 1** Coefficients of the prototype filter in (18) for $K \in \{2, 3, 4\}$

| $K = 2$ | $K = 3$ | $K = 4$ |
|---|---|---|
| $G_0 = 1$ | $G_0 = 1$ | $G_0 = 1$ |
| $G_1 = 1/\sqrt{2}$ | $G_1 = 0.911438$ | $G_1 = 0.97196$ |
| – | $G_2 = 0.411438$ | $G_2 = 1/\sqrt{2}$ |
| – | – | $G_3 = 0.235147$ |

is constant with $F$; here we have denoted with $G_k(\cdot)$ and $G(\cdot)$ the Fourier transform of $g_k[\cdot]$ and $g[\cdot]$, respectively. The equality in (17), which straightforwardly follows from the definition (15), implies that the ISI problem for even $m$ can be solved by introducing a further constraint in the prototype filter design; moreover, it implies that the transmitted signal is white when all the subcarriers are active. The prototype filter, therefore, has to be real-valued, symmetric, with the right-hand side of (17) constant with $F$, and with secondary lobes sufficiently small, of course.

A method for filter design often employed in the literature determines [4] the filter according to the following expression:

$$g[n] = G_0 + 2\sum_{q=1}^{K-1} G_q \cos\left(2\pi \frac{q}{KM} n\right) \qquad n \in I_K \tag{18}$$

and null otherwise. The coefficients optimized to achieve the design requirements are specified in Table 1; the behavior of the obtained prototype filters is reported in Fig. 6. We can notice the strong improvement in the spectral selectivity for $K = 3$ and for $K = 4$; however, such improvement corresponds to a complexity increase of the transceiver. We have also to note a poorer spectral containment for $K = 2$, which is a cheaper choice from the computational point of view. Therefore, the method considered in the present section is appropriate when severe spectral containment is required and we are ready to pay for the corresponding increase of computational complexity. A better solution for the case $K = 2$ will be presented in Sect. 4. Let us also note that the orthogonality condition is only achieved approximately both because the secondary lobes are approximately null and because the condition (17) is only approximated. However, large signal-to- interference ratios are achieved with the considered filters.

## *3.2 Transmitter Structure*

When the transmitted signal can be written as

$$s[n] \triangleq \sum_{k \in \mathscr{A}} s_k[n] \tag{19}$$

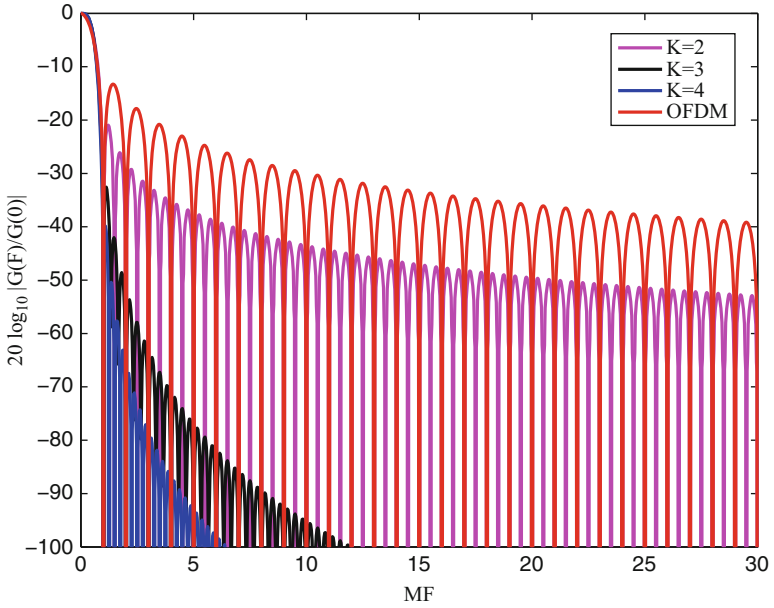


**Fig. 6** The spectral shape of the different prototype filters proposed in [4] for $K \in \{2, 3, 4\}$

with $s_k[n]$ defined in (14) and $\mathscr{A}$ denoting the set of the active subcarriers, we can construct [13] the transmitter using a traditional polyphase structure and the receiver according to the recently proposed frequency-despreading structure.

The standard structure for the implementation of the FBMC-OQAM transmitter has been first proposed in [9]; it first separates the overall signal as $s[n] = s^R[n] + js^I[n - M/2]$, we call $a^R_{i,k}$ the real-valued subsequence of the symbols $a_k[i]$ with halved bit-rate transmitted by the signal $s^R[n]$ and with $a^I_{i,k}$ the real-valued subsequence of the symbols $a_k[i]$ with halved bit-rate transmitted by the signal $s^I[n]$. A first part of the structure constructs the signal $s^R[n]$ as follows while the second part constructs the signal $s^I[n]$.

The generation of the sequence $s^R[n]$ is equivalent to the generation of the sequence of $M \times 1$ vectors $\mathbf{d}^{(R)}_i$ whose $k$th component $d^{(R)}_{i,k}$ is equal to $s^R[iM + k]$ for $k \in \{0, 1, \ldots, M - 1\}$. Such a vector can be written [13] as follows:

$$\mathbf{d}^{(R)}_i = \sum_{\ell=-\infty}^{+\infty} \mathbf{b}^{(R)}_\ell \odot \mathbf{g}_{i-\ell} \tag{20}$$

where $\odot$ denotes the Hadamard product, the vector $\mathbf{g}_i$ is defined so that its $k$th component is $g[k + iM - KM/2]$ and $\mathbf{b}^{(R)}_\ell$ is the IDFT of the vector with $k$th component equal to $j^k a^R_{\ell,k}$ ($k = 0, 1, \ldots, M - 1$). The finite length property of the

prototype filter implies that the sum in (20) can be limited to $\ell \in \{i-K+1, i-K+2, \ldots, i\}$. Consequently, (20) can be rewritten as

$$\mathbf{d}_i^{(R)} = \mathbf{g}_0 \odot \mathbf{b}_i^{(R)} + \mathbf{g}_1 \odot \mathbf{b}_{i-1}^{(R)} + \ldots + \mathbf{g}_{K-1} \odot \mathbf{b}_{i-(K-1)}^{(R)}. \tag{21}$$

The name PPN, often used for the standard structure, is related to the use of the polyphase components of the prototype filter in (21). Analogously, the vector $\mathbf{d}_i^{(I)}$ whose $k$th ($k = 0, 1, \ldots, M-1$) component is equal to $s^I[iM+k]$ can be evaluated according to the following relation:

$$\mathbf{d}_i^{(I)} = \mathbf{g}_0 \odot \mathbf{b}_i^{(I)} + \mathbf{g}_1 \odot \mathbf{b}_{i-1}^{(I)} + \ldots + \mathbf{g}_{K-1} \odot \mathbf{b}_{i-(K-1)}^{(I)} \tag{22}$$

where $\mathbf{b}_\ell^{(I)}$ is the IDFT of the vector with $k$th component equal to $j^k a_{\ell,k}^I$ ($k = 0, 1, \ldots, M-1$).

### 3.3 Receiver Structure

The previously described orthogonality property implies that the optimum receiver (in the maximum likelihood (ML) sense) in AWGN channel requires to evaluate the real part of the scalar product between the received signal $r[n]$ and the useful pulse $j^i g_k\left[n - i\frac{M}{2}\right]$ in order to estimate $a_k[i]$:

$$D_k^i \triangleq \Re\left\{\sum_{n=-\infty}^{+\infty} r[n] j^{-i} g_k^*\left[n - i\frac{M}{2}\right]\right\} = (-1)^{kK} \sum_{q=-(K-1)}^{K-1} (-1)^{qK} G_q \Re\left\{R_{i,kK+q} j^{-i-k}\right\} \tag{23}$$

where

$$r_i[n] \triangleq r\left[n + i\frac{M}{2}\right] \tag{24}$$

and with $r_{i,\ell}$ the $\ell$th component ($\ell = 0, 1, \ldots, KM-1$) of the vector

$$\mathbf{r}_i \triangleq \left[r_i\left[-\frac{KM}{2}\right], r_i\left[-\frac{KM}{2}+1\right], \ldots, r_i\left[\frac{KM}{2}-1\right]\right] \tag{25}$$

whose DFT has been denoted with $\mathbf{R}_i$. The last equality in (23) follows from the use of the expression (18) and specifies a simple structure, called frequency-despreading, for implementing the receiver. The most complex stage lies in the implementation of the DFT of $\mathbf{r}_i$ over $KM$ points every $M/2$ time steps and by the further $M$ (one for each value of the index $k$) linear combinations of $2K-1$ elements of the DFT outputs with the same coefficients $(-1)^{qK} G_q$.

In the presence of the normalized carrier frequency offset $\varepsilon$ on the received signal, the previous expression can be straightforwardly extended. In fact, in such a case,

the ML receiver in AWGN channel can be written as

$$D_k^i \triangleq \Re\left\{\sum_{n=-\infty}^{+\infty} r[n]e^{-j2\pi\frac{\varepsilon}{M}n}j^{-i}g_k^*\left[n-i\frac{M}{2}\right]\right\}$$
$$\simeq (-1)^{kK}\sum_{q=-(K_\varepsilon-1)}^{K_\varepsilon-1}(-1)^{qK_\varepsilon}\Re\left\{G_{q,\varepsilon}^* R_{i,kK+q}j^{-i-k}\right\}. \tag{26}$$

The parameters $K_\varepsilon$ and $G_{q,\varepsilon}$ are defined by the following approximation, used for deriving (26):

$$g[n]e^{j2\pi\frac{\varepsilon}{M}n} \simeq \sum_{q=-(K_\varepsilon-1)}^{K_\varepsilon-1} G_{q,\varepsilon}e^{j2\pi\frac{q}{KM}n} \tag{27}$$

and, therefore, $MG_{q,\varepsilon}$ is the value of $G(F)$ for $F = \frac{q-K\varepsilon}{KM}$. Since it can be simply evaluated by Dirichlet relation with minor complexity increase [14], the structure (26) can easily compensate in the frequency domain the CFO of the received signal. The complexity increase is usually very small since the value $K_\varepsilon$ is much close to $K$ because the spectral containment of the prototype filter makes negligible the increase $K_\varepsilon - K$ necessary to achieve a reasonable approximation in (27).

The proposed scheme exhibits an increase in terms of computational complexity with respect to the classical OFDM. In fact, the main computational weight is given by the DFT stage. In the transmitter, implemented according to the polyphase structure, we need an FFT over $M$ points to be performed with a rate $2/M$, i.e., once every $M/2$ samples. This requires to roughly double the complexity of the OFDM transceiver where an FFT over $M$ points is required every $M + M_p$ samples. The receiver, implemented according to the frequency-despreading structure, roughly requires to calculate an FFT over $KM$ samples with rate $2/M$ and, therefore, exhibits a complexity increase roughly quantified by the factor $2K$; we could simply spend only a factor 2 of receiver complexity increase by resorting to the polyphase structure also at the receiver.

When the complexity counting includes also the computational complexity required for frequency-domain CFO compensation, the complexity-increase factor becomes smaller because classical OFDM (and polyphase receiver for FBMC-OQAM) is not able to perform it with the same efficiency [5, 14]. This becomes particularly important when the receiver has to operate with signals transmitted by different users, each with its own value of CFO. While, in the case of the single user transmitting over all the subcarriers, the upper-bound for an inefficient CFO compensation in the classical OFDM system is the complexity doubling required by the time-domain CFO compensation, when the number of users becomes larger, such upper bound can become much larger than the complexity of the proposed structure. A detailed comparison of the computational complexity of the classical

OFDM receiver including CFO compensation procedures and that of the frequency-despreading structure is not available yet; however, the traditional conclusion of a complexity-increase factor equal to $2K$ (of the proposed FBMC-OQAM in comparison to classical OFDM system), which is certainly a good approximation when the complexity of the CFO compensation is not included in the counting, seems much far from the practice, especially when a large number of different users is considered at the receiver. An important reduction of the complexity-increase factor is further encountered when the procedures for timing and CFO alignment among the different users in an uplink scenario, only needed in the classical OFDM system but not in the FBMC-OQAM system, are included in the counting. For such a reason, we do not consider the complexity increase a significant reason to not adopt the FBMC-OQAM system.

### *3.4 Equalization Stage*

The equivalent channel is usually modeled as a linear filter with impulse response $h_c[n]$, assumed not null in the interval $[0, 1, \ldots, M_h]$; then, the received signal $r_h[n]$ can be written as

$$r_h[n] = \sum_{\ell=0}^{M_h} h_c[\ell] s[n-\ell] \tag{28}$$

where $s[n]$ is specified in (19). The multipath channel introduces new delayed versions of the useful transmitted signal; consequently, new delayed versions of the interfering signals, whose delay (once set to zero the delay of the useful component) is not multiple of $M/2$, are now present at the receiver and affect the basic transceiver considered in the previous stage; therefore, the receiver needs a proper equalization stage.

The equalization stage is present also in the classical OFDM receiver; however, the presence of cyclic prefix much simplifies its design: a single-tap equalizer per each FFT output is sufficient for implementing the optimum equalizer. In fact, the main purpose of introducing the cyclic prefix at the transmitter is the simplification of the equalizer stage at the receiver.

The abandon of the line of the *time separation* makes appear insufficient the simple single-tap equalization stage. Nevertheless, the standard solution for equalization stage in FBMC-OQAM closely resembles the equalization stage in OFDM; in fact, the receiver structure (26) is usually modified by replacing the FFT output $R_{i,m}$ with $R_{i,m}c_m$, i.e., a single-tap equalizer is commonly used at the receiver output; the design of the coefficient $c_m$ can follow a zero-forcing approach or an MMSE approach. When we follow a zero-forcing approach, the coefficient $c_m$ is set as $H_c^{-1}(m/(KM))$ where $H_c(F)$ is the Fourier transform of the channel impulse response. As already done with reference to the previously considered CFO compensation stage, here we are neglecting the effect of the estimation error.

It is well known in the literature (e.g., see [13]) that the single-tap equalizer is only an approximate solution, i.e., the receiver equipped with such an equalizer stage shows a performance loss with respect to that achieved on the flat channel, with the same fading statistics, unless the value of $M$ is sufficiently large. Determining the value of $M$ for which such approximate solution shows limited performance loss is often neglected in the literature; however, it is usually considered that the use of the single-tap structure as equalization stage of the FBMC-OQAM receiver provides minor performance loss only if $M$ is sufficiently large. In other terms, there may be values of $M$ for which it makes sense to consider alternative structures.

An interesting equalizer structure suited to the considered receiver scenario where an FFT bank is already present at the receiver is the frequency-domain implementation of a linear equalizer [10]. Such a linear equalizer can be designed by zero-forcing approach; then, its form can be written as $H_c^{-1}(F)$ when we resort to the zero-forcing design but we can also use a different expression provided that the noise power (or better the noise PSD) has been correctly estimated.

Since the output of such a linear equalizer represents the input of the receiver structure described in (26), we need its output values $\tilde{r}[n]$ in the interval $I_i \triangleq \{(i-K)M/2, \ldots, (i+K)M/2-1\}$ in order to be able to construct the vector $\mathbf{r}_i$ in (25) and evaluate its DFT $\mathbf{R}_i$. When the equalization stage is perfect, the useful component of $\tilde{r}[n]$ coincides with $s[n]$ while the noise may be amplified in correspondence of the weaker subcarriers.

To evaluate the output samples of the signal $\tilde{r}[n]$ in the interval $I_i$, we need to extract from the received signal $r_h[n]$ the samples in the interval $\overline{I}_i \triangleq \{(i-K)M/2 - n_c, \ldots, (i+K)M/2 - 1 + n_a\}$. Such an interval includes (in its central part) the interval $I_i$ and is extended using $n_c$ samples on the left side and $n_a$ samples on the right side. We denote its length with $\overline{K}M \triangleq KM + n_c + n_a$ and, consequently, define the input signal $\overline{r}_h^{(i)}[n]$ as follows:

$$\overline{r}_h^{(i)}[n] = \begin{cases} r_h[n] & n \in \overline{I}_i \\ 0 & n \notin \overline{I}_i \end{cases}. \tag{29}$$

Let us assume that we design the linear filter according to the zero-forcing criterion as the filter with impulse response $H_c^{-1}(F)$ evaluated as follows:

$$H_c^{-1}(F) \simeq \sum_{n=-M_a}^{M_c} h_{\text{inv}}[n] e^{-j2\pi nF} \tag{30}$$

where $M_c$ denotes the causal memory and $M_a$ the anticausal memory of the inverse filter to be implemented with impulse response $h_{\text{inv}}[n]$. Since the channel is modeled as an FIR filter, its inverse will have an infinite length; we assume that the length of the memories $M_a$ and $M_c$ is sufficient to reasonably neglect the unavoidable differences between the FIR structure and the actual IIR structure so that the

equality holds for any $F$ with good approximation. Note that the values of $M_a$ and $M_c$ mainly depend on the zeros of the channel closest to the unit circle on the internal and external side, respectively. Once accepted that we are trying to implement a noncausal filter with finite memory specified in (30), we first consider the implementation of the causal FIR filter with impulse response $h_{\text{inv}}[n-M_a]$, whose output can be therefore written as $\tilde{r}[n-M_a]$. To evaluate such an output, let us first collect the nonnull values of the signal $\overline{r}_h^{(i)}[n]$ in the vector $\overline{\mathbf{r}}_i$ of length $\overline{K}M$ and let us define the vector $\mathbf{h}_{\text{inv}}^{(a)} \triangleq \mathbf{h}_{\text{inv}} \odot \mathbf{w}$ of the same size where the $\ell$th component ($\ell = 0, 1, \ldots, \overline{K}M-1$) of $\mathbf{h}_{\text{inv}}$ is equal to the Fourier transform of $h_{\text{inv}}[n]$ (i.e., $H_c^{-1}(F)$) and the $\ell$th component of the vector $\mathbf{w}$ is equal to $e^{-j2\pi M_a F}$, in both cases for $F = \ell/(\overline{K}M)$. Consequently, it is well known that the vector $\widetilde{\mathbf{r}}_i^{\text{w}}$ of the same size defined as follows:

$$\widetilde{\mathbf{r}}_i^{\text{w}} = \text{IDFT}\left[\text{DFT}\,[\overline{\mathbf{r}}_i] \odot \mathbf{h}_{\text{inv}}^{(a)}\right] = \text{IDFT}\left[\text{DFT}\,[\overline{\mathbf{r}}_i] \odot \mathbf{h}_{\text{inv}} \odot \mathbf{w}\right] \tag{31}$$

is such that its $\ell$th component $\tilde{r}_{i,\ell}^{\text{w}}$ is equal to

$$\tilde{r}_{i,\ell}^{\text{w}} = \tilde{r}[n-M_a]_{|n=(i-K)M/2-n_c+\ell} \qquad \ell \geq M_a + M_c \tag{32}$$

with obviously $\ell \leq \overline{K}M-1$; the number of points of the DFT and IDFT in (31) is equal to $\overline{K}M$. Moreover, the vector

$$\widetilde{\mathbf{r}}_i = \text{IDFT}\left[\text{DFT}\,[\overline{\mathbf{r}}_i] \odot \mathbf{h}_{\text{inv}}\right] \tag{33}$$

is the circularly shifted version of $\tilde{r}_{i,\ell}^{\text{w}}$ of $M_a$ samples on the left. Therefore, its $\ell$th component satisfies the following relation:

$$\tilde{r}_{i,\text{mod}_{\overline{K}M}(\ell-M_a)} = \tilde{r}_{i,\ell}^{\text{w}} = \tilde{r}[(i-K)M/2 - n_c + \ell - M_a] \qquad \ell \geq M_a + M_c \tag{34}$$

or, equivalently,

$$\tilde{r}_{i,\ell-M_a} = \tilde{r}[(i-K)M/2 - n_c + \ell - M_a] \qquad \ell \geq M_a + M_c \tag{35}$$

and, also,

$$\tilde{r}_{i,\ell} = \tilde{r}[(i-K)M/2 - n_c + \ell], \qquad M_c \leq \ell \leq \overline{K}M - 1 - M_a. \tag{36}$$

The relation (36) shows that we can extract the desired values of $\tilde{r}[n]$ for $n \in I_i$, provided that $n_c \geq M_c$ and $n_a \geq M_a$. In such a case, the desired vector of length $KM$ is obtained from the values of $\tilde{r}_{i,\ell}$ from $\ell = n_c$ to $\ell = n_c + KM - 1$. Therefore, the frequency domain equalizer (FDE) requires

- to select a block of length $KM$ samples of the received signal, centered on the sample $iM/2$;

- to further increase its length by including at least $M_c$ samples on the left of the block and $M_a$ samples at its right side, so determining the vector $\overline{\mathbf{r}}_i$ of length $\overline{K}M$;
- to determine the vector $\widetilde{\mathbf{r}_i}$ according to (33); here, we use the channel response since the $\ell$th component of $\mathbf{h}_{\text{inv}}$ is equal to $H_c^{-1}(F)$ for $F = \ell/(\overline{K}M)$;
- to extract from it the central part constituted by the desired $KM$ components starting from the component with index $n_c$.

The choice of the parameters $n_c$ and $n_a$ is critical because they have to be larger than $M_c$ and $M_a$, which depend on the specific channel realization; therefore, they have to be chosen sufficiently larger than the values usually encountered in practice; moreover, we have to consider that the value $\overline{K}M$ should be chosen to be a power of two. For example, one could choose $K = 3$ and $n_c = n_a = M/2$ or one could choose $K = 2$ and $n_c = n_a = M$ or, also, when $K = 1$, $n_c = n_a = M/2$.

The increase of the computational complexity is due to the DFT and IDFT stages described in (33), whose size is also slightly larger than the DFT already present in the receiver. Therefore, the considered receiver structure with its computational cost could be also used in the classical OFDM system to avoid the loss of spectral efficiency due to the cyclic prefix, as well as with reference to the transceiver described in the next section. In fact, it operates on the received signal trying to restore the transmitted signal independently of the transmitter and the receiver structure.

An adaptive behavior of such equalizer could allow to reduce the computational complexity of the structure. In fact, as soon as the channel impulse response is estimated, the values of $M_c$ and $M_a$ needed to satisfy (30) can be evaluated and the FFT and IFFT size $\overline{K}M$ modified accordingly; this would minimize the number of FFT (and IFFT) points required in correspondence of each channel realization.

The selected vector represents the input of the standard receiver structure, called $\mathbf{r}_i$ in (25); therefore, we have to perform again the DFT of the $KM$ vector with the selected components. Therefore, we have a series of three stages, first the DFT in (31), then the IFDT in (31), and finally the DFT in the standard receiver. The first two stages are over $\overline{K}M$ points while the last one is over $KM$ points. In the special case where the number of points of the longer DFT is the double of the shorter one, we can perform also the IDFT on the shorter number of points provided that a few further processing is included. In particular, let us denote with $N$ the shorter length and assume that the central part of the IDFT has to be extracted and processed according to the DFT; in such a case, we should calculate the following quantities

$$y[n] = \sum_{k=0}^{2N-1} X_k e^{j2\pi \frac{k}{2N} n} \qquad Y_k = \frac{1}{N} \sum_{n=N/2}^{3N/2-1} y[n] e^{-j2\pi \frac{k}{N}(n-N/2)} \tag{37}$$

where $X_k$ represents the $k$th component of the IDFT input while $Y_k$ can be also expressed as

$$Y_k = \frac{(-1)^k X_{2k}}{N} + \frac{j}{N} \sum_{n=0}^{N-1} \left[ \sum_{q=0}^{N-1} (-1)^q X_{2q+1} e^{j2\pi \frac{q}{N} n} \right] e^{j\pi \frac{n}{N}} e^{-j2\pi \frac{k}{N} n}. \tag{38}$$

Therefore, also the IDFT can be performed over $N$ points. In such a case, we pass from a single FFT over $N$ points usually employed with single-tap equalizer to an FFT over $2N$ followed by an IFFT and an FFT over $N$ points; the complexity-increase factor due to the equalization stage is therefore roughly equal to 4.

## 4 FBMC-PAM

The good spectral behavior of the prototype filter designed in (18) for $K = 4$ motivated its frequent use in various contributions appeared in the literature. However, the choice $K = 3$ is particularly interesting when the choice $\overline{K} = 4$ is sufficient for the equalization stage. The choice $K = 4$, instead, is not much suited to the considered equalization stage when the FFT length is imposed to be a power of two; in fact, the choice $\overline{K} = 8$ would become mandatory, with negative effects on the transceiver latency. Let us note that the transceiver latency is given by $\overline{K}M - n_c = KM + n_a$ plus the time interval needed to implement the receiver processing; it is true that pulse lengths longer than $M$ are needed to achieve the spectral containment but the further length introduced for the specific purpose of FBMC-OQAM implementation represents a contribution to the network latency that could be avoided; for such a reason, the choice $\overline{K} = 8$ should be avoided. Obviously, we can also employ a receiver structure where $\overline{K}M$ is not a power of two to maintain the transceiver latency at its minimum possible value. Also the choice $K = 4$ with single-tap equalizer is often employed in the literature and may represent a good compromise.

In order to decrease the network latency, the choice $K = 2$ is particularly interesting; however, as shown in Fig. 6, the spectral behavior of the prototype filter obtained in (18) for the case $K = 2$ is particularly poor, and this can give problems in many practical applications since many properties of the FBMC-OQAM system are related to the significant spectral containment of the desired filter. For such a reason, we consider a different multicarrier system, the FBMC-PAM [15], able of improving the spectral containment of the FBMC-OQAM system when the choice $K = 2$ is adopted.

Let us consider the following extension

$$T_c[k,n] \triangleq \sin\left[\frac{\pi}{2M}\left(n+\frac{1}{2}\right)\right] \exp\left[j\frac{\pi}{M}\left(k+\frac{1}{2}\right)\left(n+\frac{1}{2}+\frac{M}{2}\right)\right] \tag{39}$$

of the $k$th real-valued modulated lapped transform $T[k,n] = \Re\{T_c[k,n]\}$ introduced by Malvar in its seminal contribution [12]. The transform in (39), which is defined

for $k, n \in \mathscr{K}_{2M} \triangleq \{0, 1, \ldots, 2M-1\}$ and is assumed null otherwise, can be used to transmit $2M$ real data symbols at rate $1/M$. The transmitted signal is

$$s[n] = \sum_{k=0}^{2M-1} s_k[n] \qquad s_k[n] = \sum_{i=-\infty}^{+\infty} a_k[i] T_c[k, n - iM]. \tag{40}$$

When the data symbols are statistically independent with zero-mean and variance $P_a$, it can be shown [15] that $s[n]$ is wide-sense stationary and white while the PSD $S_k(F)$ of the signal $s_k[n]$ transmitted on the $k$th subcarrier is

$$S_k(F) = \frac{P_a}{M} \left| \sin\left(\frac{\pi}{4M}\right) \frac{\cos\left(2\pi M \left(F - \frac{k+1/2}{2M}\right)\right) \cos\left(\pi \left(F - \frac{k+1/2}{2M}\right)\right)}{\cos\left(2\pi \left(F - \frac{k+1/2}{2M}\right)\right) - \cos\left(\frac{\pi}{2M}\right)} \right|^2. \tag{41}$$

The study of $S_k(F)$ is useful when not all the subcarriers are active; it is centered around $F_k \triangleq \frac{k+1/2}{2M}$ and is practically separated from $S_{k+2}(F)$ and $S_{k-2}(F)$ as shown in Fig. 7, which reports the behavior of $S_0(F)$, $S_1(F)$, and $S_2(F)$ under the assumption that these PSDs at the central frequency are unit. Moreover, the property of orthogonality of the Malvar transform guarantees (see [15] for the details) that the signals transmitted on adjacent subcarriers do not interfere. Therefore, subcarrier

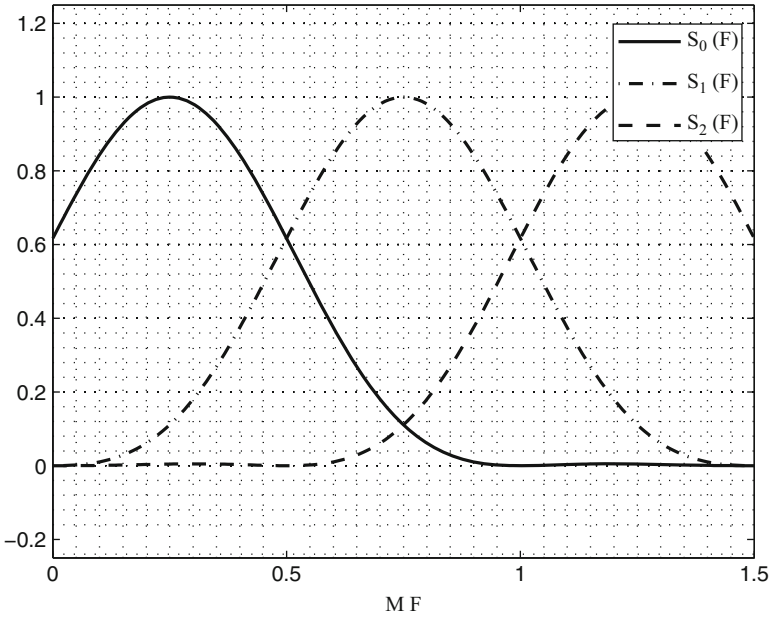


**Fig. 7** PSDs of the signal $s_k[n]$ for $k = 0$, $k = 1$, and $k = 2$

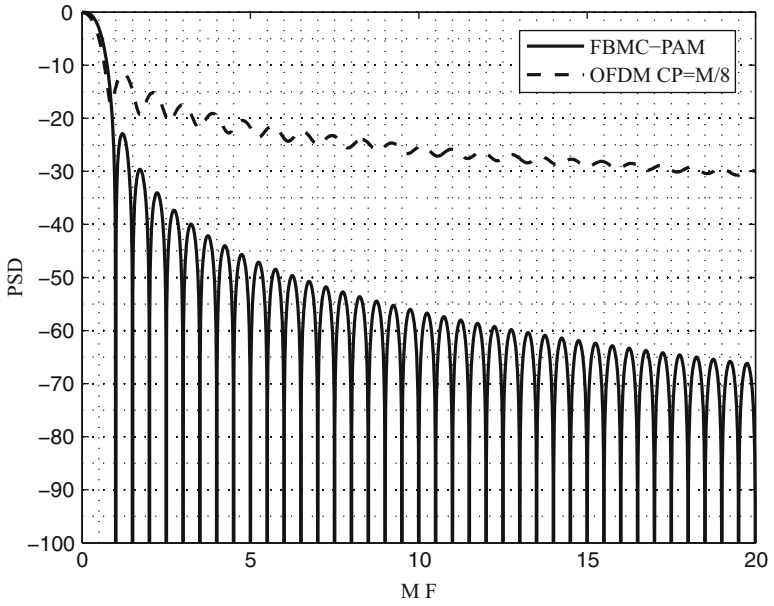


**Fig. 8** PSD in dB of the transmitted signal at the edge of the spectral bandwidth (FBMC-PAM). The PSD of the classical OFDM signal with the cyclic prefix $M_p = M/8$ is inserted for comparison

spacing is set to $1/(2M)$ and the subcarriers are used by the $2M$ real data symbols to be transmitted in a multicarrier symbol period. For such a reason, the transceiver is called FBMC-PAM. Figure 8, which reports the PSD of the transmitted signal at the edge of the spectral bandwidth, shows that the achieved levels of spectral containment are similar to those obtained by the signals considered in Figs. 3 and 4; in particular, if the out-of-band attenuation is set to $-40$ dB, a guard band equal to $2 \times 3/M$ is sufficient. Therefore, FBMC-PAM spectral efficiency is expressed by (10) with $\rho = 1$ and $M_p = 0$.

When the signal $s[n]$ is transmitted on a flat AWGN channel with channel gain $\alpha$ and $r[n]$ denotes the received signal, the ML receiver of the symbol $a_k[i]$ can be written as

$$\Re\left\{\frac{1}{\alpha M}\sum_{n=0}^{2M-1} r[iM+n]\sin\left(\frac{\pi}{2M}\left(n+\frac{1}{2}\right)\right)e^{-j\frac{\pi}{M}\left(k+\frac{1}{2}\right)\left(n+\frac{1}{2}+\frac{M}{2}\right)}\right\} \tag{42}$$

provided that no synchronization error is present and the data symbols are statistically independent. The proof, which is provided in [15], is based on the properties of the Malvar transform.

### 4.1 Efficient Transceiver Implementation

Transceiver implementation represents a crucial system issue. The structures in (39) for the transmitter side and in (42) for the receiver side represent a basic step: they are mainly useful for understanding the behavior of the proposed transceiver but they are not an effective implementation tool. For implementation purposes, it is important to verify, as proved in [15], that efficient transceiver structures can be obtained. In particular, at the transmitter side, the $i$th block $[s[iM], s[iM+1], \ldots, s[iM+2M-1]]$ of the FBMC-PAM signal in (40) is the IFFT of the vector whose components, for $k \in \mathscr{K}_{2M}$, are

$$(1-j)\left[\theta\, T_i[\mathrm{mod}_{2M}[k-1]] - T_i[k]\right] \tag{43}$$

with

$$T_i[k] \triangleq \theta^k j^k \frac{a_k[i]}{2\sqrt{2}} \qquad \theta \triangleq \exp\left(j\frac{\pi}{2M}\right). \tag{44}$$

The structure is named [15] frequency-spreading transmitter. Also an efficient receiver structure can be derived; more specifically, at the receiver side, the decision variable in (42) can be obtained as the real part of $D_k[i]$ defined as follows:

$$D_k[i] = \frac{(1+j)j^{-k}}{2\sqrt{2}\alpha}\left[\theta^{-(k+1)} R_i[\mathrm{mod}_{2M}[k+1]] - \theta^{-k} R_i[k]\right] \tag{45}$$

where $R_i[k]$ represents the DFT of $r[iM+n]$:

$$R_i[k] \triangleq \frac{1}{2M}\sum_{n=0}^{2M-1} r_i[n] e^{-j\frac{2\pi}{2M}nk} \qquad r_i[n] \triangleq r[iM+n] \qquad n = 0, 1, \ldots, 2M-1. \tag{46}$$

Such a structure is named frequency-despreading; its implementation complexity is sufficiently small so that the proposed system appears a good compromise between the spectral containment described by Fig. 7 and the computational increase.

Let us observe that the obtained expression (45) for the decision variable in (42) in terms of the FFT of the vector $r_i[n]$ in (46) suggests that frequency domain equalization can be performed by exploiting the FFT of the channel impulse response. Moreover, it also suggests that a CFO compensation procedure analogous to that introduced with reference to FBMC-OQAM can be implemented.

For illustration, the block diagram of the corresponding multicarrier receiver is shown in Fig. 9.

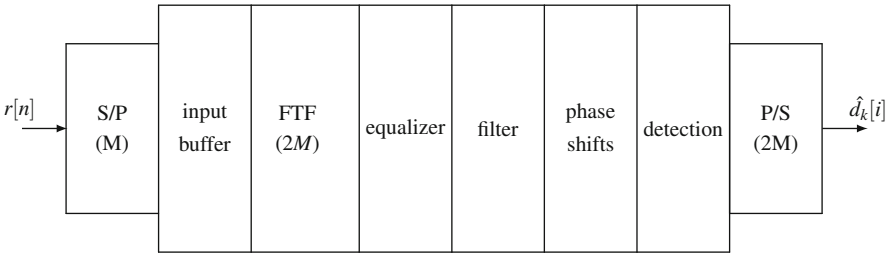


**Fig. 9** Efficient implementation of the FBMC-PAM receiver

## *4.2 Evaluating the Computational Complexity*

The computational complexity of the transceiver structure in (44) and (45) has been evaluated in [15]. The total number of real-valued flops (per discrete-time sample) required by the transmitter in (43) can be approximately written as

$$8\log(M) + 20. \tag{47}$$

Moreover, the total number of real-valued flops (per discrete-time sample) required by the receiver in (45) can be approximately written as

$$8\log(M) + 28. \tag{48}$$

In the case of the transmitter for classical OFDM the number of real-valued flops per discrete-time sample can be approximately written as follows:

$$4\log(M). \tag{49}$$

The same expression approximately represents the complexity of the receiver with the single-tap equalizer for classical OFDM. We can therefore verify that the proposed system has a computational complexity more than double than that of the classical OFDM system.

Further improvements of the transmitter structure can be achieved by using a slightly modified structure proposed in [15]; in such a case, it can be shown that the complexity-reduction factor is around 35 % in comparison with that of the implementation structure in (43) and (44). This makes appear practically negligible the increase of the implementation complexity at the transmitter of the proposed structure (in comparison with the classical OFDM system). The comparison with the complexity of the FBMC-OQAM system with $K = 2$ has to take into account that such a system needs to perform a DFT of the same size at a double rate. Therefore, the complexity of the FBMC-OQAM with $K = 2$ is double than that of the FBMC-PAM here considered.

### 4.3 Further Stages of the FBMC-PAM Transceiver

Analogously to the FBMC-OQAM transceiver, the FBMC-PAM transceiver needs to be equipped with a stage for CFO compensation and with a stage for channel equalization at the receiver side. We do not discuss in detail such stages since they can be implemented following the lines discussed in Sect. 3. In particular, it is reasonable that the CFO compensation procedure described in Sect. 3.3 can be extended to the FBMC-PAM receiver by following an analogous derivation. Consequently, we can assume that, unlike the classical OFDM system, the FBMC-PAM can compensate the CFO at the receiver with marginal increase of its computational complexity. This implies that the doubling of the required complexity discussed in the previous section can become much more limited when the procedure for CFO compensation is taken into account.

The equalization stage can be easily implemented by adopting the FDE described in Sect. 3.4; in fact, we have already discussed its independence of the specific used transceiver. In particular, in the considered receiver, we need to extract a filter output block of length $N = 2M$, as the central part of the extended block of length $\overline{K}M$; here a reasonable choice is $n_c = n_a = M$ which corresponds to the choice $\overline{K}M = 4M$; the resulting transceiver latency amounts to $\overline{K}M - n_c = 3M$. In such a case the computational cost of the equalization stage is given by an FFT over $4M$ points, an IFFT over $2M$ points according to (38), and, finally, the complex-valued multiplications for $H_c^{-1}(\cdot)$ and for the term $e^{j\pi \frac{n}{N}}$ in (38). The number of real-valued flops per each discrete-time sample can be written as:

$$16\log(4M) - 24 + 8\log(2M) - 12 + 12 = 24\log(M) + 16. \tag{50}$$

If we add such value to the receiver complexity in (48) we arrive at about $32\log(M)$, which is eight times larger than the complexity of the classical OFDM receiver reported in (49).

## 5 CP-Less OFDM and Compatibility

The perfect reconstruction property obtained with the sine filter in the previous section can be extended to other kinds of filters and, particularly, a square filter having $M$ non-zero coefficients equal to unit. Then, the emitted symbols no longer overlap, as in OFDM. Since there is no cyclic prefix, frequency domain equalization is critical and the performance of the scheme is linked to the size of the FFT window employed in the receiver.

The operation of the equalizer is illustrated in Fig. 10. A sequence of symbols generated by the transmitter is applied to the channel and, in the receiver, symbol $i$ has to be restored. The length of the FFT window is $2M$, the length of the channel impulse response is denoted with $M_h$, the length of the non-negligible causal part

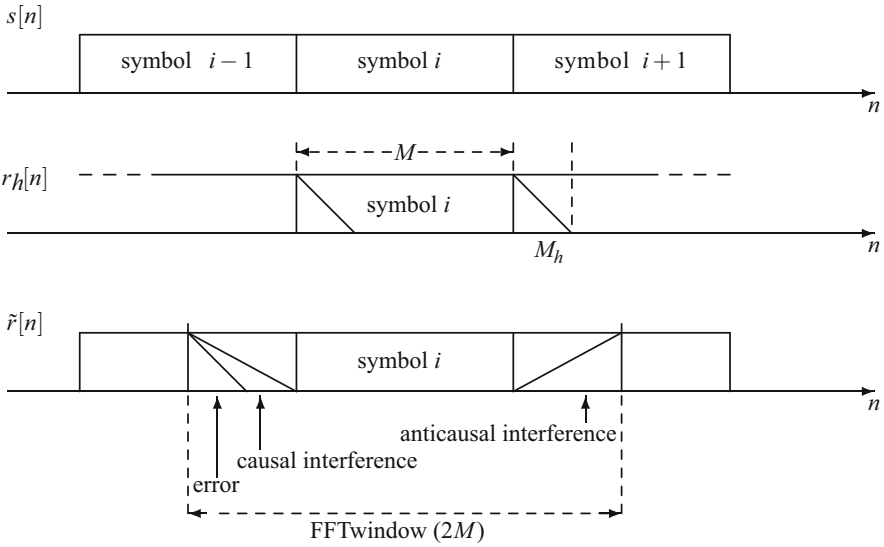


**Fig. 10** Frequency domain equalizer operation in CP-less OFDM: the transmitted signal $s[n]$ is a sequence of symbol intervals of length $M$, not separated by a cyclic prefix. The received signal $r_h[n]$ is the output of the linear convolution where $M_h$ is the channel memory. The error term in the equalizer input is due to the difference between the linear and the circular convolution and has length $M_h$. The intervals named causal interference of length $M_h + M_c$ and anticausal interference of length $M_a$ represent the samples of the equalizer output $\tilde{r}[n]$ affected by the error at the equalizer input

of its inverse is denoted with $M_c$, and, finally, the length of the non-negligible anticausal part of the channel inverse is denoted with $M_a$. At the receiver input, the samples of the symbol of interest $i$ are spread over $M + M_h$ samples. Then, the equalizer restores symbol $i$ in its proper length as well as the sections of the adjacent symbols which falls in the FFT window. However, due to the circular property of the FFT, an error signal of length $M_h$ is present at the equalizer input: it is the sum of the missing samples of symbol $i + 1$ and the overlapping samples of symbol $i - 1$. Then, the circular convolution of this error signal with the inverse channel impulse response generates a causal interference signal of length $M_h + M_c$ and an anticausal signal of length $M_a$; both the interference samples should not overlap with the desired symbol.

If the desired symbol is positioned in the middle of the FFT window, the condition for perfect restoration of the signal is

$$M_h + M_c < (\overline{K} - 1)M/2, \qquad M_a < (\overline{K} - 1)M/2. \tag{51}$$

In fact, the optimal position for the desired symbol is determined by the channel. A wireless transmission channel can be of the line-of-sight (LOS) type or non

line-of-sight (NLOS). Moreover, some residual timing offset may be added. Therefore, the error signal can occur at both ends of the FFT window and, in such conditions, it is reasonable to center the desired symbol.

Condition (51) clearly shows that the inverse channel impulse response has a critical impact on the performance of the system. If necessary, the oversize factor $\overline{K}$ can be increased but at the cost of more computational complexity and processing delay. In practice, $\overline{K}$ is mostly limited to 2, 3 or 5, with delays $1.5M$, $2M$ and $3M$ respectively, and some performance degradation will have to be tolerated for large $M_a$ and $M_c$ values. An important observation is that the error signal in Fig. 10 disappears when symbols $i-1$, $i$ and $i+1$ are identical. This can be used for channel identification with the help of a preamble. The main reason for using CP-less OFDM instead of the classical OFDM is that the scheme is robust to timing offsets, as shown in the next section, and, therefore, it is appropriate for asynchronous access.

## 6 Simulation Results and Comparison

The objective of the simulations is to compare the above techniques in a scenario where all the multicarrier systems can operate, but with CP-OFDM as the reference. Therefore, we are considering a scenario where the OFDM system is working with optimum performance and minimal receiver complexity and transceiver latency while we are interested to evaluate the loss of the alternative systems in terms of computational complexity, transceiver latency, and FDE equalizer performance needed to achieve the improvement in terms of spectral efficiency. In particular, we are interested in determining the values of $n_c = n_a$ needed to achieve a satisfactory performance in the case $K = 2$ because we are interested to limit the transceiver latency.

The main settings of our experiments follow:

- we have set the parameter $M = 256$; note that the subcarrier spacing is equal to $1/M$ in OFDM and FBMC-OQAM systems and $1/(2M)$ in FBMC-PAM;
- we have set the prototype filter for the FBMC-OQAM system according to (18) with $K = 2$;
- we have used as channel model the ITU-R vehicular B [16] where the different paths have the following discrete delays [0 1 25 36 48 56] and the following rms amplitudes [0.75 1 0.23 0.316 0.055 0.16];
- we have set the percentage of active subcarriers in OFDM, FBMC-OQAM, and FBMC-PAM transceivers as 82, 89, and 89 % of the overall number of subcarriers, respectively. Taking into account also the presence of the cyclic prefix in the OFDM system (we have set $M_p = M/4$), the ratio between the FBMC spectral efficiency and the OFDM spectral efficiency is equal to 1.36;
- the BER is obtained by averaging over all the active sub-carriers of a given multicarrier symbol and over $10^4$ independent channel realizations;

- we have used the single-tap equalizer for OFDM and for FBMC-OQAM transceiver equipped with the standard polyphase receiver while we have used the FDE structure described in Sect. 3.4 with different values of $n_c = n_a$ for the FBMC-PAM and for the CP-less OFDM. More specifically, we have denoted the different solutions under test as follows:
  - *one-tap-FBMC-2M*: the FBMC-OQAM system with $K = 2$, using the standard polyphase receiver and the single-tap equalizer;
  - *FS-FBMC-2M*: the FBMC-OQAM system with $K = 2$, using the frequency-despreading receiver and the single-tap equalizer;
  - *OFDM-CP=0*: the CP-less OFDM system using the single-tap equalizer;
  - *FBMC-PAM-2M*: the FBMC-PAM system using the single-tap equalizer;
  - *OFDM-3M*: the CP-less OFDM system using the FDE with $n_c = n_a = M$;
  - *FBMC-PAM-2M-4M*: the FBMC-PAM using the FDE with $n_c = n_a = M$;
  - *OFDM-5M*: the CP-less OFDM system using the FDE with $n_c = n_a = 2M$;
  - *OFDM CP=M/4*: the classical OFDM system with cyclic prefix $M_p = 64 = M/4 > M_h = 56$ using the single-tap equalizer.

Figure 11 shows the performance of the considered systems when the 4QAM/2PAM symbol constellation is used. The first obvious result that can be noted from the figure is that the CP-less OFDM cannot work with the simple single-tap equalizer; the other solutions using the single-tap equalizer provide performance

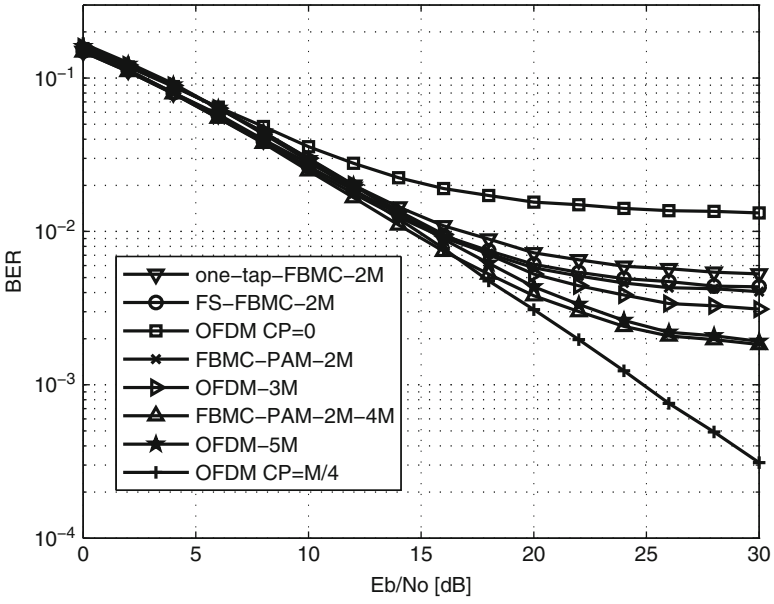


**Fig. 11** Performance of the different systems under test in the considered scenario when the 4QAM/2PAM symbol constellation is used

much similar among them, except for the solution with polyphase structure, with slightly worse performance. Such solutions exhibit significant performance loss in comparison with the classical OFDM when $E_b/N_0$ becomes larger than 15 dB. The solutions using the FDE structures exhibit the best performance; among them, the CP-less OFDM exhibits the same performance of the FBMC-PAM only when the same transceiver latency is accepted (in fact, the solution "OFDM-5M" uses $n_a = 2M$ so arriving at a total transceiver latency of $3M$, which is the same latency of the solution "FBMC-PAM-2M-4M") while the solution "OFDM-3M" with minor latency also exhibits worse performance. It is interesting to note that the choice $n_a = M$ is sufficient to achieve good performance for the FBMC-PAM structure; in practice, it shows significant performance loss in comparison with the OFDM system, which is transmitting at a minor rate on the same bandwidth, only when $E_b/N_0 > 20$ dB. Analogous results are shown in Fig. 12: the use of the 64QAM/8PAM constellation makes notice more clearly the differences among the different structures. We can observe in particular that the polyphase structure of the FBMC-OQAM receiver with single-tap equalizer is much worse than the other structures using the single-tap equalizer. Moreover, the performance of the CP-less OFDM receiver with $n_a = M$ is similar to that achieved by the single-tap structures while also the choice $n_a = 2M$ pays about 4 dB in terms of $E_b/N_0$ in comparison with the classical OFDM when the required BER is set to $10^{-2}$.

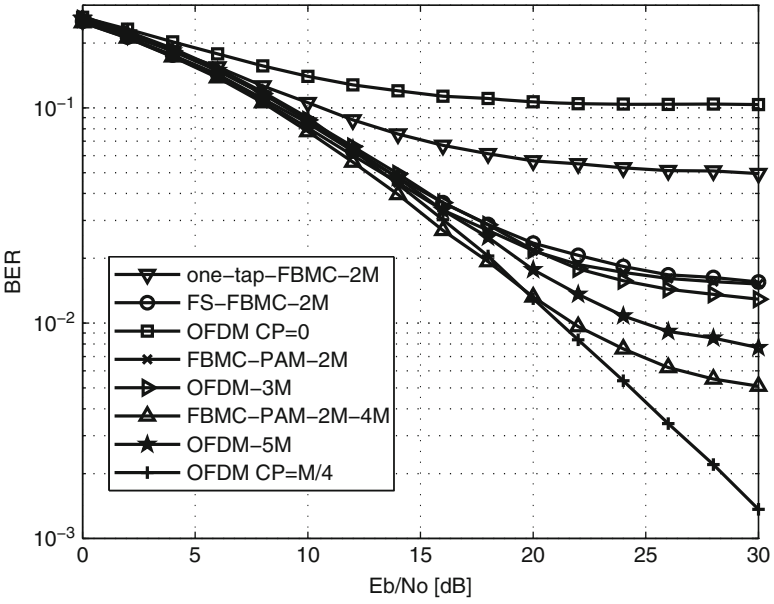


**Fig. 12** Performance of the different systems under test in the considered scenario when the 64QAM/8PAM symbol constellation is used

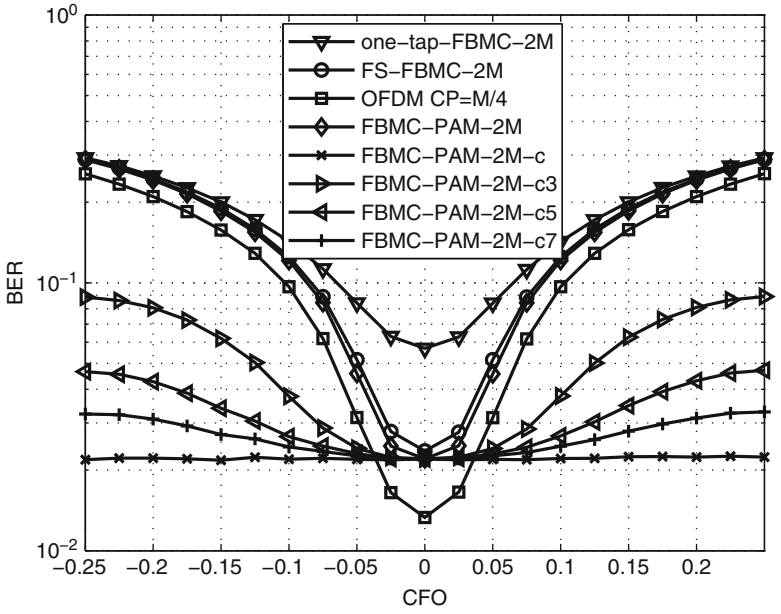


**Fig. 13** Performance of the different systems under test as a function of the CFO with $E_b/N_0 =$ 20 dB and using a 64QAM/8PAM constellation

Finally, note that practically no loss in terms of $E_b/N_0$ is payed by the FBMC-PAM-2M-4M solution for a target BER of $10^{-2}$. Therefore, such solution, which starts to pay a performance loss in comparison with the classical OFDM system only for $E_b/N_0 > 22$ dB, is able to gain its spectral-efficiency advantage in comparison with the classical OFDM system by paying with a factor three in terms of transceiver latency and with a factor eight in terms of computational complexity.

The considered complexity comparison does not take into account the cost of the CFO compensation procedure and the cost of the procedures for timing and CFO alignment among the different users, also because the procedures for CFO compensation to be employed in FBMC-PAM systems have not been defined yet in their details. However, we have used a general less efficient procedure described in [5] for the CFO compensation. We have reported in Fig. 13 the results of the simulation experiments (carried out for $E_b/N_0 = 20$ dB and using a 64QAM/8PAM constellation) aimed at verifying the effect of an uncompensated CFO on the most important of the considered solutions but also the performance achieved by applying to FBMC-PAM the CFO compensation procedure described in [5]. We denote with "FBMC-PAM-2M-c" the perfectly compensated structure with an heavy computational requirement and with "FBMC-PAM-2M-cB" the structure employing $B$ coefficients for the compensation procedure. We can notice that all the solutions analyzed in Fig. 13 are sensitive to the uncompensated CFO but, however, a few coefficients are sufficient for CFO compensation in the FBMC-PAM

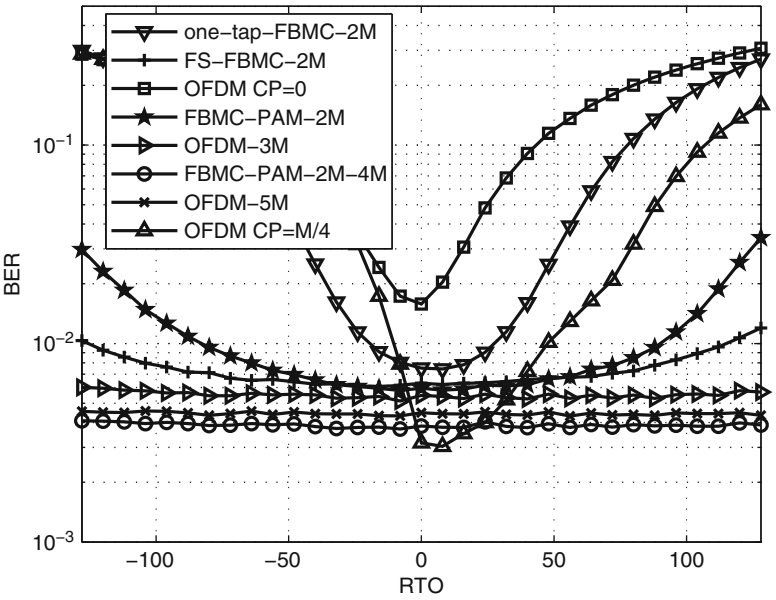


**Fig. 14** Performance of the different systems under test as a function of the residual timing offset (RTO) with $E_b/N_0 = 20$ dB and using a 4QAM/2PAM constellation

system. When a procedure analogous to that described in (26) for FBMC-OQAM will be available, the increase of the computational complexity in comparison with OFDM systems will become smaller; the results show that there is space for such improvement in CFO compensation for FBMC-PAM systems which is able to converge also with a minor number of coefficients.

Finally, we have analyzed the effect of a timing offset on the received signal. We have considered that the residual timing offset is estimated after that the receiver FFT has been performed and we analyzed the performance of the different methods for frequency-domain timing-offset compensation. The results of the experiment carried out by setting $E_b/N_0 = 20$ dB and using a 4QAM/2PAM constellation, reported in Fig. 14, confirm an obvious result: the single-tap equalizer structures cannot compensate the residual timing offset while those employing the FDE can compensate almost perfectly and without significantly increasing their computational complexity. It is worth noting that, among the solutions employing single-tap equalizer, the FBMC-OQAM and the FBMC-PAM exhibit a minor performance loss due to imperfections in timing-offset compensation. Moreover, the results of the experiment carried using the 64QAM/8PAM constellation, reported in Fig. 15, confirm the previous results and show that the advantage of the structures using the FDE is also larger.

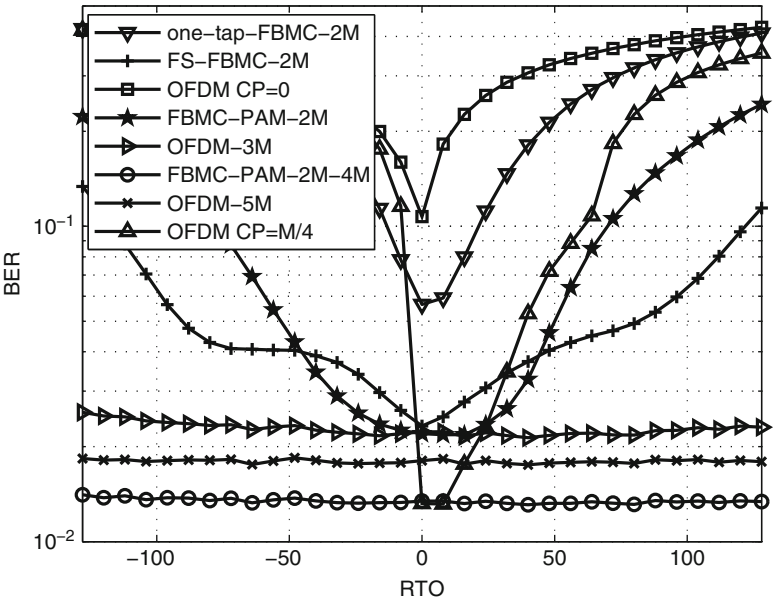


**Fig. 15** Performance of the different systems under test as a function of the residual timing offset (RTO) with $E_b/N_0 = 20$ dB and using a 64QAM/8PAM constellation

## 7 Conclusions

Some evolutions of OFDM, particularly per-user filtering and pulse shaping, lead to significant improvements in spectral separation, with moderate loss in spectral efficiency and limited increase in complexity. Although asynchronous access may remain an issue with these techniques, they are likely to be strong candidates for 5G standardization. A different perspective, which can be complementary, has been proposed above, with different trade-offs in terms of performance and complexity. It aims at maximum spectral efficiency, with adjustable performance in spectral separation and transmission characteristics. The principle is to use filter banks with frequency domain processing, exploiting the FFT. Of the two approaches presented, namely FBMC-OQAM and FBMC-PAM, the latter seems particularly appropriate to meet the objectives of 5G in terms of spectral separation, limited latency increase with respect to CP-OFDM and asynchronous access capability. In the baseline option, the increase in computational complexity is limited. If necessary, the equalizer performance can be enhanced, at the cost of more computations in the receiver, as well as more processing delay. Overall, the FBMC approach, and particularly the FBMC-PAM technique, is likely to be another strong candidate for 5G standardization.

## References


1. J. Abdoli, M. Jia, J. Ma, Filtered OFDM: a new waveform for future wireless systems, in *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)* (2015), pp. 66–70
2. E. Bala, J. Li, R. Yang, Shaping spectral leakage: a novel low-complexity transceiver architecture for cognitive radio. IEEE Veh. Technol. Mag. **8**(3), 38–46 (2013)
3. P. Banelli, S. Buzzi, G. Colavolpe, A. Modenini, F. Rusek, A. Ugolini, Modulation formats and waveforms for 5G networks: who will be the heir of OFDM? An overview of alternative modulation schemes for improved spectral efficiency. IEEE Signal Process. Mag. **31**(6), 80–93 (2014)
4. M. Bellanger, Specification and design of a prototype filter for filter bank based multicarrier transmissions, in *2001 Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. IV (2001), pp. 2417–2420
5. J. Choi, C. Lee, H.W. Jung, Y.H. Lee, Carrier frequency offset compensation for uplink OFDM-OFDMA systems. IEEE Commun. Lett. **4**(12), 414–416 (2000)
6. J. Erfanian, R.E. Hattachi, B. Daly et al., NGMN 5G white paper (2015). Available at www.dieti.unina.it/dieti/nostro.pdf
7. B. Farhang-Boroujeny, C.H. Yuen, Cosine modulated and offset QAM filter bank multicarrier techniques: a continuous-time prospect. EURASIP J. Adv. Signal Process. **2010**, 16 (2010)
8. M. Gudmundson, P.O. Anderson, Adjacent channel interference in an OFDM system, in *IEEE 46th Vehicular Technology Conference Mobile Technology for the Human Race*, vol. 2 (1996), pp. 918–922
9. B. Hirosaki, An orthogonally multiplexed QAM system using the discrete fourier transform. IEEE Trans. Commun. **29**(7), 982–989 (1981)
10. J. Li, E. Bala, R. Yang, Sliding-window frequency-domain equalization for multi-mode communication systems, in *Proceedings of 9th Annual Conference on Systems Applications and technology (LISAT 2013), Long Island, NY, USA, May 3* (2013), pp. 1–6
11. A. Loulou, M. Renfors, Enhanced OFDM for fragmented spectrum use in 5G systems. Trans. Emerg. Telecommun. Technol. **26**, 31–45 (2015). doi:10.1002/ett.2898
12. H. Malvar, Modulated QMF filter banks with perfect reconstruction. Electron. Lett. **26**(13), 906–907 (1990)
13. D. Mattera, M. Tanda, M. Bellanger, Analysis of an FBMC/OQAM scheme for asynchronous access in wireless communications. EURASIP J. Adv. Signal Process. **2015**, 23 (2015)
14. D. Mattera, M. Tanda, M. Bellanger, Frequency domain CFO compensation for FBMC systems. Signal Process. **114**, 183–197 (2015)
15. D. Mattera, M. Tanda, M. Bellanger, Filter bank multicarrier with PAM modulation for future wireless systems. Signal Process. **120**, 594–606 (2016)
16. Recommendation ITU-R M. 1225, Guidelines for evaluation of radio transmission technologies for IMT-2000 (1997)
17. V. Vakilian, T. Wild, F. Schaich, S. ten Brink, J.F. Frigon, Universal-filtered multi-carrier technique for wireless systems beyond LTE, in *IEEE Globecom Workshops (GC Wkshps)* (2013), pp. 223–228
18. J.J. van de Beek, P.O. Brjesson, M.L. Boucheret, D. Landstrom, J.M. Arenas, P. Odling, S.K. Wilson, Three non-pilot based time- and frequency estimators for OFDM. Signal Process. **80**(7), 1321–1334 (2000)
19. G. Wunder, M. Kasparick, S. ten Brink, F. Schaich, T. Wild, I. Gaspar, E. Ohlmer, S. Krone, N. Michailow, A. Navarro, G. Fettweis, D. Ktenas, V. Berg, M. Dryjanski, S. Pietrzyk, B. Eged, 5GNOW: challenging the LTE design paradigms of orthogonality and synchronicity, in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)* (2013), pp. 1–5
20. G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. Brink, I. Gaspar, N. Michailow, A. Festag, L. Mendes, N. Cassiau, D. Ktenas, M. Dryjanski, S. Pietrzyk, B. Eged, P. Vago, F. Wiedmann, 5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications. IEEE Commun. Mag. **52**(2), 97–105 (2014)

# Fundamentals of Faster-than-Nyquist Signaling

**Angelos D. Liveris and Costas N. Georghiades**

**Abstract** This chapter presents the fundamentals of Faster-than- Nyquist (FTN) signaling. As originally introduced, FTN increases the bit-rate in the signaling bandwidth by packing symbols closer in time, at the cost of introducing intersymbol interference (ISI). We begin with the Euclidean distance properties of bandwidth efficient pulses at FTN rates and describe receivers that mitigate the severe ISI. The FTN achievable information rate is compared with the Nyquist information rate for practical pulses. We then discuss FTN extension to multicarrier systems with not only time packing but also subcarrier, optimizing both the time and frequency packing.

## 1 Introduction

The basic idea behind Faster-than-Nyquist signaling is based on challenging the time orthogonality assumption in a typical single carrier system shown in Fig. 1 for the binary case. In such a system the transmit signal is

$$s_{\mathbf{a}}(t) = \sum_{k} a_k g(t - k\tau), \tag{1}$$

where $\boldsymbol{a} = [a_k]$ is the sequence of transmitted pulse-amplitude modulation (PAM) symbols and $g(t)$ is the transmit pulse shaping filter. The transmit signal is then passed through the channel, such as an additive white Gaussian noise (AWGN) channel considered here with zero-mean noise $n(t)$ and power-spectral density $N_0/2$.

A.D. Liveris (✉)
Microwave Networks Inc., 4000 Greenbriar, Stafford, TX 77477
e-mail: alive@ieee.org

C.N. Georghiades
Electrical and Computer Engineering Department, Texas A&M University, College Station, TX 77843–3128
e-mail: georghiades@tamu.edu



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_8

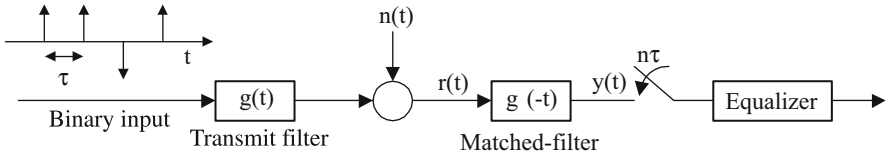


**Fig. 1** Single carrier communication system for binary faster-than-Nyquist signaling

The received signal $r(t)$ is first passed through a receive matched filter $g^*(-t)$, resulting in the output signal

$$y(t) = \sum_k a_k x(t - k\tau) + n'(t), \tag{2}$$

where $x(t) = g(t) * g^*(-t)$ and $n'(t)$ is the filtered noise with autocorrelation $R_{n'n'}(\Delta t) = x(\Delta t)N_0/2$. The matched filter output signal $y(t)$ is then sampled at the signaling rate $1/\tau$

$$y(n\tau) = \sum_k a_k\, x\left((n-k)\tau\right) + n'(n\tau). \tag{3}$$

To limit the bandwidth occupied by the transmit signal, a limited bandwidth transmit filter $g(t)$ is used. The ideal filter is the sinc filter with a strictly bandlimited, square pulse shaped frequency response. This results in the end-to-end response being a sinc pulse, i.e.,

$$x(t) = E_s\ \mathrm{sinc}\left(\frac{t}{T}\right) = E_s\ \frac{\sin(\pi t/T)}{\pi t/T}, \tag{4}$$

where assuming baseband transmission $W = 1/2T$ is the bandwidth it occupies. $E_s$ is the average PAM symbol energy.

If we use the ideal sinc pulse and the successive signal pulses are transmitted at $T = \tau$, then the matched filter output samples become

$$y(nT) = a_n\, x\,(0) + n'(nT), \tag{5}$$

i.e., each sample only depends on a single transmitted symbol and the noise samples become independent, significantly simplifying the PAM symbol detection process. This is the Nyquist constraint or time orthogonality principle and the $1/T$ signaling rate is called the Nyquist rate.

To improve bandwidth efficiency, we could signal faster with $\tau < T$ so that the signal bandwidth, e.g., using the sinc pulse, does not change, but more PAM symbols are transmitted in the same time as shown in Fig. 2. In this case the matched filter output samples in (3) will include interference from the other symbols in the

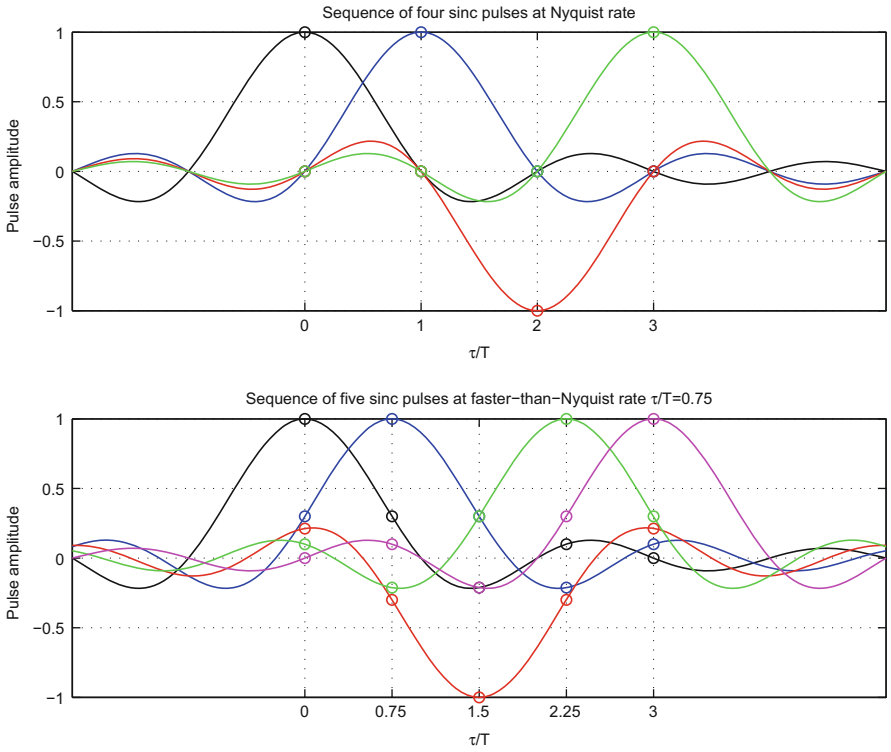


**Fig. 2** Sequence of sinc pulses at the Nyquist rate (*top*) and at the faster-than-Nyquist rate $\tau/T = 0.75$ (*bottom*). The values of each pulse at the sampling instants is also shown in both cases

sequence as the Nyquist constraint no longer holds as the sampling instants in Fig. 2 no longer correspond to the zero crossings of the neighboring pulses.

Surprisingly, with the ideal sinc pulses Mazo [1] was the first to realize that the minimum Euclidean distance between binary sequences transmitted at a rate up to 25 % higher than Nyquist was the same as the minimum distance of sequences transmitted at the Nyquist rate. This minimum distance limit is referred to as the Mazo limit in the FTN literature.

Since the minimum Euclidean distance between sequences referred to maximum likelihood detection and the intersymbol interference (ISI) between symbols when signaling faster than the Nyquist rate was severe, theoretically infinite for the sinc pulse, achieving the promised minimum distance gains meant significant receiver complexity. Given that at about the same time the Viterbi algorithm was just being introduced for ISI receivers [2, 3], it seemed unrealistic to come up with a limited complexity receiver that takes advantage of Mazo's minimum distance results [4]. Mazo himself admitted that his FTN research was triggered by "curiosity" [5].

In the decades that followed several developments changed this landscape:

(1) First, reduced complexity receivers started being introduced with performance close to the maximum likelihood receivers even with severe intersymbol interference [6–10]. This development allowed handling the FTN intersymbol interference with reasonable complexity.
(2) Second, the introduction of turbo codes [11] led to applying the turbo concept to several other scenarios, such as turbo equalization [12]. This allowed achieving coding gains over channels with severe ISI with an iterative receiver at reasonable complexity. The channel capacity approaching turbo codes also directed more focus on the achievable information rates than just the maximum likelihood performance. This further helped strengthen the FTN approach as a way to take advantage of all available frequencies in the channel as we will see next.
(3) The extension of the "time-packing" approach of FTN to frequency-packing in a multicarrier system [13] as well as to combined time and frequency packing, the two-dimensional Mazo limit [14], showed even more promising minimum distance gains.
(4) Modern telecommunication systems and their constant quest for increased throughput started reaching a limit of resources, especially bandwidth, even though processing power kept growing.

All these developments led to renewed interest in FTN lately as signal processing techniques used in current telecommunication systems keep advancing. Meanwhile, resources, such as bandwidth, have become more scarce and their efficiency must improve even at the cost of increased processing complexity. So, FTN is considered for the 5G standard [15] and for satellite and optical communications [16, 17].

In the following sections we address several key aspects of faster-than-Nyquist (FTN) signaling.

## 2 Minimum Distance

The ideal sinc pulse discussed in the Introduction is a special case of the general family of raised-cosine pulses [18]

$$x(t) = E_s \operatorname{sinc}(t/T) \frac{\cos(\pi \alpha t/T)}{1 - (2\alpha t/T)^2}, \tag{6}$$

with $0 \leq \alpha \leq 1$ being the roll-off factor and $W = (1 + \alpha)/(2T)$ the bandwidth they occupy. The ideal sinc pulse has $\alpha = 0$. In practical communications systems a non-zero roll-off factor $\alpha$ is used, typically with $\alpha$ in the range of 0.1–0.5 for higher bandwidth efficiency.

ISI free transmission satisfying the Nyquist constraint for these pulses means $\tau = T = (1 + \alpha)/(2W) \geq 1/(2W)$. In [19] Mazo's $d_{min}$ observations for the binary case were numerically extended to non-zero roll-off $\alpha$ as shown in Fig. 3 for roll-off

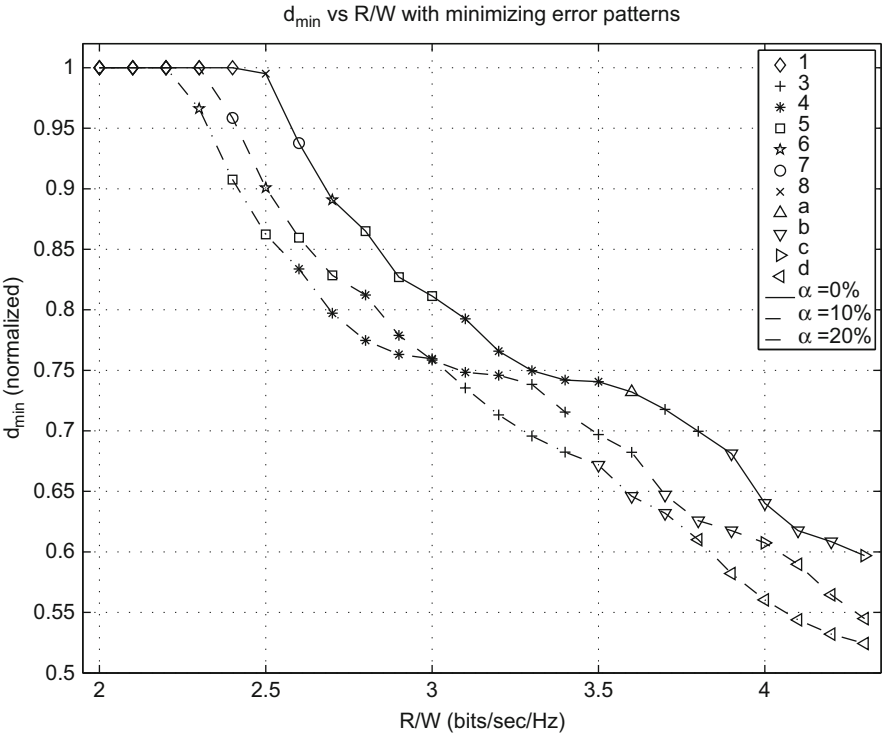


**Fig. 3** Minimum Euclidean distance $d_{min}$ as a function of the bandwidth efficiency for binary faster-than-Nyquist signaling (FNS) for the sinc pulse ($\alpha = 0\,\%$) and the raised-cosine pulses with roll-off $\alpha = 10, 20\,\%$. The error sequences $\boldsymbol{e}_{min}$ are also shown. The numbers show the length of the single alternating block $\boldsymbol{e}_{min}$ [e.g., $5 = \pm(\ldots 000 - + - + - 000\ldots)$] and the letters stand for $\boldsymbol{e}_{min}$ containing more than one block [$a = \pm(\ldots 000 - + - 0 + - + 000\ldots)$, $b = \pm(\ldots 000 - +0 - +0 - +000\ldots)$, $c = \pm(\ldots 000 - +0 - +00 - +0 - +000\ldots)$, $d = \pm(\ldots 000 - +0 - +000\ldots)$]

factors of $\alpha = 0, 10, 20\,\%$. The minimum distance is plotted as a function of the bandwidth efficiency $R/W$, i.e., the ratio of the signaling rate $R = 1/\tau$ to the utilized bandwidth $W = (1+\alpha)/(2T)$ (measured in bits/s/Hz). The Nyquist (orthogonal) signaling rate is $1/T$ corresponding to a bandwidth efficiency of $2/(1+\alpha)$ bits/s/Hz. The highest Nyquist rate bandwidth efficiency is 2 bits/s/Hz, achieved by the ideal sinc pulse ($\alpha = 0$).

We define the normalized minimum Euclidean distance $d_{min}$ as

$$d_{min} = \min_{\boldsymbol{e}\neq\boldsymbol{0}} \frac{d(\boldsymbol{e})}{2\sqrt{E_b}}, \tag{7}$$

where $\boldsymbol{e} = \boldsymbol{a} - \boldsymbol{b}$ is the difference (error) sequence between the transmitted sequences $\boldsymbol{a}$ and $\boldsymbol{b}$, with $e_k = (a_k - b_k) \in \{0, \pm 2\}$ for the binary case, and the distance $d(\boldsymbol{e})$ defined as

$$d^2(\boldsymbol{e}) = \int_{-\infty}^{\infty} [s_{\boldsymbol{a}}(t) - s_{\boldsymbol{b}}(t)]^2 \, dt = \sum_k \sum_l e_k e_l x((k-l)\tau) \, . \tag{8}$$

Figure 3 shows that the $d_{min}$ between binary sequences is not reduced below the $d_{min}$ for the Nyquist rate case. Even for non-zero roll-off this $d_{min}$ behavior extends to high enough bandwidth efficiency $R/W$ that overcomes the bandwidth efficiency loss caused by the non-zero roll-off.

As long as there is no change in $d_{min}$ compared to the ISI-free case, the error sequence $\boldsymbol{e}_{min}$ that yields $d_{min}$ is one with only a single nonzero component; i.e., a pair of input binary sequences is at $d_{min}$ if the two sequences differ at a single position. Surprisingly, the error patterns $\boldsymbol{e}_{min}$ that first cause degradation in $d_{min}$ contain a single long block of alternating polarity and, more generally, they consist of one or more blocks of alternating polarity, as Fig. 3 shows. This was also observed in [1, 20, 21] for sinc pulses.

These error events $\boldsymbol{e}_{min}$ are depicted for all three roll-off factors. The numbers in the legend stand for error sequences containing a single block of alternating polarity of this length; for example, 5 represents the error pattern $\pm(\ldots 000 - + - + - 000 \ldots)$, where $+$ and $-$ correspond to $e_k = 2$ and $e_k = -2$, respectively. Each letter corresponds to an error pattern consisting of more than one block. Thus, $a = \pm(\ldots 000 - + - 0 + - + 000 \ldots)$, $b = \pm(\ldots 000 - +0 - +0 - +000 \ldots)$, $c = \pm(\ldots 000 - +0 - +00 - +0 - +000 \ldots)$, $d = \pm(\ldots 000 - +0 - +000 \ldots)$. Close to the "knee", the point where $d_{min}$ starts decreasing, the $d_{min}$ error event is a single block of alternating polarity, while for higher rates, first its length is reduced and then it starts splitting into two or three alternating blocks.

Such patterns can only arise from a pair of binary sequences that include alternating blocks of some length with opposite polarity. What this means for the single block of alternating polarity close to the "knee" is that the probability of such sequences is rather low. Therefore, the $d_{min}$ reductions close to the "knee" are not likely to affect the maximum likelihood performance of systems operating at this FTN rate [1]. The lower bound on the bit error rate is given by combining the error event lower bound from [22]

$$P_{event} \geq \Pr[\boldsymbol{e}_{min}] \, Q\left( d_{min} \sqrt{\frac{2E_b}{N_0}} \right) \tag{9}$$

with the number of bit errors resulting from $\boldsymbol{e}_{min}$, i.e.,

$$P_e \geq \sum_k |e_{min,k}| \Pr[\boldsymbol{e}_{min}] \, Q\left( d_{min} \sqrt{\frac{2E_b}{N_0}} \right) \tag{10}$$

where $\boldsymbol{e}_{min} = [e_{min,k}]$ and $Q(\cdot)$ is the area under the tail of the Gaussian probability density function.

For example, at the bandwidth efficiency of $R/W = 2.4$ bits/s/Hz with $\alpha = 10\,\%$, $d_{min} = 0.959$ and

$$\boldsymbol{e}_{min} = \pm[\ldots 0,\ 0,\ 0,\ -2,\ +2,\ -2,\ +2,\ -2,\ +2,\ -2,\ 0,\ 0,\ 0,\ldots]. \quad (11)$$

For this example the lower bound of the maximum likelihood performance is given by

$$P_e \geq \frac{7}{64} Q\left(0.959\sqrt{\frac{2E_b}{N_0}}\right). \quad (12)$$

Due to the low multiplicity of the $d_{min}$ error event in this example, this lower bound is less than the no-ISI $Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$ performance even for medium-to-high SNR as we will see in the next section.

These results were extended to non-binary PAM symbols in [23], where it was found that for the same roll-off factor $\alpha$ that the same $\tau$ caused the binary, the 4-PAM, and the 8-PAM minimum distance to start decreasing from the corresponding binary, 4-PAM, and 8-PAM pairwise (no-ISI) distance. Of course, at the same $\tau$ 8-PAM and then 4-PAM achieve better bandwidth efficiency. The target bandwidth efficiency determines the best combination of signaling alphabet and $\tau$, which is operating close to the $d_{min}$ curve "knee" of each signaling alphabet. For example, with $\alpha = 30\,\%$ at bandwidth efficiency of 4–5 bits/s/Hz the 4-PAM faster-than-Nyquist $d_{min}$ is larger than the binary and 8-PAM faster-than-Nyquist $d_{min}$ [23].

Another approach based on the structured form of error events and the associated binary sequences is to use constrained codes that do not allow the occurrence of input sequences to the transmit filter that generate such error events [19]. This keeps $d_{min}$ constant for even higher signaling rates at the expense of rate loss and also allows the use of limited complexity receivers.

An alternative view taken in [23] is to use a short linear precoding filter with 4–8 taps in the transmitter that does not add any redundancy. Optimizing its coefficients can improve the binary and 4-PAM $d_{min}$ without the rate loss penalty of the constrained coding approach. This approach adds additional complexity to the receiver side.

## 3 Reduced Complexity Receivers for Uncoded FTN

The intersymbol interference (ISI) introduced by FTN signaling in a single matched filter output sample extends to a large number of symbols. So, the maximum likelihood approach [2, 3] of considering all the interfering symbols results in significant complexity that is not possible to implement.

The first approach to limiting the complexity of a detector handling the severe ISI caused by FTN is to truncate the number of states in the Viterbi algorithm. This

means that only the ISI from the neighboring symbols, which typically cause most of the ISI, as shown in Table 1, is managed by the detector while the rest of the ISI is left untreated leading to performance degradation.

Approximating this residual ISI as Gaussian noise leads to the realization that it can cause an error floor at high signal-to-noise ratio (SNR) as the residual ISI is constant and does not change with the SNR [19].

Fortunately, there are several approaches to considering the complete ISI or to mitigating the effect of the residual ISI within the detector. One of the simplest approaches is to include decision feedback cancelation of the residual ISI inside the Viterbi algorithm as proposed in [10].

Results for both these approaches, truncating the Viterbi algorithm memory without and with decision feedback, can be seen in Fig. 4. The $d_{min}$ lower bound for FTN is based on the $d_{min}$ discussion of the previous section.

Another detection approach presented in [19] was based on a soft-output reduced state Viterbi detection followed by linear filtering to cancel the residual ISI.

Other reduced complexity approaches are based on the M-algorithm, which can handle non-binary modulation with lower complexity than reduced state detection [23]. In [23] it was also observed that the M-algorithm does not work well with the Ungerboeck receiver model used in this section [3, 10], which operates directly on the matched filter output samples. Instead, an approximate whitened matched filter (WMF) is employed, following Forney's receiver model [2], with which the M-algorithm performs well despite the minor mismatch required to implement the WMF. Further analysis and advantages of the WMF model for FTN signaling are explored in [24].

## 4 Turbo Equalization for Coded FTN

The previous section presented low-complexity receivers for uncoded FTN transmission. In this section we address low-complexity receivers for coded FTN transmission. Figure 5 shows such a binary coded FTN system. The channel encoder output is interleaved before it is passed to the FTN pulse shaping. At the receiver the matched filter output samples are now processed by a turbo equalizer, which performs iterative detection and decoding [12].

When combining FTN ISI detection and decoding in the turbo equalizer, there are two aspects that need to be considered:

(1) A reduced complexity soft-input soft-output (SISO) detector is needed for the severe ISI introduced by FTN signaling. This way the soft output can be input into the channel decoder and allow iterations between the two receiver modules.

**Table 1** Raised cosine ISI coefficients for the closest ten pulses from one side for $\alpha = 10\,\%$ and $\tau/T = 0.75$

| $x(0)$ | $x(0.75T)$ | $x(1.5T)$ | $x(2.25T)$ | $x(3T)$ | $x(3.75T)$ | $x(4.5T)$ | $x(5.25T)$ | $x(6T)$ | $x(6.75T)$ | $x(7.5T)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.0000 | 0.2985 | −0.2078 | 0.0954 | 0.0000 | −0.0525 | 0.0582 | −0.0328 | −0.0000 | 0.0212 | −0.0240 |

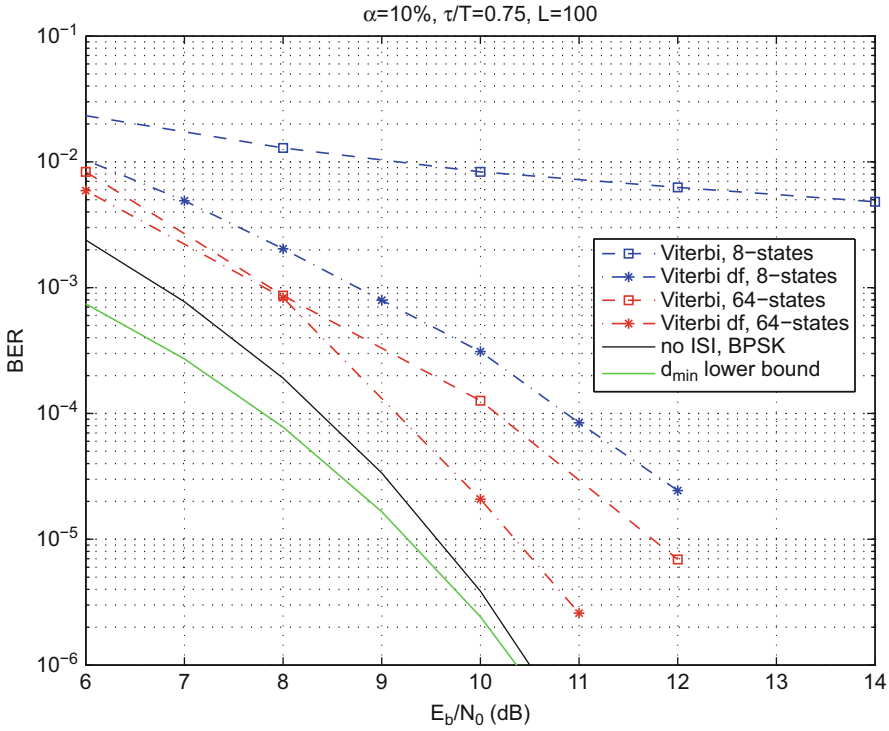


**Fig. 4** Reduced state Viterbi BER for the raised-cosine pulses with roll-off $\alpha = 10\,\%$ with bandwidth efficiency 2.42 bits/s/Hz

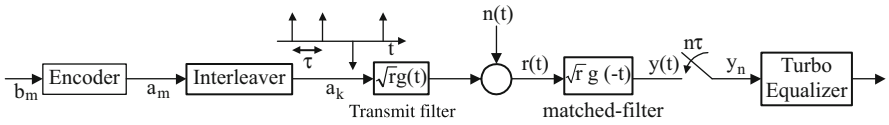


**Fig. 5** Communication system for coded binary faster-than-Nyquist signaling

(2) The effect of the interleaver is important not just due to its ability to decorrelate the iterative equalizer detector and decoder input data but also due to the nature of the long error events with FTN signaling. The interleaver will allow breaking the long alternating polarity error events, so, they can more easily be corrected by the decoder.

An advantage of the turbo equalizer is that the channel decoder helps make the suboptimum ISI detection more reliable, especially, its residual ISI cancelation, which in some reduced complexity detectors, such as the decision feedback we considered in the uncoded FTN detection section, could partly be based on tentative decisions.

What the analysis of the information rates suggests (more details in the next section) is that even the use of a very low complexity ISI detector combined with a channel decoder in an iterative (turbo) equalization scheme can work as well as a more complex ISI detector. The ISI detector beyond the SISO requirement must also consider the whole length of the ISI, even if it just performs tentative-decision cancelation of the interference.

To demonstrate this, we consider here the approach from [25] simplified for the binary case and with some adjustments from [19]. The basic idea is that the mean value of the FTN received symbols is used to cancel their effect on neighboring symbols and the uncertainty in these FTN received symbols is included in the generation of the soft-output probabilities.

When the FTN symbols are received, their statistics consist only of their received values. So, rewriting (3)

$$y_n = y(n\tau) = x(0)a_n + \sum_{k \neq n} a_k \, x\left((n-k)\tau\right) + n'(n\tau). \tag{13}$$

the intersymbol interference in $y_n$ from the neighboring symbols can be approximated as an additional source of Gaussian noise when considering only the current sample to make a decision on $a_n$. The channel information can be written as the following log-likelihood ratio

$$L_n^{ch} = \log \frac{\Pr[a_n = +1|y_n]}{\Pr[a_n = -1|y_n]} = 4\frac{E_s}{N_0 + \rho} y_n, \tag{14}$$

where $E_s = x(0)$ and $\rho$ is defined as the scaled variance of the residual ISI in $y_n$

$$\rho = \frac{2}{x(0)} E\left[\sum_{k \neq n} a_k \, x\left((n-k)\tau\right)\right]^2 = \frac{2}{x(0)} \sum_{k \neq 0} x^2\left(k\tau\right). \tag{15}$$

The channel log-likelihood ratio (LLR) is the only piece of information available initially at the receiver. This can provide an initial estimate of the probabilities for $a_n$ as

$$\Pr[a_n = +1|y_n] = \frac{e^{L_n^{ch}}}{1 + e^{L_n^{ch}}} \text{ and } \Pr[a_n = -1|y_n] = \frac{1}{1 + e^{L_n^{ch}}}. \tag{16}$$

Based on these probabilities the mean value and the variance of $a_n$ can be derived as

$$\mu_n = \frac{e^{L_n^{ch}} - 1}{e^{L_n^{ch}} + 1} \tag{17}$$

and

$$\sigma_n^2 = \frac{4e^{L_n^{ch}}}{\left(e^{L_n^{ch}}+1\right)^2}. \tag{18}$$

The soft-output ISI detector can then use the mean values $\mu_n$ to cancel the interference

$$\tilde{y}_n = y_n - \sum_{k\neq n} \mu_k \, x\left((n-k)\tau\right) \tag{19}$$

and the variance values $\sigma_n^2$ to scale the soft-output log-likelihood ratios

$$L_n^{det} = 4\frac{E_s}{N_0+\tilde{\rho}}\tilde{y}_n, \tag{20}$$

where

$$\tilde{\rho} = \frac{2}{x(0)}\sum_{k\neq 0} x^2\left(k\tau\right)\sigma_k^2. \tag{21}$$

The soft-output LLR $L_n^{det}$ is then deinterleaved and passed to the channel decoder. The soft-output of the channel decoder can then be interleaved and the output of the interleaver $L_n^{dec}$ can be passed back to the ISI detector to improve the ISI cancelation by adding the extrinsic part of this LLR $L_n^{ap}$ to $L_n^{ch}$ in (16), where $L_n^{ap}$ is given from

$$L_n^{ap} = L_n^{dec} - L_n^{det}. \tag{22}$$

The block diagram of the turbo equalizer is shown in Fig. 6.

The performance of this turbo equalizer has been simulated in Fig. 7 using the rate 1/2 [5,7] convolutional code and an s-random 4000-bit interleaver. The convolutional code decoder is the standard BCJR decoder [26]. The lower bound for the performance without ISI shown in Fig. 7 is based on the minimum Hamming distance $d_{H,min}$ of the [5,7] convolutional code, i.e.,

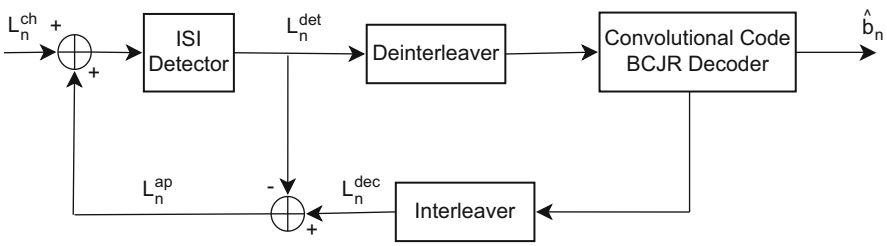


**Fig. 6** Block diagram of the turbo equalizer used for coded Faster-than-Nyquist

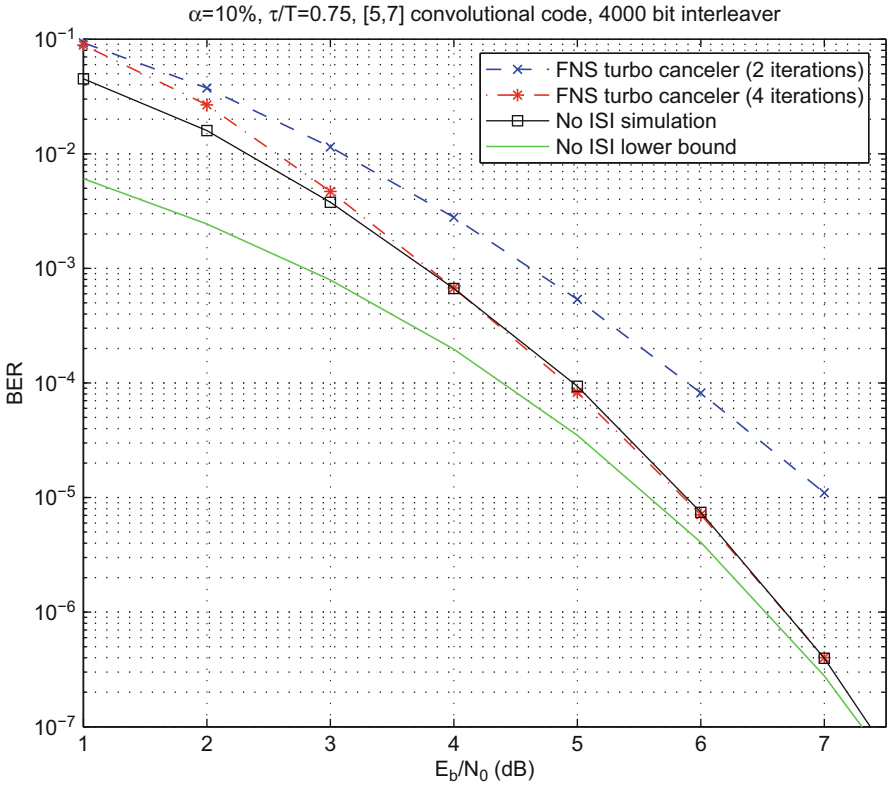


**Fig. 7** Performance of convolutionally coded Faster-than-Nyquist signaling for the raised-cosine pulses with roll-off $\alpha = 10\,\%$ with bandwidth efficiency 1.21 bits/s/Hz

$$P_e \geq Q\left(\sqrt{rd_{H,min}\frac{2E_b}{N_0}}\right), \tag{23}$$

where $r = 1/2$ is the rate of the [5,7] convolutional code and $d_{H,min} = 5$.

A similar FTN interference cancelation approach with a turbo code instead of a convolutional code was used for one of the receivers in [27].

Another approach with both serial and parallel concatenation of convolutional codes [28] used the more complex ISI detector from [29] and truncated the number of states. The serial concatenation used a rate-1 precoder, whose decoding was performed in the same BCJR as the ISI detector without any increase in complexity, and resulted in improved performance compared to not including the precoder.

The reduced complexity BCJR from [29] can also be used for binary and nonbinary transmitted alphabets and was combined with LDPC codes in [30]. In [30] it is found that it is beneficial to increase spectrum efficiency by keeping the code rate constant and signaling at higher rates than by using higher code rates. In addition, for limited ISI detector complexity, which means a reduced number of states in the ISI detector, there is a spectrum efficiency level, about 1.4 bits/s/Hz for the parameters used in [30], at which LDPC-coded FTN 16QAM starts outperforming LDPC-coded FTN QPSK.

Frequency-domain equalization generating soft output for a serial concatenation of convolutional codes was proposed in [31]. The rate-1 precoder in this case was followed by a second interleaver showing significantly improved performance compared to not including the precoder and the second interleaver.

## 5 Achievable Information Rates

The ultimate limit of the FTN gains is determined by the channel capacity based on Shannon's capacity formula [32]

$$C = W\log_2\left(1 + \frac{P}{WN_0}\right), \tag{24}$$

where $P$ is the transmit signal power and $W$ is the signal bandwidth. Shannon's formula assumes that the ideal sinc pulses are used in transmission.

However, in practical systems the roll-off of the pulse bandwidth causes some of this capacity to remain unused, as for the rolloff to be contained within the same bandwidth the signaling rate needs to be reduced. For example, for Nyquist transmission with a root raised cosine pulses of roll-off $\alpha$, the achievable information rate becomes:

$$I_{\text{Nyquist},\alpha} = \frac{1}{2T}\log_2\left(1 + \frac{2PT}{N_0}\right), \tag{25}$$

where $T$ is the signaling rate and $W = (1+\alpha)/2T$.

The last two equations are compared in Fig. 8, which shows that taking into account the pulse roll-off reduces the achievable information rate when maintaining the Nyquist orthogonality constraint and the transmit signal power $P$. The larger the roll-off the larger this information rate loss. Note that even though this constrained information rate is affected by the pulse roll-off, it does not take into account the actual pulse shape. The information rate (25) is the same even if different pulses with the same roll-off were used, e.g., triangular.

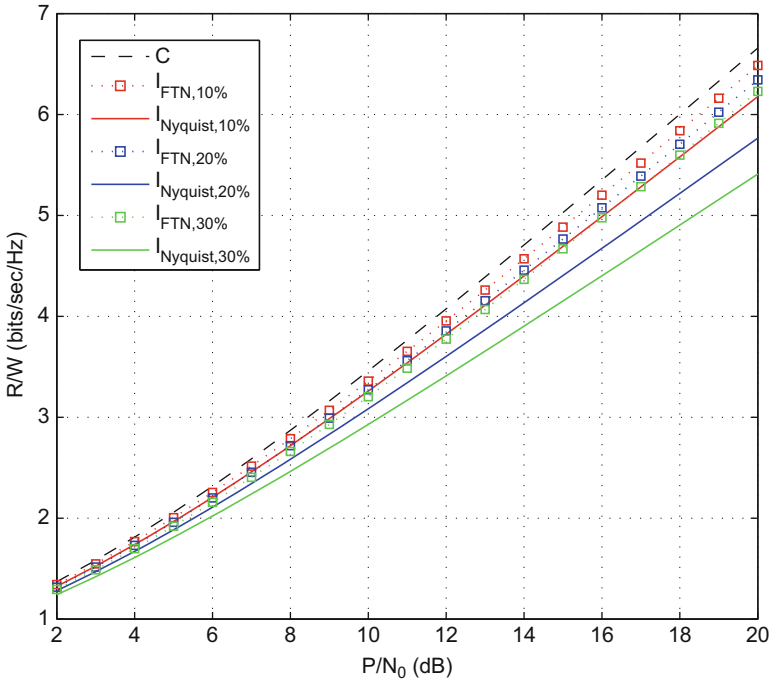


**Fig. 8** Achievable information rates for faster-than-Nyquist signaling and Nyquist signaling with roll-off $\alpha = 10, 20, 30\,\%$

FTN can be viewed as a method to regain some of the lost information rate by taking advantage of the complete pulse shape of the signal. It was shown in [33] that the achievable information rate with FTN using independently and identically distributed (i.i.d.) input is

$$I_{\mathrm{FTN},\alpha} = \int_0^W \log_2\left(1 + \frac{2P}{N_0}|G(f)|^2\right) df, \tag{26}$$

where $G(f)$ is the Fourier transform of the transmit pulse shape $g(t)$ with $\int_{-W}^{W} |G(f)|^2 df = 1$. This information rate can be achieved by FTN when $\tau \geq \frac{1}{2W} = \frac{T}{1+\alpha}$ [16, 33], i.e., by signaling fast enough to make up for the non-zero roll-off factor. The main observation from (26) is that at high SNR the weaker parts of the spectrum can also contribute noticeably to the achievable information rate.

Figure 8 shows how the FTN information rate with raised cosine pulses compares to the Nyquist information rate for different roll-off. For zero roll-off, i.e., when FTN also uses the sinc pulse, it was proven in [34] that the FTN information rate equals the Shannon capacity.

Figure 8 establishes a clear advantage of FTN signaling in practical systems, which employ non-zero roll-off transmit pulses. This comparison is based on Gaussian input signals. Since we signal faster than the Nyquist rate, the assumption that the transmitted pulses have to be orthogonal at the Nyquist rate is no longer relevant. Dropping this assumption [35] further bridged the gap between the channel capacity $C$ in Fig. 8 and the FTN information rate at high SNR by optimizing with respect to the pulse shape.

For more practical input alphabets the constrained input information rate has so far been established with lower-bounds. These lower-bounds are based on constraining the receiver structure, such as symbol-by-symbol detection [25] or MMSE-DFE detection [36], and finding the achievable rate of this suboptimal receiver.

## 6 Frequency Packing

Bringing the pulses closer together in time, as described previously, is also called time packing. Another approach to improving the bandwidth efficiency is with frequency packing in the context of multicarrier modulation. In frequency packing the neighboring carriers start overlapping generating inter-carrier interference, but as in the case of time packing, less overall bandwidth is used resulting in improved bandwidth efficiency.

The transmitted multicarrier signal is expressed as follows:

$$s(t) = \sum_{m} \left[ \sum_{k} a_{m,k} g(t - k\tau) \right] e^{-j2\pi mFt}, \tag{27}$$

where $F$ is the frequency spacing of the carriers.

Starting with the filter bank multicarrier approach in [37], one can require that the frequency content of neighboring subcarriers does not overlap to guarantee that no intercarrier interference (ICI) occurs. This frequency spacing constraint together with the Nyquist time constraint are together called the generalized Nyquist constraints [37].

Following the time approach, one can start bringing the subcarriers closer together causing them to overlap and cause ICI but becoming more bandwidth efficient. Such approaches were introduced in [13, 38] as a method to increase the OFDM bandwidth efficiency as shown in Fig. 9.

It is possible to further improve bandwidth efficiency in (27) by allowing both time and frequency packing and optimizing the trade-off between them [14]. This yields additional gains in the sense that the two-dimensional time-frequency limit, where the minimum distance between transmitted signals becomes smaller than the minimum distance for the orthogonal time-frequency transmission, extends to higher spectral efficiency than in time-packing or frequency-packing alone.

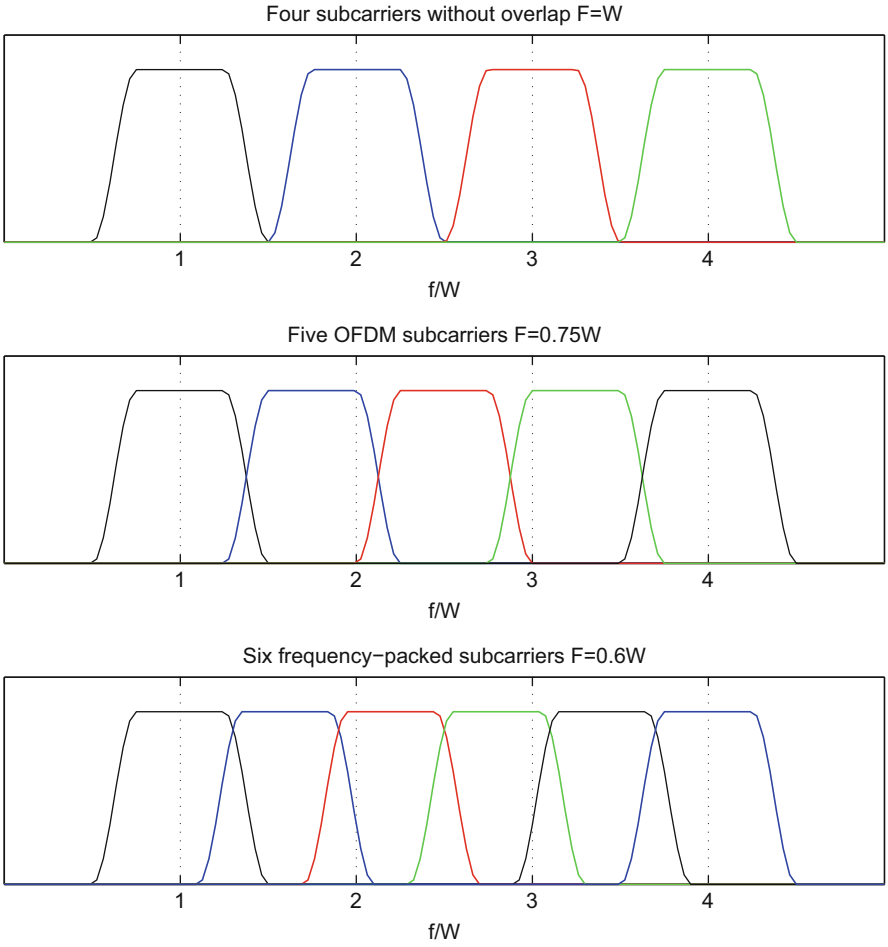


**Fig. 9** Subcarriers in multicarrier transmission when the subcarrier spectrum does not overlap (*top*), in OFDM (*middle*), and in the frequency packed approach. $W$ is the bandwidth of each subcarrier and in this example all subcarriers fit within a total bandwidth of $4W$

Orthogonal multicarrier modulation, such as OFDM, is more bandwidth efficient than single carrier modulation, because the unused roll-off is only that of the subcarriers at the edges of the whole band. Therefore, the information rate gains of single carrier modulation are more limited in multicarrier systems [15, 39], but combining both time and frequency offers other advantages, mainly additional flexibility in pulse shape design and rate adaptivity and use in the single carrier uplink [15] as well as less overhead and better latency in other 5G applications [40]. These advantages of FTN have made it an alternative to the time and frequency orthogonality in the 5G standard [15].

The same low complexity interference canceler that was described in the Turbo Equalization section earlier can be used to mitigate the effects of both the inter-symbol and intercarrier interference from the neighboring symbols. Such schemes approach the coded performance without interference with both convolutional codes [39] and LDPC codes [25]. Similar interference cancelation approaches dealing with intercarrier interference when only frequency packing is used can perform about 1dB worse than coded OFDM at the same bit rate with the frequency-packed signal using 60 % of the OFDM bandwidth [41].

## 7 Discussion

In this section we outline some of the recent research activity in directions relevant to modern communication systems.

Rusek [42] proved that the AWGN $d_{min}$ Mazo limit can be extended to the MIMO case and also reported results on the fading channel. Recent work on the broadcast channel suggests that FTN signaling is also an alternative approach to achieving the Gaussian input capacity for two users [43]. A turbo-coded broadcast scheme was introduced approaching this capacity [27]. An alternative uncoded scheme with more than two users and detecting all users' signals was proposed in [44]. Very promising sum capacity gains with FTN signaling were also reported for the synchronous and asynchronous multiple access channels in [45, 46] but no receiver design has been proposed yet to take advantage of these gains. This multiuser work focuses only on the classical time-packed FTN signaling.

Practical directions of faster-than-Nyquist signaling have been investigated recently with hardware implementations [5, 47] considering several aspects of the algorithms involved in an FTN system. The focus of these implementations has been on multicarrier systems as these seem to be more relevant to 5G communication systems. Other potential benefits of FTN signaling in practical systems, such as the peak-to-average power ratio (PAPR), have been explored in [48].

## 8 Conclusion

As the increasing amount of recently published FTN work shows, FTN signaling has been a rather active research area in the last few years because of its more efficient use of spectrum resources. There are several open questions that remain unanswered, especially with multiple transmit and receive antennas, in multiuser systems and in practical implementations, and therefore, the associated research activity is expected to intensify in the coming years.

**Acknowledgements** The authors would like to thankfully acknowledge support from Texas Instruments for their original work on faster-than-Nyquist signaling.

## References


1. J.E. Mazo, Faster-than-Nyquist signaling. Bell Syst. Tech. J. **54**, 1451–1462 (1975)
2. G.D. Forney, Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference. IEEE Trans. Inf. Theory **IT-18**, 363–378 (1972)
3. G. Ungerboeck, Adaptive maximum-likelihood receiver for carrier-modulated data-transmission systems. IEEE Trans. Commun. **COM-22**, 624–636 (1974)
4. G.J. Foschini, Contrasting performance of faster binary signaling with QAM. Bell Syst. Tech. J. **63**, 1419–1445 (1984)
5. D. Dasalukunte, V. Owall, F. Rusek, J.B. Anderson, *Faster than Nyquist Signaling: Algorithms to Silicon* (Springer, Cham, 2014)
6. J. Anderson, S. Mohan, Sequential coding algorithms: a survey and cost analysis. IEEE Trans. Commun. **32**, 169–176 (1984)
7. M.V. Eyuboglu, S.U.H. Qureshi, Reduced-state sequence estimation with set partitioning and decision feedback. IEEE Trans. Commun. **COM-36**, 13–20 (1988)
8. A. Duel-Hallen, C. Heegard, Delayed decision-feedback sequence estimation. IEEE Trans. Commun. **COM-37**, 428–436 (1989)
9. K. Balachandran, J.B. Anderson, Reduced complexity sequence detection for nonminimum phase intersymbol interference channels. IEEE Trans. Inf. Theory **43**, 275–280 (1997)
10. A. Hafeez, W.E. Stark, Decision feedback sequence estimation for unwhitened ISI channels with applications to multiuser detection. IEEE J. Select. Areas Commun. **JSAC-16**, 1785–1795 (1998)
11. C. Berrou, A. Glavieux, P. Thitimajshima, Near Shannon limit error-correcting coding and decoding: turbo-codes, in *Proceedings of ICC '93*, vol. 2 (1993), pp. 1064–1070
12. C. Douillard, M. Jezequel, C. Berrou, A. Picart, P. Didier, A. Glavieux, Iterative correction of intersymbol interference: turbo-equalization. Eur. Trans. Telecommun. **6**, 507–511 (1995)
13. M. Rodrigues, I. Darwazeh, A spectrally efficient frequency division multiplexing based communication system. In: *8th International OFDM-Workshop, Hamburg, Germany*, pp. 70–74 (2003)
14. F. Rusek, J.B. Anderson, The two dimensional Mazo limit, in *Proceedings of ISIT '05* (2005), pp. 970–974
15. P. Banelli, S. Buzzi, G. Colavolpe, A. Modenini, F. Rusek, A. Ugolini, Modulation formats and waveforms for 5G networks: Who will be the heir of OFDM?: An overview of alternative modulation schemes for improved spectral efficiency. IEEE Signal Process. Mag. **31**, 80–93 (2014)
16. A. Modenini, F. Rusek, G. Colavolpe, Faster-than-Nyquist signaling for next generation communication architectures, in *2014 Proc. 22nd Europ. Signal Proc. Conf. (EUSIPCO)* (2014), pp. 1856–1860
17. M. Secondini, T. Foggi, F. Fresi, G. Meloni, F. Cavaliere, G. Colavolpe, E. Forestieri, L. Poti, R. Sabella, G. Prati, Optical time-frequency packing: principles, design, implementation, and experimental demonstration. J. Lightwave Technol. **33**(17), 3558–3570 (2015)
18. J.G. Proakis, *Digital Communications*, 4th edn. (McGraw-Hill, New York, 2000)
19. A.D. Liveris, C.N. Georghiades, Exploiting faster-than-Nyquist signaling. IEEE Trans. Commun. **51**, 1502–1511 (2003)
20. J.E. Mazo, H.J. Landau, On the minimum distance problem for faster-than-Nyquist signaling. IEEE Trans. Inf. Theory **IT-34**, 1420–1427 (1988)
21. D. Hajela, On computing the minimum distance for faster than Nyquist signaling. IEEE Trans. Inf. Theory **IT-36**, 289–295 (1990)
22. G.D. Forney, Lower bounds on error probability in the presence of large intersymbol interference. IEEE Trans. Commun. **COM-20**, 76–77 (1972)
23. F. Rusek, J.B. Anderson, Non binary and precoded faster than Nyquist signaling. IEEE Trans. Commun. **56**, 808–817 (2008)

24. A. Prlja, J.B. Anderson, F. Rusek, Receivers for faster-than-Nyquist signaling with and without turbo equalization, in *Proc. IEEE Int. Symp. Inf. Theory* (2008), pp. 464–468
25. A. Barbieri, D. Fertonani, G. Colavolpe, Time-frequency packing for linear modulations: spectral efficiency and practical detection schemes. IEEE Trans. Commun. **COM-57**, 2951–2959 (2009)
26. L.R. Bahl , J. Cocke , F. Jeinek, J. Raviv, Optimal decoding of linear codes for minimizing symbol error rate. IEEE Trans. Inf. Theory **20**, 284–287 (1974)
27. Y.J.D. Kim, J. Bajcsy, An architecture for faster than Nyquist turbo broadcasting, in *Proc. 7th Int. Symp. on Turbo Codes and Iterative Inform.* (2012), pp. 170–174
28. F. Rusek, J.B. Anderson, Serial and parallel concatenations based on faster than Nyquist signaling, in *Proceedings of ISIT '06* (2006), pp. 1993–1997
29. G. Colavolpe, A. Barbieri, On MAP symbol detection for ISI channels using the Ungerboeck observation model. IEEE Commun. Lett. **9**, 720–722 (2005)
30. J. Yu, J. Park, F. Rusek, B. Kudryashov, I. Bocharova, High order modulation in faster-than-Nyquist signaling communication systems, in *Proceedings of 2014 IEEE Vehicular Technology Conference (VTC Fall)* (2014), pp. 1–5
31. S. Sugiura, L. Hanzo, Frequency-domain-equalization-aided iterative detection of faster-than-Nyquist signaling. IEEE Trans. Veh. Technol. **64**, 2122–2128 (2015)
32. C.E. Shannon, Communication in the presence of noise. Proc. IRE **37**, 10–21 (1949)
33. F. Rusek, J.B. Anderson, Constrained capacities for faster-than-Nyquist signaling. IEEE Trans. Inf. Theory **55**, 764–775 (2009)
34. A.D. Liveris, On Distributed Coding, Quantization of Channel Measurements and Faster-than-Nyquist Signaling, Ph.D. thesis, Texas A&M University, 2004
35. J. Zhou, D. Li, X. Wang, Generalized faster-than-Nyquist signaling, in *IEEE Int. Symp. Inform. Theory* (2012), pp. 1–5
36. Y.G. Yoo, J.H. Cho, Asymptotic optimality of binary faster-than-Nyquist signaling. IEEE Commun. Lett. **14**, 788–790 (2010)
37. B. Farhang-Boroujeny, OFDM versus filter bank multicarrier. IEEE Signal Process. Mag. **28**, 92–112 (2011)
38. I. Kanaras, A. Chorti, M.R.D. Rodrigues, I. Darwazeh, Spectrally efficient FDM signals: bandwidth gain at the expense of receiver complexity, in *IEEE Int. Conf. Commun. (ICC)* (2009), pp. 1–6
39. J.B. Anderson, F. Rusek, V. Owall, Faster-than-Nyquist signaling. Proc. IEEE **101**, 1817–1830 (2013)
40. G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. Brink, I. Gaspar, M. Michailow, A. Festag, L. Mendes, N. Cassiau, D. Ktenas, M. Dryjanski, S. Pietrzyk, B. Eged, P. Vago, F. Wiedmann, 5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications. IEEE Commun. Mag. **52**, 97–105 (2014)
41. T. Xu, I. Darwazeh, Spectrally efficient FDM: spectrum saving technique for 5G?, in *Int. Conf. 5G for Ubiquitous Connectivity (5GU)* (2014), pp. 273–278
42. F. Rusek, On the existence of the Mazo-limit on MIMO channels. IEEE Trans. Wirel. Commun. **8**, 1118–1121 (2009)
43. Y.J.D. Kim, J. Bajcsy, Faster than Nyquist broadcast signaling, in *Proc. 26th Bienn. Symp. Commun., Kingston* (2012), pp. 186–189
44. M.J. Abdoli, M. Jia, Trellis decoding for multi-user faster-than-Nyquist transmission, in *IEEE Veh. Tech. Conf. (VTC Fall)* (2014), pp. 1–5
45. Y. Feng, J. Bajcsy, On faster-than-Nyquist transmission over a Multiple-Access Channel, in *IEEE Mil. Commun. Conf. (MILCOM)* (2014), pp. 824–829
46. Y. Feng, J. Bajcsy, Improving throughput of faster-than-Nyquist signaling over Multiple-Access Channels, in *IEEE Veh. Tech. Conf. (VTC Spring)* (2015), pp. 1–5
47. P.N. Whatmough, M.R. Perrett, S. Isam, I. Darwazeh, VLSI architecture for a reconfigurable spectrally efficient FDM baseband transmitter. IEEE Trans. Circuits Syst. I Regul. Pap. **59**, 1107–1118 (2012)
48. C. Le, M. Schellmann, M. Fuhrwerk, J. Peissig, On the practical benefits of faster-than-Nyquist signaling, in *Int. Conf. Advanced Techn. for Commun. (ATC)* (2014), pp. 208–213

# Generalized Frequency Division Multiplexing: A Flexible Multi-Carrier Waveform for 5G

**Maximilian Matthé, Ivan Simões Gaspar, Luciano Leonel Mendes, Dan Zhang, Martin Danneberg, Nicola Michailow, and Gerhard Fettweis**

**Abstract** The next generation of wireless networks will face different challenges from new scenarios. The conventional Orthogonal Frequency Division Multiplexing (OFDM) has shown difficulty in fulfilling all demanding requirements. This chapter presents Generalized Frequency Division Multiplexing (GFDM) as a strong waveform candidate for future wireless communications systems which can be combined with several techniques such as precoding or Offset Quadrature Amplitude Modulation (OQAM) and which offers the flexibility to emulate a variety of other popular waveforms as corner cases. This property suggests GFDM as a key technology to allow reconfiguration of the physical layer (PHY), enabling a fast and dynamic evolution of the infrastructure. Additionally, multicarrier transmission theory is covered in terms of Gabor theory. Details on synchronization, channel estimation algorithms and MIMO techniques for GFDM are presented and a description of a proof-of-concept demonstrator shows the suitability of GFDM for future wireless networks.

## 1 GFDM: A General Solution for 5G

The fifth generation (5G) of mobile cellular systems needs to deal with a wide range of services, each of which can have very different requirements. For instance, the four main scenarios illustrated in Fig. 1 respectively require loose synchronization for Internet of Things (IoT); low latency for Tactile Internet; reliable, efficient, and robust high throughput for bitpitpe communication; high coverage and dynamic spectrum allocation with low out-of-band (OOB) emission and Cognitive Radio (CR) techniques for Wireless Regional Area Network (WRAN) applications.

M. Matthé (✉) • I.S. Gaspar • D. Zhang • M. Danneberg • N. Michailow • G. Fettweis
Vodafone Chair Mobile Communication Systems, Technische Universität Dresden, Dresden, Germany
e-mail: maximilian.matthe@ifn.et.tu-dresden.de; ivan.gaspar@ifn.et.tu-dresden.de; dan.zhang@ifn.et.tu-dresden.de; martin.danneberg@ifn.et.tu-dresden.de; nicola.michailow@ifn.et.tu-dresden.de; gerhard.fettweis@ifn.et.tu-dresden.de

L.L. Mendes
Instituto Nacional de Telecomunicações, Santa Rita do Sapucaí, MG, Brazil
e-mail: luciano@inatel.br



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_9

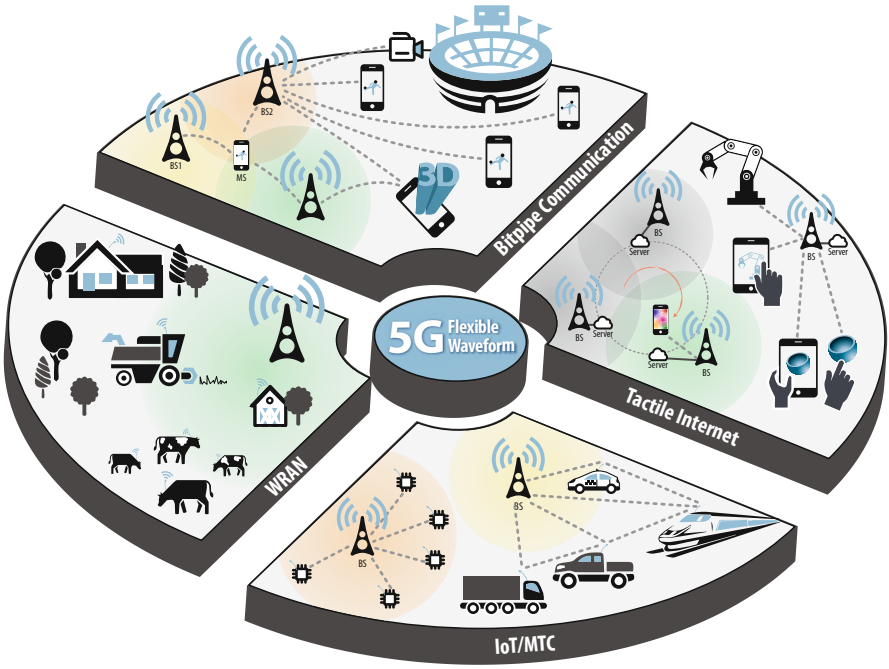


**Fig. 1** Four of the main scenarios envisioned for 5G networks

A physical layer (PHY) with unprecedented flexibility will be of great benefit to users such that various applications can be efficiently and effectively addressed by proper reconfiguration. Certainly, all these scenarios will benefit from multiple-input multiple-output (MIMO)-techniques, such as increased data rate, enhanced robustness and the additional degree of freedom in space for multi-user and multi-cell interference management. In order to maximize MIMO profits, the 5G PHY must also be able to cope with multi-antenna interference and related algorithms need to be implemented with an affordable complexity. 5G related topics have been intensely researched lately and several projects were funded by the European Union [1], e.g. METIS and 5GNOW.

Instead of selecting one specific waveform for each 5G scenario, it is much more desirable to adopt a single flexible waveform that can be easily reconfigured to address a multitude of applications. More importantly, such a *software-defined waveform (SDW)* builds a foundation at PHY for preparing the paradigm shift towards software-defined virtualization. Namely, the SDW shall be generated in programmable hardware, based on Application Specific Integrated Processors (ASIP), field programmable gate array (FPGA), or software with manageable cost. By means of a cost effective approach that exposes the time-frequency resource grid and waveform engineering capabilities to software, it becomes feasible to customize PHY such that it can be seen as a virtual service for upper layers.

In this context, GFDM [2] provides a very flexible time-frequency structure that favors software exposition, controlling, and virtualization. Starting from introducing some basics of GFDM, the main goal of this section is to explore the GFDM flexibility to achieve a framework that covers all major waveform candidates for the different 5G scenarios. Furthermore, we will show that the main waveforms considered for 5G are, in fact, corner cases for GFDM. Also, the flexibility offered by GFDM is sufficient and necessary to fulfill diverse quality of service (QoS) requirements in 5G.

The remainder of this chapter is organized as follows: This section introduces the basics of the GFDM system, including modulation and demodulation equations, waveform engineering, precoding and application of OQAM modulation. The theoretic foundation for multicarrier systems in terms of Gabor theory is described in the subsequent Sect. 2. Section 3 covers development of synchronization and channel estimation algorithms for GFDM in detail whereas Sect. 4 is dedicated to a detailed description of MIMO techniques for GFDM. Finally, Sect. 5 considers implementation aspects for a real-world system demonstrator and the chapter is concluded in Sect. 6.

## 1.1 Basics

GFDM has a block structure where $N = KM$ complex-valued data symbols $d_{k,m}$ are transmitted on $K$ subcarriers and $M$ time-slots (or subsymbols). Each subcarrier is pulse-shaped by a filter impulse response with $S = RT$ samples, where $T$ is the number of periods of the filter and $R$ is the number of samples per period. The use of a circular pulse shaping filter is a key feature of GFDM when it was initially proposed [3]. Its advantages include (1) keeping the GFDM signal compact in time; and (2) enabling low complexity signal processing, e.g. [4, 5]. Denoting modulo $S$ operation as $\langle\rangle_s$, the GFDM transmit sequence is given by

$$x[n] = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} d_{k,m} g\left[\langle n - mP \rangle_S\right] \exp\left(j2\pi \frac{kQ}{S} n\right), \quad n = 0, 1, \ldots, S-1, \tag{1}$$

where $P$ is the spacing between adjacent subsymbols and $Q$ is the distance between adjacent subcarriers. The subsymbol and subcarrier spacing factors are respectively given by

$$\begin{aligned} \nu_{\mathrm{t}} &= \frac{P}{R} \\ \nu_{\mathrm{f}} &= \frac{Q}{T}, \end{aligned} \tag{2}$$

**Table 1** Terminology

| Variable | Meaning |
|---|---|
| $R$ | Samples per period in the filter |
| $T$ | Periods in the filter |
| $S = RT$ | Total number of samples in the signal |
| $P$ | Subsymbol spacing in time domain |
| $Q$ | Subcarrier spacing in frequency domain |
| $\nu_\mathrm{t} = P/R$ | Subsymbols distance factor |
| $\nu_\mathrm{f} = Q/T$ | Subcarriers distance factor |
| $K = RT/Q = R/\nu_\mathrm{f} = S/Q$ | Subcarriers per block |
| $M = TR/P = T/\nu_\mathrm{t} = S/P$ | Subsymbols per block |
| $N = KM$ | Number of data symbols per block |

which allows us to rewrite (1) as

$$x[n] = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} d_{k,m} g\left[\langle n - m\,\nu_t R\rangle_S\right] \exp\left(j2\pi \frac{k\,\nu_f T}{S} n\right), \quad n = 0, 1, \ldots, S-1. \tag{3}$$

When $\nu_t < 1$ the overlapping between the subsymbols increases, leading to a higher intersymbol interference (ISI) and better spectrum efficiency. For $\nu_t > 1$, the subsymbols are taken apart from each other, reducing the time overlapping and decreasing the spectrum efficiency. The same reasoning is valid for $\nu_f$, but now in the frequency domain. Table 1 summarizes the parameters used to describe the GFDM signal.

An interesting configuration is achieved when $\nu_t = \nu_f = 1$, which will be called "critically sampled GFDM". In this case, the GFDM sequence can be written as

$$x[n] = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} d_{k,m} g\left[\langle n - mK\rangle_N\right] \exp\left(j2\pi \frac{k}{K} n\right), \quad n = 0, 1, \ldots, N-1 \tag{4}$$

where Fig. 2 depicts the block diagram of such GFDM modulation. A cyclic prefix (CP) is added to protect the GFDM sequence from the multipath channel. Assuming that the CP length is larger than the channel impulse response, the received signal after CP removal (see Fig. 3) is given by

$$y[n] = x[n] \circledast h[n] + w[n], \tag{5}$$

where $\circledast$ denotes the circular convolution operation, $h[n]$ is the channel impulse response and $w[n]$ is the additive white Gaussian noise (AWGN) with variance $\sigma_w^2$. After synchronization and channel estimation, a frequency-domain equalizer (FDE) is used to compensate the effects of the multipath channel, leading to a equalized

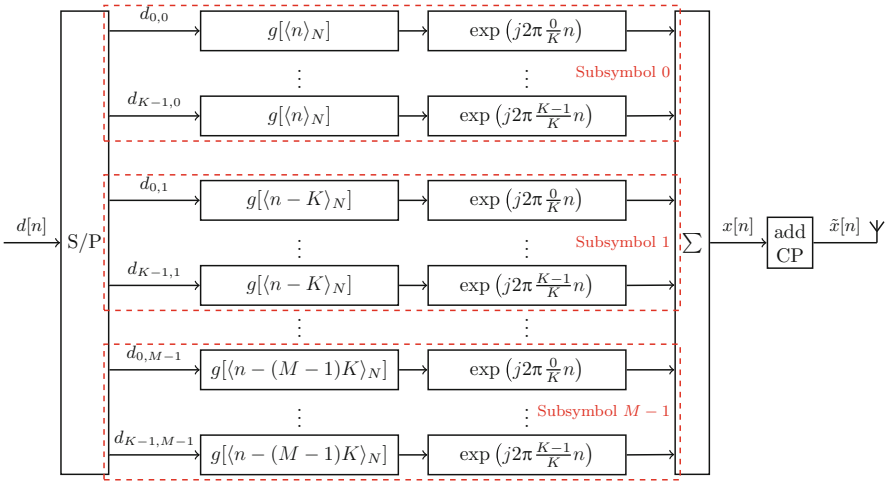


**Fig. 2** Block diagram of the GFDM modulator

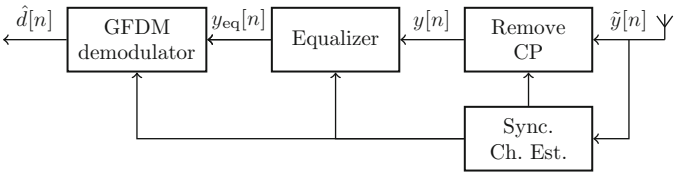


**Fig. 3** Block diagram of the GFDM receiver

received sequence $y_{\text{eq}}[n]$. A set of filters is derived from a prototype receive filter $\gamma[n]$ as

$$\gamma_{k,m} = \gamma[\langle -n \rangle_N] \exp(-j2\pi \tfrac{k}{K} n), \tag{6}$$

and the recovered data symbols are given by

$$\hat{d}_{k,m} = \gamma^*_{k,m} \circledast y_{\text{eq}}[n]|_{n=mK}. \tag{7}$$

Different receive filters can be used to recover the data symbols and matrix notation can be useful to design linear receivers for GFDM. The column vectors containing the samples of the transmit filters are given by

$$[\mathbf{g}_{k,m}]_n = g_{k,m}[n], \tag{8}$$

and the modulation matrix is arranged by

$$\mathbf{A} = [\mathbf{g}_{0,0}\ \mathbf{g}_{1,0} \ldots \mathbf{g}_{K-1,0}\ \mathbf{g}_{0,1} \ldots \mathbf{g}_{K-1,M-1}]. \tag{9}$$

Let us further organize the $KM$ data symbols of one GFDM block into a $(K \times M)$ data matrix, given by

$$\mathbf{D} = \begin{bmatrix} d_{0,0} & d_{0,1} & \dots & d_{0,M-1} \\ d_{1,0} & d_{1,1} & \dots & d_{1,M-1} \\ \vdots & \vdots & \ddots & \vdots \\ d_{K-1,0} & d_{K-1,1} & \dots & d_{K-1,M-1} \end{bmatrix} = \begin{bmatrix} \mathbf{d}_{\mathrm{c},0} & \mathbf{d}_{\mathrm{c},1} & \dots & \mathbf{d}_{\mathrm{c},M-1} \end{bmatrix} = \begin{bmatrix} \mathbf{d}_{\mathrm{r},0} \\ \mathbf{d}_{\mathrm{r},1} \\ \vdots \\ \mathbf{d}_{\mathrm{r},K-1} \end{bmatrix} \tag{10}$$

Then, the GFDM transmit vector is given by

$$\mathbf{x} = \mathbf{A}\mathbf{d}, \tag{11}$$

where the data vector $\mathbf{d}$ is obtained by stacking the columns of $\mathbf{D}$ one after another

$$\mathbf{d} = \begin{bmatrix} \mathbf{d}_{\mathrm{c},0}^{\mathrm{T}} & \mathbf{d}_{\mathrm{c},1}^{\mathrm{T}} & \dots & \mathbf{d}_{\mathrm{c},M-1}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}. \tag{12}$$

The prototype filter used on the transmit side can lead to a non-orthogonal system, which means that ISI and intercarrier interference (ICI) can be introduced among the data symbols by the modulation process. The GFDM demodulator must be able to tackle this self-interference to minimize the impact on the symbol error rate (SER) performance. After FDE, the demodulation process can be performed as

$$\hat{\mathbf{d}} = \mathbf{B}\mathbf{y}_{\mathrm{eq}}, \tag{13}$$

where $\mathbf{B}$ is the demodulation matrix. Different approaches can be used to demodulate the GFDM signal based on the transmit matrix, as presented in Table 2. The Matched filter (MF) demodulator maximizes the signal-to-noise ratio (SNR), but it is not able to remove the self-interference. The zero-forcing (ZF) demodulator can remove the interference, but it enhances the noise. The minimum mean square error (MMSE) demodulator achieves a good balance between MF and ZF by taking the SNR into account. Also, the MMSE incorporates the channel matrix $\mathbf{H}$, which means that the FDE is not necessary prior the demodulation in this case.

Figure 4 shows two possible configurations of GFDM and the corresponding OFDM counterpart. Configuration (a) concatenates several OFDM symbols, while keeping the subcarrier spacing and symbol duration. Hence, the multitude of CPs of the several OFDM symbols can shrinked to one single CP for the GFDM signal. However, as the number and distance of subcarriers is kept, the PAPR for

**Table 2** Linear demodulation matrices

| Type | Definition |
|---|---|
| MF | $\mathbf{B}_{\mathrm{MF}} = \mathbf{A}^{\mathrm{H}}$ |
| ZF | $\mathbf{B}_{\mathrm{ZF}} = \mathbf{A}^{-1}$ |
| MMSE | $\mathbf{B}_{\mathrm{MMSE}} = (\mathbf{R}_w + \mathbf{A}^{\mathrm{H}}\mathbf{H}^{\mathrm{H}}\mathbf{H}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{H}}\mathbf{H}^{\mathrm{H}}$ |

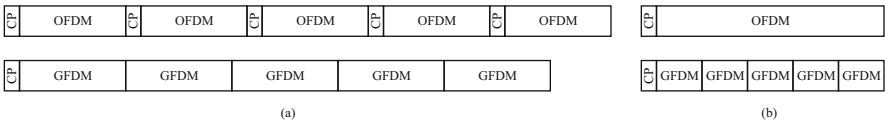


**Fig. 4** Different configurations for GFDM with relation to equivalent OFDM system. (**a**) Configuration suitable for downlink, where GFDM and OFDM have equal number of subcarriers and same bandwidth. The CP overhead of OFDM is saved by GFDM. (**b**) Configuration suitable for uplink, where GFDM has fewer but wider subcarriers than OFDM. Using such configuration, GFDM can achieve lower PAPR than OFDM. One special case is to apply a Dirichlet kernel for pulse shaping, emulating SC-FDMA

**Table 3** Comparison of implementation complexity for GFDM equivalent OFDM, with configurations (a) and (b) referring to Fig. 4

| | Complex multiplications per TX symbol | |
|---|---|---|
| | Transmitter | Receiver (incl. FDE) |
| GFDM | $\frac{1}{KM}(\frac{KM}{2}(M+\log_2 K))$ | $\frac{1}{KM}(\frac{KM}{2}(M+3\log_2 K)+KM)$ |
| OFDM (a) | $\frac{1}{KM}\frac{KM}{2}\log_2 K$ | $\frac{1}{KM}(\frac{KM}{2}\log_2 K+KM)$ |
| OFDM (b) | $\frac{1}{KM}\frac{KM}{2}\log_2 KM$ | $\frac{1}{KM}(\frac{KM}{2}\log_2 KM+KM)$ |

Values taken from [6]

this configuration is as high as for OFDM, making this configuration preferable for a downlink transmission from the base station. On the other hand, PAPR can be reduced by increasing the subcarrier bandwidth and reducing the number of subcarriers. This configuration is shown in Fig. 4b, where one GFDM block is designed to have the same length as one OFDM symbol. Hence, the subcarriers become wider and the block contains fewer subcarriers. Accordingly the PAPR reduces, similar to the principle of SC-FDMA, where several OFDM subcarriers are combined to a single channel for the SC-FDMA system. Hence, this configuration is suitable for uplink transmission.

In the literature, several algorithms for a reduced complexity implementation of the GFDM modulator and demodulator are given [4–8]. Table 3 shows the number of required complex multiplications for current proposals for implementation of GFDM, compared to the FFT/IFFT approach of OFDM. As shown, the additionally to FFT/IFFT operations for frequency domain conversion and FDE, the GFDM implementation requires an extra amount of roughly $M/2$ multiplications per data symbol which accounts for the overlapping of subsymbols in time or frequency domain.

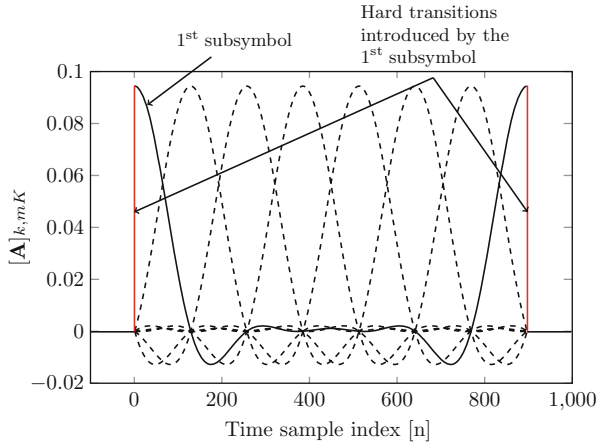


**Fig. 5** The first subsymbol of the GFDM block introduces abrupt variations in the time-domain signal, which leads to high OOB emissions

## *1.2 Waveform Engineering*

The flexibility of GFDM allows for designing a signal that has a very low OOB radiation. This section contributes with solutions for low OOB radiation.

### 1.2.1 GFDM with Guard Symbol

In order to reduce the OOB emission, it is necessary to smooth the transitions between the GFDM blocks. The circularity of the signal in time-domain allows a simple and elegant solution to reduce the abrupt change between GFDM blocks. As can be seen in Fig. 5, the first subsymbol wraps around the edges of the block, introducing the abrupt amplitudes discontinuities.

By erasing the first sub-symbol, a guard symbol is introduced between the GFDM blocks and the edges of the signal fade out towards zero, making the transitions between blocks smooth. This technique is called guard-symbol GFDM (GS-GFDM) and Fig. 6 shows the time-domain signal and the corresponding spectrum.

The addition of the CP would introduce hard transitions between blocks once again. One solution to avoid this problem is to make also the last subsymbol null and making $N_{\mathrm{CP}} = K$. The drawback of this approach is the throughput reduction, given by

$$R_{\mathrm{GS}} = \frac{M-2}{M} \times \frac{KM}{KM+K} = \frac{M-2}{M+1}. \tag{14}$$

From (14), it is clear that this technique becomes interesting for scenarios where $M$ is large.

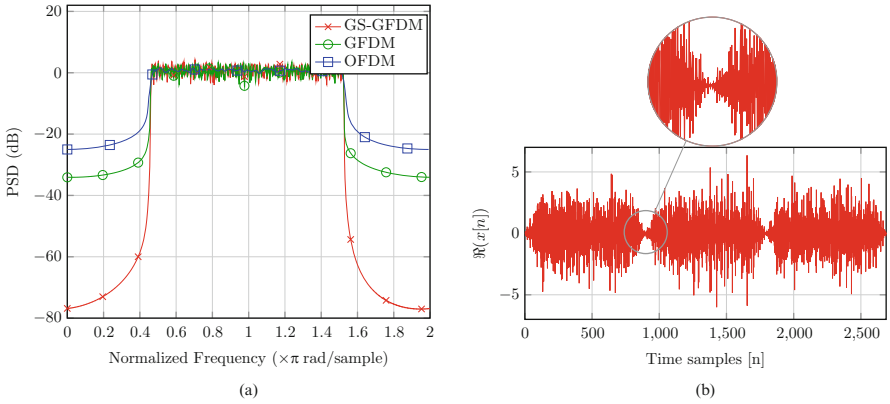


**Fig. 6** GS-GFDM signal. (**a**) Comparison of the GS-GFDM, GFDM and OFDM PSDs for the following parameters: $K = 128$ with 68 active subcarriers, $M = 7$, RC with $\alpha = 0.5$. (**b**) Real part of the GS-GFDM signal, highlighting the smooth transition between blocks

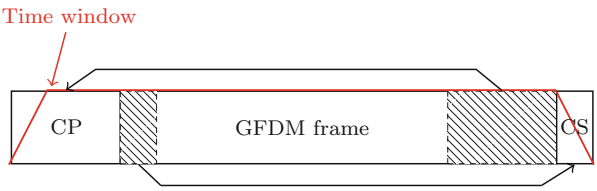


**Fig. 7** W-GFDM time-domain signal. A time-window is used to smooth the transitions between GFDM blocks

### 1.2.2 Windowed-GFDM

For windowed GFDM (W-GFDM), a time-window is applied for smoothing the transition between GFDM blocks, as depicted in Fig. 7.

A CP with length $N_{\text{CP}} = N_{\text{CH}} + N_{\text{W}}$ and a cyclic suffix (CS) with length $N_{\text{CS}} = N_{\text{W}}$, where $N_{\text{CH}}$ is the length of the channel impulse response and $N_{\text{W}}$ is the length of the time window transition, are added to the GFDM block. Notice that the CS is just the copy of the first $N_{\text{CS}}$ samples of the GFDM block to its end. The time window is defined as

$$w[n] = \begin{cases} w_{\text{rise}}[n] & 0 \leq n < N_{\text{W}} \\ 1 & N_{\text{W}} \leq n \leq N_{\text{CP}} + N \\ w_{\text{fall}}[n] & N_{\text{CP}} + N < n < N_{\text{CP}} + N + N_{\text{W}} \end{cases} \tag{15}$$

where $w_{\text{up}}[n]$ and $w_{\text{down}}[n]$ are the ramp-up and ramp-down segments of the time window, respectively. The ramp-up and ramp-down segments can assume different shapes. The most common cases are linear, RC or fourth order RC [2]. Figure 8 shows the PSD achieved by W-GFDM when linear and cosine ramp-up and

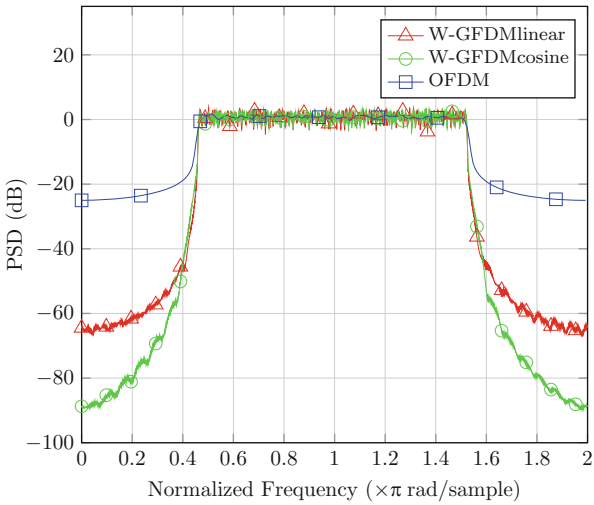


**Fig. 8** PSD of the W-GFDM using linear and cosine time windows compared with OFDM

ram-down, with 32 samples each, are employed. Clearly, varying the ramp-up and ramp-down sequences impacts the OOB emission. Sequences that introduce low derivative inflexion points in the edges of the signal provide lower OOB emission.

W-GFDM can be used to achieve low OOB emission and still keep high spectral efficiency, even when $M$ is low. The rate loss introduced by the ramp-up and ramp-down edges is given by

$$R_{\mathrm{W}} = \frac{N}{N + N_{\mathrm{CP}} + N_{\mathrm{W}}}. \tag{16}$$

Since $N_{\mathrm{CP}}$ is defined as a function of the channel impulse response, it equally affects the rate loss of GS-GFDM and W-GFDM. The ramp-up and the ramp-down sequences are much smaller than the GFDM block length, meaning $R_{\mathrm{W}}$ is typically higher than $R_{\mathrm{GS}}$. From Figs. 6 and 8, it is possible to conclude that both GS-GFDM and W-GFDM can achieve similar OOB emission. Therefore, W-GFDM can be seen as a more promising solution for a high efficient PHY with low OOB emissions.

## *1.3 Precoding: WHT*

One challenge for low-latency scenarios is to achieve reliable communication with single shot transmission over frequency selective channels. In this case, relatively small packages must be received with low probability of error since the low latency requirement does not allow for retransmissions of missed packages. Precoding using WHT can efficiently increase the GFDM robustness over multipath channels [9].

The main idea is to spread the data symbols over all subcarriers, so data symbols can be correctly detected on the receiver side even when a subset of subcarriers suffers from severe attenuation.

In order to achieve this goal, for each subsymbol, the data symbols $\mathbf{d}_{c,m}$ are linearly combined using the Walsh-Hadamard matrix

$$\boldsymbol{\Omega}_K = \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{\Omega}_{\frac{K}{2}} & \boldsymbol{\Omega}_{\frac{K}{2}} \\ \boldsymbol{\Omega}_{\frac{K}{2}} & -\boldsymbol{\Omega}_{\frac{K}{2}} \end{bmatrix}, \tag{17}$$

with $\boldsymbol{\Omega}_1 = [1]$. Hence, the data coefficients transmitted in the $m$th subsymbol are given by

$$\mathbf{c}_m = \boldsymbol{\Omega}_K \mathbf{d}_{c,m}. \tag{18}$$

On the receiver side, after the GFDM demodulation, the data symbols can be reconstructed as

$$\hat{\mathbf{d}}_{c,m} = \boldsymbol{\Omega}_K^{\mathrm{H}} \hat{\mathbf{c}}_m = \boldsymbol{\Omega}_K \hat{\mathbf{c}}_m. \tag{19}$$

Figure 9 shows the WHT-GFDM SER performance assuming the parameters presented in Table 4 and the channel delay profiles shown in Table 5. The figure shows that the gain introduced by the WHT highly depends on the channel delay profile. Higher gain is expected when the channel presents narrow and deep notches in the frequency response, while smaller gain is obtained under channel with mild frequency responses. Therefore, WHT-GFDM is an interesting scheme to be used in scenarios where low OOB emission and robustness must be combined.

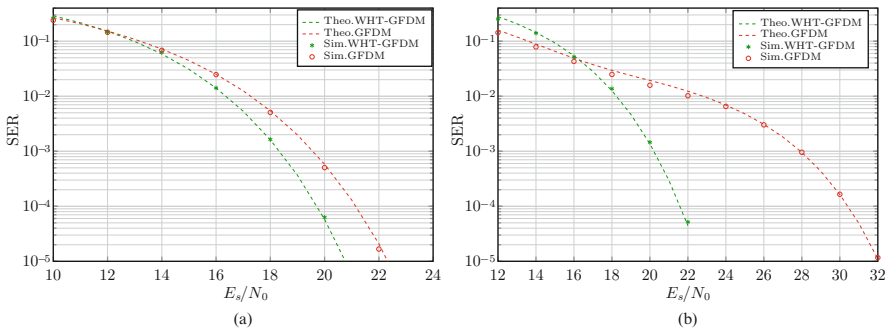


**Fig. 9** SER performance of WHT-GFDM over FSC using the ZF demodulator. (**a**) SER performance over channel A. (**b**) SER performance over channel B

**Table 4** Simulation parameters

| Parameter | Value |
|---|---|
| Mapping | 16-QAM |
| Transmit filter | RC |
| Roll-off ($\alpha$) | 0.25 |
| Number of subcarriers ($K$) | 64 |
| Number of subsymbols ($M$) | 7 |
| GFDM block duration | 256 µs |
| CP duration | 32 µs |
| Windowing | Not used |

**Table 5** Channel delay profiles

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Ch. A | Gain (dB) | 0 | −8 | −14 | – | – | – | – |
| | Delay (µs) | 0 | 4.57 | 9.14 | – | – | – | – |
| Ch. B | Gain (dB) | 0 | −10 | −12 | −13 | −16 | −20 | −22 |
| | Delay (µs) | 0 | 2.85 | 4.57 | 6.28 | 9.71 | 15.43 | 20 |

## *1.4 OQAM*

In general, GFDM is a non-orthogonal waveform. However, it is possible to make it orthogonal by choosing some special pulse shaping filters, such as Dirichlet pulses. Another solution is to combine GFDM with OQAM [10].

Instead of direct modulation of the in-phase (I) and quadrature-phase (Q) symbol input, OQAM can be seen as two independent Gabor expansions where the symbol input is restricted to be either only the real or only the imaginary component that achieves the orthogonality condition.

For square root Nyquist prototype filters with bandwidth limited to two subcarriers, the I-input introduces only real interference in the adjacent subcarriers, while the Q-input produces only imaginary interference. This aspect is illustrated in Fig. 10 using the ambiguity function $\mathscr{X}(m,k)$ to calculate the contribution from a transmitted symbol to every other position of the half-symbol spaced time-frequency grid. In Fig. 10, the I-input is used to modulate a root raised cosine (RRC) prototype filter with roll-off 0.5. The magnitude of real and imaginary part of $\mathscr{X}(m,k)$ reveals regions free of interference. Therefore, orthogonality can be achieved if every even and odd subcarrier is modulated with I and Q inputs, respectively.

Figure 10 also shows that I-inputs produce zero real interference in the adjacent subcarriers at every half subsymbol shift while Q-inputs produce zero imaginary interference. Hence, a second Gabor expansion with a time-shift of half subsymbol duration can be used to transmit data free of interference.

The use of OQAM considering pulse shaping filters with arbitrary length and overlapping factors greatly increases the flexibility of the waveform. Recently, frequency-shift OQAM has also being introduced as an alternative for using shorter

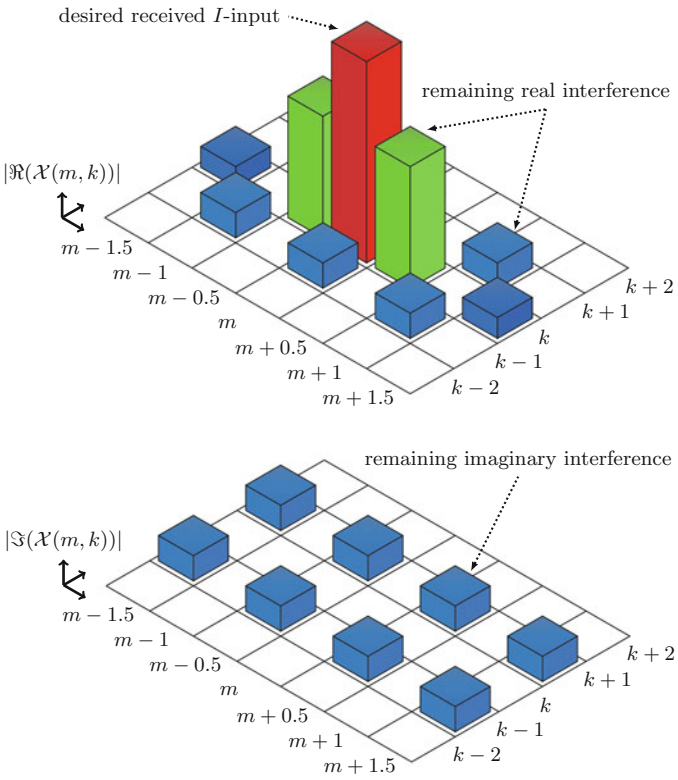


**Fig. 10** Illustration of the ambiguity function $\mathcal{X}(m,k)$, with the magnitude of the residual real (*top*) and imaginary (*bottom*) values along the neighbors subsymbols and subcarriers, for a GFDM transmitted $I$-input using RRC prototype filter with roll-off $= 0.5$

pulses in time [11]. Also, low complexity implementation that fully explores the cyclic principles has been addressed in [2] and allows to harvest the benefits of OQAM.

## *1.5 Waveform Framework*

This section illustrates how GFDM can be used as a framework for multicarrier waveforms. To this end, the parameters and properties of GFDM necessary to emulate prominent waveforms are listed in Table 6. All contestants have common roots in the filtered multicarrier systems proposed by Chang [12] and Saltzberg [13]. Although this framework covers the most popular waveforms, note that some candidates may not be covered. For example, Universal Filtered Multicarrier (UFMC) [14] applies linear filtering to a set of subcarriers and does not use CP, making it incompatible with GFDM.

**Table 6** GFDM parameters for emulating 5G waveform candidates

| Design space | GFDM | Block | | | | FBMC OQAM | FBMC FMT | FBMC COQAM | CB-FMT | FTN | SEFDM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | OFDM | OFDM | SC-FDE | SC-FDM | | | | | | |
| # subcarriers | $K$ | $K$ | $K$ | 1 | $K$ | $K$ | $K$ | $K$ | $K$ | $K$ | $K$ |
| # subsymbols | $M$ | 1 | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ | $M$ | 1 |
| Scaling freq. | $\nu_\mathrm{f}$ | 1 | 1 | 1 | 1 | 1 | $> 1$ | 1 | $> 1$ | 1 | $< 1$ |
| Scaling time | $\nu_\mathrm{t}$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | $< 1$ | 1 |
| Silent subsym. | $M_\mathrm{s}$ | – | – | – | – | $M_\mathrm{p}$ | $M_\mathrm{p}$ | – | – | $M_\mathrm{p}$ | – |
| Filter imp. resp. | Cyclic | Rect | Rect | Dirichlet | Dirichlet | $\sqrt{\text{Nyquist}}$ | $\sqrt{\text{Nyquist}}$ | Cyclic | Cyclic | IOTA | Rect |
| Offset mod. | (Yes) | (Yes) | No | No | No | Yes | No | Yes | No | Yes | No |
| Cyclic prefix | Yes | Yes | Yes | Yes | Yes | No | No | Yes | Yes | No | Yes |
| Orthogonal | (Yes) | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | No | No |
| Application scenarios | All | Legacy syst. | Bitpipe | IoT, MTC | IoT, MTC | WRAN, bitpipe | WRAN | Tactile internet | Tactile internet | Bitpipe | Bitpipe |
| Beneficial features | Flex. | Orth. | Small CP overhead | Low PAPR | Low PAPR | Low OOB | Low OOB | No filter tail | No filter tail | Spectral eff. | Spectral eff. |

In the context of this framework, the different waveforms are characterized by two aspects. First, parameters related to the dimensions of the underlying resource grid are explored. This includes the number of subcarriers $K$ and subsymbols $M$ in the system. The scaling factor in time $\nu_t$ and frequency $\nu_f$ can theoretically take values of any rational number larger than zero, while numbers close to one are meaningful because they relate to critically sampled Gabor frames. Additionally, the option to have guard symbols, i.e., the presence of silent subsymbols, is relevant for some candidates. The second set of features is related to the properties of the signal. Here, the choice of the pulse shaping filter is a significant attribute and the presence or absence of circularity constitutes a characteristic feature. Moreover, the use of OQAM is needed for some waveforms, aiming to achieve real orthogonality. Further, some waveforms rely on a CP to allow transmission of a block based frame structure in a time dispersive channel, while others don't use CP in order to achieve higher spectrum efficiency.

The family of *classical waveforms* includes OFDM, block OFDM, Single Carrier Frequency Domain Equalization (SC-FDE) and single carrier frequency division multiplexing (SC-FDM). Particularly OFDM and SC-FDM have been relevant for the development of the fourth generation (4G) cellular standard Long-Term Evolution (LTE). All four waveforms in this category have in common that $\nu_f = 1$ and $\nu_t = 1$, which allows to meet the Nyquist criterion. Silent subsymbols are not employed, the CP and regular quadrature amplitude modulation (QAM) are used in the default configuration. OFDM and block OFDM are corner cases of GFDM, where a rectangular pulse is used. Additionally, OFDM is restricted to one subsymbol, while block OFDM constitutes the concatenation of multiple OFDM symbols in time to create a block with a single common CP. Similarly, SC-FDE and SC-FDM can also be considered as corner cases of GFDM. However, here a Dirichlet pulse is used and analogously, the number of subcarriers in SC-FDE is $K = 1$, while SC-FDM is a concatenation in frequency of multiple SC-FDE signals. All waveforms in this category share property of orthogonality, but with different sensitivities towards various radio frequency (RF) imperfections, for instance SC-FDE is well known for its low peak-to-average power ratio (PAPR), which greatly benefits the mobile stations (MS) in terms of transmit power efficiency and reduced cost of the power amplifier.

The family of *filter bank waveforms* revolves around filtering the subcarriers in the system and still retaining orthogonality. As the names suggest, Filterbank multicarrier (FBMC)-OQAM [15] and its cyclic extension FBMC-COQAM [16] rely on offset modulation, while in FBMC-FMT and cyclic block filtered multitone (CB-FMT) [17] the spacing between the subcarriers is increased such that they do not overlap, i.e. $\nu_f > 1$. Also, a separation between cyclic and non-cyclic prototype filters can be made. In this context, silent subsymbols become relevant. The best spectral efficiency is achieved with no silent subsymbols $M_s = 0$, while $M_s > 0$ helps to improve the spectral properties of the signal. Using a sufficiently large number of silent subsymbols at the beginning and the end of a block allows to emulate non-cyclic filters from a cyclic prototype filter response, in order to generate

FBMC-OQAM and FBMC-FMT bursts. More precisely, $M_\mathrm{p}$ is the length of the prototype filter and $M_\mathrm{s} = M_\mathrm{p}$. Lastly, the CP is only compatible with cyclic filters.

Generally, the waveform can become non-orthogonal depending on the use of specific filters and for a given value of $\nu_\mathrm{f}$ and $\nu_\mathrm{t}$. This is addressed in the final category that consists of the *non-orthogonal multicarrier techniques* Faster than Nyquist (FTN) [15] and spectrally efficient frequency division multiplexing (SEFDM) [18]. The key property of FTN is $\nu_\mathrm{t} < 1$, which results in an increment of the subsymbol data rate. The isotropic orthogonal transform algorithm (IOTA) pulse, in combination with OQAM, has been proposed in order to avoid the use of CP. Since the impulse response of the filter is not cyclic, $M_\mathrm{p}$ subsymbols are silent. Analogously, the idea of SEFDM is to increase the density of subcarriers in the available bandwidth, i.e. $\nu_\mathrm{f} < 1$. Here, $M = 1$ because each block consists of a single subsymbol that is filtered with a rectangular pulse and a CP is prepended to combat multipath propagation. In this case, regular QAM is employed. Clearly, the amount of squeezing without severely impacting the error rate performance is limited. The Mazo limit states that this threshold is around 25 % for both schemes.

### *1.6 Multi-Service and Multi-User Systems*

In [14], a unified framework was proposed for delivering various services. GFDM has good time and frequency properties to fit into that framework. Namely, one subcarrier as a guard band is sufficient to divide the resource grid for serving different applications in an interference free manner. Benefiting from this feature, a single resource grid can convey various GFDM settings, including frame structure and pilot pattern, for target applications. This concept is particularly of interest for base station (BS) which can support programmable transceivers. On contrary, a MS does not need to have all possible features of GFDM implemented on its PHY, but only the features that are requested for the services supported by this equipment.

The GFDM time and frequency properties are also beneficial to mitigate the multi-user interference coming from asynchronous users sharing the resource grid [19]. With a single subcarrier as the guard band, the spectrum of the different users does not overlap. As such, the time misalignment among users does not lead to multi-user interference. Since the side lobes of the subcarriers spectrum are negligible, only coarse frequency synchronization among the devices is required to avoid spectrum overlapping due to frequency misalignments.

## 2 Gabor Transform: Theory Behind GFDM

In order to better understand the common principles of multicarrier modulation, consider a signal $s$. When looking at its time-domain representation $s(t)$, exact information about the behavior at any time instant is revealed. However, this

representation contains no information at all about frequency components at these positions. Instead, the Fourier transform (FT) of the signal provides exact information about frequency components, but on the other hand, no information on time-domain behaviour is obtained. The short-time Fourier transform (STFT) provides time-localized information about frequency components of a given signal by looking at the FT of the signal multiplied with a time window. However, the output of the STFT is highly redundant, since time and frequency sampling is kept independent.

To remove this redundancy, Dennis Gabor proposed in 1947 to represent a signal as a linear combination of Gaussian functions that are shifted in time and frequency to positions in a regular grid given by

$$s(t) = \sum_{k\in\mathbb{Z}} \sum_{m\in\mathbb{Z}} a_{k,m} g(t - m\Delta T) \exp(j2\pi \Delta F k t). \tag{20}$$

where $\Delta T$, $\Delta F$ describes grid distance in time and frequency, respectively and $g(t)$ describes the Gaussian bell function. Gabor proposed $\Delta F \Delta T = 1$. This technique is illustrated in Fig. 11. Gabor selected the Gaussian function as the basis function due to its best simultaneous localization in time and frequency, so that local behavior of the signal is most accurately described. According to Gabor, the original signal is fully characterized by the coefficients $a_{k,m}$ multiplying the Gaussian functions, establishing the foundation of time-frequency analysis [20]. More precisely, the uniqueness and existence of such an expansion was later analyzed and it was shown to critically depend on the density of the grid of time-frequency shifts, which is defined as the product $\rho \stackrel{\Delta}{=} \Delta T \Delta F$ of spacing in time $\Delta T$ and frequency $\Delta F$. Oversampled densities $\rho > 1$ imply non-unique expansions whereas with undersampled densities $\rho < 1$, the expansion coefficients only exist for certain signals. For critical densities $\rho = 1$, the expansion is unique and only exists for certain windows.

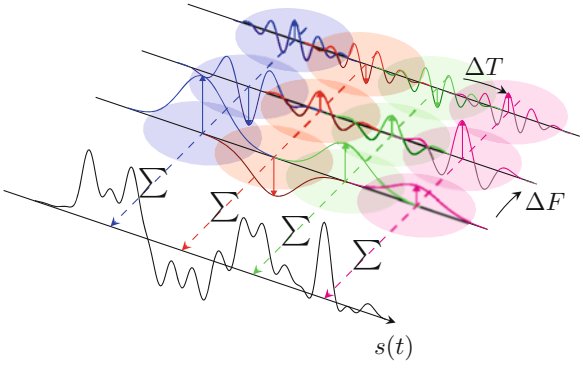


**Fig. 11** Illustration of Gabor expansion. The expanded signal is the sum of scaled time-frequency shifts of a prototype window. The scaling factors are given by the Gabor expansion coefficients

Nowadays, the linear combination of time-frequency shifted windows (20) is known as a Gabor expansion and the calculation of the STFT with a certain window at a regular grid is known as a Gabor transform with window $\gamma(t)$ given by [21]

$$a_{k,m} = \int_{\mathbb{R}} s(t)\gamma^*(t - m\Delta T)\exp(-j2\pi\Delta Fkt)dt \tag{21}$$

$$= \langle s(t), \gamma(t - m\Delta t)\exp(j2\pi\Delta Fkt)\rangle_{\mathbb{C}}\,. \tag{22}$$

The expansion and transform windows are in a dual relation. Accordingly, the coefficients used to expand to a certain signal with a given window $g(t)$ are provided by the Gabor transform of that signal with the dual window $\gamma(t)$. In case the window and its dual are equal, i.e. $\gamma(t) = g(t)$, the window is said to be orthogonal and expansion and transform reduce to well-known orthogonal expansion series

$$s(t) = \sum_{k\in\mathbb{Z}}\sum_{m\in\mathbb{Z}} \langle s(t), g_{k,m}(t)\rangle_{\mathbb{C}}\, g_{k,m}(t), \tag{23}$$

with $g_{k,m}(t) = g(t - m\Delta T)\exp(j2\pi\Delta Fkt)$.

A prominent example for Gabor expansion and transform is OFDM, which performs a Gabor expansion using a finite discrete set of rectangular window functions with length $T_S + T_{CP}$ in time and shifts of $1/T_S$ in the frequency grid. At the receiver, the dual receiver window are rectangular windows with length $T_S$ in order to achieve orthogonality between subcarriers.

## 3 Synchronization and Channel Estimation

As depicted in Fig. 2, synchronization and channel estimation are required for coherent equalization and demodulation. In this section, we introduce some basics of synchronization and channel estimation for GFDM systems.

### *3.1 Synchronization*

Synchronization can be achieved in GFDM on a block basis, allowing adaptation of fundamental OFDM solutions to estimate symbol time offset (STO) and carrier frequency offset (CFO) [22, 23], but low OOB emission is a factor to be specifically considered in GFDM.

Algorithms originally proposed to OFDM [23–25] can achieve one-shot synchronization using a straightforward proposal of a separated preamble. For instance, this concept can be exploited with low OOB in the GFDM case by using a windowed preamble (Fig. 12) for best isolation of data and training sequence, or by embedding

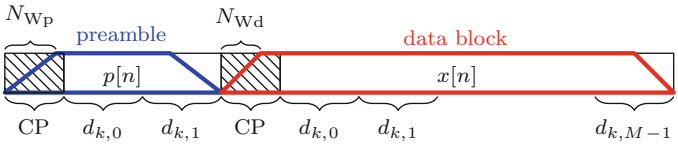


**Fig. 12** Windowed preamble preceding a GFDM data block

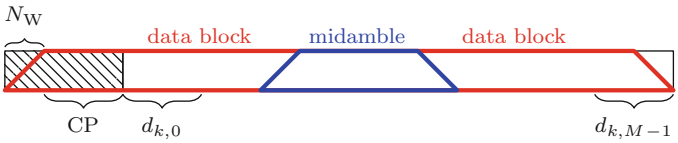


**Fig. 13** Embedded midamble in a GFDM data block

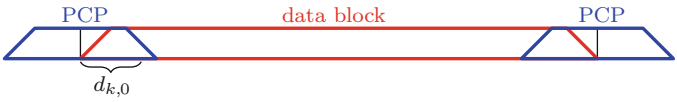


**Fig. 14** PCP in a GFDM data block

the training sequence as a midamble [26] (Fig. 13) or by a pseudo circular preamble (PCP) approach [27] (Fig. 14) for a self containing block transmission.

Classically, consecutive transmission of identical signals allows the exploration of robust coarse autocorrelation metrics [23, 24]. This can be achieved with subsymbols configured to carry the same pseudo-noise (PN) sequence data in all subcarriers in the first two time slots, $\mathbf{d}_{\mathrm{c},0}$ and $\mathbf{d}_{\mathrm{c},1}$, in the preamble case, or two consecutive time slots in the center of the block for the midamble case, or by the use of only even or odd subcarriers in the first time slot for the PCP scheme. In the latter one, the training information can be used to replace the traditional use of the CP with known cyclic data-aided information and easily achieve soft concatenation of blocks in continuous transmissions.

In the isolated preamble case, low OOB emission is obtained by pinching the block boundary, where a windowed preamble precedes a data block and forms a burst with a double pinching pattern. Different pinching lengths $N_{\mathrm{Wp}}$ and $N_{\mathrm{Wd}}$ can be applied respectively to the preamble and data blocks in order to achieve a desired emission mask. Regarding the midamble design, low OOB emission can be obtained by using guard subsymbols to smooth the block boundary, in this case the midamble and data forms a single burst with one pinching pattern. The midamble is placed in the center of the block, which minimizes its time difference to any subsymbol and can later also be used for equalization purposes. And with respect to the PCP approach, low OOB emission can be obtained by concatenating consecutive blocks, once in this case the training sequence forms a continuous transmission without transitions between blocks.

In the following the isolated preamble case will be presented in details.

Given that $r[n]$ is a set of received samples containing at least one complete windowed preamble, the two identical halves are identified with an autocorrelation metric, which is integrated along the CP and CS length to remove plateau effects [23, 24], leading to the metric

$$\mu[n] = \sum_{\iota=-N_{\mathrm{CP}}}^{0} \sum_{k=0}^{N-1} r[n+\iota+k]^* r\left[n+\iota+k+\frac{N}{2}\right], \tag{24}$$

where $N = MK$ denotes the length of the GFDM preamble. The argument that maximizes the absolute value of the metric $\mu[n]$ is taken as a coarse STO

$$\hat{n}_{\mathrm{c}} = \underset{n}{\mathrm{argmax}} |\mu[n]|, \tag{25}$$

while the angle of $\mu[\hat{n}_{\mathrm{c}}]$ is used to estimate the CFO as

$$\hat{\epsilon} = \frac{\angle \mu[\hat{n}_{\mathrm{c}}]}{\pi}. \tag{26}$$

The value $\hat{\epsilon}$ is employed to correct the CFO in the received sequence and a cross-correlation operation is then performed as

$$c[n] = \frac{1}{N} \sum_{k=0}^{N-1} r[n+k]^* e^{j2\pi \frac{\hat{\epsilon}}{N}(n+k)} p[k], \tag{27}$$

where $p[k]$ represents the known GFDM preamble.

To suppress side peaks that arise from the two halves and from the CP and CS parts, $c[n]$ is combined with $\mu[n]$ and an optimized estimation of the STO is obtained by searching the peak value in the range $\mathscr{S} = \{\hat{n}_{\mathrm{c}} - \frac{N}{2}, \ldots, \hat{n}_{\mathrm{c}} + \frac{N}{2}\}$ around the coarse STO estimation by

$$\hat{n}_{\mathrm{o}} = \underset{n \in \mathscr{S}}{\mathrm{argmax}} \left( |c[n]||\mu[n]| \right). \tag{28}$$

Depending on SNR and channel conditions, the search range $\mathscr{S}$ can be reduced in order to decrease implementation complexity. This synchronization procedure is robust for single path channels, but in a time-variant FSC, the primary echo can be lower than other echoes and the strongest peak will not represent the correct STO. Thus, an additional search before $\hat{n}_{\mathrm{o}}$ can reveal if there is another yet undetected peak to be considered as the primary one. For samples that do not belong to the preamble, the output of the cross-correlation can be interpreted as a complex Gaussian random sequence and a threshold criteria, depending on an acceptable probability of false alarm $p_{\mathrm{FA}}$, can reveal the presence of multipaths before $\hat{n}_{\mathrm{o}}$.

Considering that the first peak of (27) is within the range $(\hat{n}_{\mathrm{o}} - \lambda, \hat{n}_{\mathrm{o}}]$ , where $\lambda < N_{\mathrm{CP}}$ is an adjustable parameter, the threshold is defined as

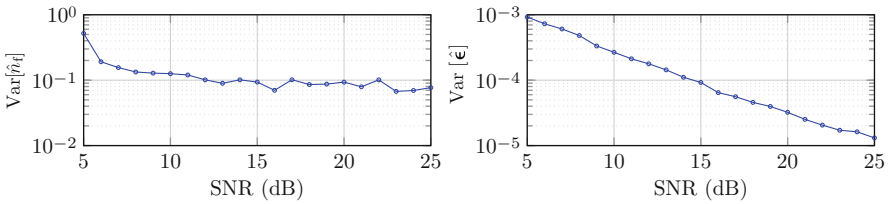


**Fig. 15** Performance evaluation of variance of the fine STO estimation $\hat{n}_\mathrm{f}$ and CFO estimation $\hat{\epsilon}$ for the GFDM preamble in a time-variant multipath channel

**Table 7** Isolated preamble parameters

| Parameter | Value |
|---|---|
| Mapping | QPSK |
| Transmit filter | Rect |
| Transmit window | RC |
| Roll-off ($\alpha$) | 0 |
| Number of subcarriers ($K$) | 64 |
| Number of subsymbols ($M$) | 2 |
| CP length ($N_\mathrm{CP}$) | 16 samples |
| CS length ($N_\mathrm{CS}$) | 0 samples |

$$T_\mathrm{Th} = \sqrt{-\frac{4}{\pi}\ln\left(p_\mathrm{FA}\right)}\left(\frac{1}{\frac{N}{2}-2\lambda}\sum_{k=-\frac{N}{2}+\lambda}^{-\lambda}\left|c[\hat{n}_\mathrm{o}+k]\right|\right) \tag{29}$$

and the fine STO estimation for the first multipath is finally given by

$$\hat{n}_f = \underset{\hat{n}_\mathrm{o}-\lambda<n\leq\hat{n}_\mathrm{o}}{\mathrm{argfirst}}\left(|c[n]| > T_\mathrm{Th}\right). \tag{30}$$

The performance evaluation in terms of variance of normalized STO and CFO estimations is presented in Fig. 15 for the W-GFDM preamble following the parameters presented in Table 7 but with $M = 2$.

A $w_{RC}[n]$ window function with $N_\mathrm{W} = 16$ is used, $\lambda = 16$, $p_\mathrm{FA} = 10^{-4}$ and the time-variant FSC with exponential decaying taps from 0 to −10 dB. For an SNR range higher than 5 dB the variance of the STO estimation stabilizes within tenths of a sample due the time variant fading effect in the multipath channel. The variance of the CFO estimation starts from thousandths of the subcarrier bandwidth and gets linearly better (in log scale) with increasing SNR.

The results obtained with the double pinching configuration show that burst synchronization can be achieved without penalties to the performance when compared with the results presented in [25]. The smooth block boundaries at the edges of the preamble and the data portion of the signal are particularly important as keeping OOB emission low is an important feature of GFDM.

### 3.2 Channel Estimation

Coherent detection used in modern mobile wireless communication standards requires knowledge of the wireless channel between transmit and receive terminals. Such channel state information (CSI) is usually obtained by transmitting known data (a.k.a. pilots) between the transmitted payload. At the receiver the effect of the wireless channel onto the known data is analyzed and the underlying CSI is extracted. Wireless channels are usually modelled as time-variant multipath channels, which are again modelled with the tapped-delay line principle, where the delays and gains represent the corresponding channel impulse response (CIR). Because of the variability of the wireless channel due to mobility or changes in the environment, the CIR changes over time. In this chapter, we assume a block-fading channel, which implies that the CIR is assumed constant for the duration of one GFDM block.

Generally, two methods for channel estimation exist: Preamble-based and pilot-based methods. In the former, a dedicated signal portion is only allocated with known data. Payload is subsequently transmitted independently of the preamble. The approach has the advantage that the preamble can be processed separately, hence reducing estimation complexity. The estimated CIR is then used for the payload block. In pilot-based methods, known data and payload are multiplexed and transmitted within the same resource block. In this case, channel estimation needs to consider if pilots and payload are orthogonal and if not, the payload needs to be additionally considered in the estimation, increasing estimation complexity. However, with the multiplexed approach, faster-changing wireless channels can be utilized, as the payload is virtually transmitted at the same time as the pilots, compared to the preamble-based approach.

In this chapter, we analyze the pilot-based channel estimation techniques for GFDM. Consider the GFDM data block vector $\mathbf{d}$ to be shared between payload and pilots such that

$$\mathbf{d} = \mathbf{d}_d + \mathbf{d}_p, \tag{31}$$

where $\mathbf{d}_d$ contains the payload and $\mathbf{d}_p$ describes the pilot symbols. Note that a single time-frequency resource can only be allocated to either pilot or payload, hence $\mathbf{d}_d$ and $\mathbf{d}_p$ cannot be non-zero at the same index and $\mathbf{d}_d \circ \mathbf{d}_p = \mathbf{0}$, where $\circ$ denotes the Hadamard (elementwise) product. Accordingly, assuming perfect synchronization, the received signal $\mathbf{y}$ in the time domain is given by

$$\mathbf{y} = \widetilde{\mathbf{H}}\mathbf{A}(\mathbf{d}_d + \mathbf{d}_p) + \mathbf{n}, \tag{32}$$

where $\mathbf{n}$ is the AWGN at the receiver and $\widetilde{\mathbf{H}} = \text{circ}(\mathbf{h})$ is the channel matrix performing circular convolution with the CIR $\mathbf{h}$. Transforming the received signal into the frequency domain yields

$$\mathbf{Y} = \mathbf{F}\mathbf{y} = \mathbf{F}\widetilde{\mathbf{H}}\mathbf{F}^{\mathrm{H}} \underbrace{\mathbf{F}\mathbf{A}\mathbf{d}_p}_{\mathbf{x}_p} + \mathbf{F}\widetilde{\mathbf{H}}\mathbf{F}^{\mathrm{H}} \underbrace{\mathbf{F}\mathbf{A}\mathbf{d}_d}_{\mathbf{x}_d} + \mathbf{F}\mathbf{n}, \tag{33}$$

with $\mathbf{F}$ being the unitary Fourier transform matrix. With denoting the channel frequency response by $\mathbf{H} = \mathbf{F}\mathbf{h}$, this is transformed into

$$\mathbf{Y} = \mathrm{diag}(\mathbf{H})\mathbf{x}_p + \mathrm{diag}(\mathbf{H})\mathbf{x}_d + \mathbf{N} \tag{34}$$

$$= \underbrace{\mathrm{diag}(\mathbf{x}_p)}_{\mathbf{X}_p}\mathbf{H} + \underbrace{\mathrm{diag}(\mathbf{x}_d)}_{\mathbf{X}_d}\mathbf{H} + \mathbf{N}. \tag{35}$$

With the introduced notation, the final system model used for deriving the estimation algorithms is hence given by

$$\mathbf{Y} = \mathbf{X}_p\mathbf{H} + \mathbf{X}_d\mathbf{H} + \mathbf{N}, \tag{36}$$

where $\mathbf{X}_p$ is a diagonal matrix with the known pilot signal in the frequency domain (FD) on the diagonal. $\mathbf{X}_d$ is a random diagonal matrix with the payload signal in the frequency domain on the diagonal. Hence, the linear model in (36) describes the received signal $\mathbf{Y}$ to be constructed of a deterministic part $\mathbf{X}_p\mathbf{H}$, a payload interference part $\mathbf{X}_d\mathbf{H}$ and a noise part $\mathbf{N}$. The influence of the payload interference is a significant difference from channel estimation for OFDM systems, where pilots and payload are separated with the Fourier transform. This on the one hand increases estimation complexity, but on the other hand enables to achieve a higher resolution in frequency domain for the estimation. From the model (36) we can now derive several basic estimators.

### 3.2.1 Least-Squares Channel Estimation

Direct application of Least Squares (LS) or ZF estimation results in

$$\widehat{\mathbf{H}}_{LS} = \mathbf{X}_p^{+}\mathbf{Y} \tag{37}$$

$$= \mathbf{H} + \mathbf{X}_p^{+}\mathbf{X}_d\mathbf{H} + \mathbf{X}_p^{+}\mathbf{N}, \tag{38}$$

where $\mathbf{X}_p^{+}$ denotes the pseudoinverse of $\mathbf{X}_p$. This estimator is unbiased, however it does not use any a-priori knowledge of the channel behaviour. Apparently, the noise is directly added to the estimation and noise enhancement occurs due to the multiplication with the pseudo inverse. Let $\boldsymbol{\chi}_p = \mathbf{X}_p^{+}$, the mean-squared error (MSE) of this estimator is given by

$$\mathrm{MSE}(\widehat{\mathbf{H}}_{LS}) = E[(\widehat{\mathbf{H}}_{LS} - \mathbf{H})^{\mathrm{H}}(\widehat{\mathbf{H}}_{LS} - \mathbf{H})] \tag{39}$$

$$= E[(\mathbf{H}^{\mathrm{H}}\mathbf{X}_d^{\mathrm{H}}\boldsymbol{\chi}_p^{\mathrm{H}} + \mathbf{N}^{\mathrm{H}}\boldsymbol{\chi}_p^{\mathrm{H}})(\boldsymbol{\chi}_p\mathbf{X}_d\mathbf{H} + \boldsymbol{\chi}_p\mathbf{N})] \tag{40}$$

$$= E[\mathbf{H}^{\mathrm{H}}\mathbf{X}_d^{\mathrm{H}}\boldsymbol{\chi}_p^{\mathrm{H}}\boldsymbol{\chi}_p\mathbf{X}_d\mathbf{H}] + E[\mathbf{N}^{\mathrm{H}}\boldsymbol{\chi}_p^{\mathrm{H}}\boldsymbol{\chi}_p\mathbf{N}], \tag{41}$$

where the last step follows since $\mathbf{H}$ and $\mathbf{N}$ are uncorrelated. In the first term, both $\mathbf{H}$ and $\mathbf{X}_d$ are random variables, however as $\mathbf{X}_d$ is diagonal, only element-wise multiplication of the random variables appears. Accordingly, the MSE is expressed as [28, Chap. 5.0.5]

$$\mathrm{MSE}(\widehat{\mathbf{H}}_{LS}) = \mathrm{tr}(\boldsymbol{\chi}_p^{\mathrm{H}} \boldsymbol{\chi}_p (\mathbf{R}_{\mathbf{HH}} \circ \mathbf{R}_{\mathbf{X}_d\mathbf{X}_d})) + \sigma_n^2 \mathrm{tr}(\boldsymbol{\chi}_p^{\mathrm{H}} \boldsymbol{\chi}_p), \tag{42}$$

where $\mathbf{R}_{\mathbf{HH}}$ and $\mathbf{R}_{\mathbf{X}_d\mathbf{X}_d}$ denote the covariance matrices of the channel $\mathbf{H}$ and the data $\mathbf{X}_d$, respectively. Apparently, the MSE consists of two terms, where only one depends on the channel noise. Hence, an error floor due to interference is expected, coming from overlapping payload and pilot signals. In order to improve the estimation performance one can apply the knowledge of the maximum channel length, which is given by the CP of the used system. Hence, the CIR is obtained by IDFT of the estimated frequency response, truncated after the $N_{CP}$th sample and transformed back into the frequency domain. The improved estimate $\widehat{\mathbf{H}}_{LS,t}$ is then given by

$$\widehat{\mathbf{H}}_{LS,t} = \mathbf{F}\mathbf{T}\mathbf{F}^{\mathrm{H}} \boldsymbol{\chi}_p \mathbf{Y}, \tag{43}$$

where $\mathbf{T}$ performs the truncation, being a diagonal matrix where only the first $N_{CP}$ elements of the diagonal are 1. Let $\mathbf{U} = \mathbf{F}\mathbf{T}\mathbf{F}^{\mathrm{H}}$, the resulting MSE is calculated by

$$\mathrm{MSE}(\widehat{\mathbf{H}}_{LS,t}) = \mathrm{tr}(\boldsymbol{\chi}_p^{\mathrm{H}} \mathbf{U}^{\mathrm{H}} \mathbf{U} \boldsymbol{\chi}_p (\mathbf{R}_{\mathbf{HH}} \circ \mathbf{R}_{\mathbf{X}_d\mathbf{X}_d})) + \sigma_n^2 \mathrm{tr}(\boldsymbol{\chi}_p^{\mathrm{H}} \mathbf{U}^{\mathrm{H}} \mathbf{U} \boldsymbol{\chi}_p) \tag{44}$$

### 3.2.2 Linear MMSE Channel Estimation

Starting from the linear model (36), a linear MMSE (LMMSE) estimator can be directly derived, given by

$$\widehat{\mathbf{H}}_{LM} = \mathbf{R}_{\mathbf{HY}} \mathbf{R}_{\mathbf{YY}}^{-1} \mathbf{Y}, \tag{45}$$

where $\mathbf{R}_{\mathbf{YY}}$ and $\mathbf{R}_{\mathbf{HY}}$ are given by

$$\mathbf{R}_{\mathbf{HY}} = \mathbf{R}_{\mathbf{HH}} \mathbf{X}_p^{\mathrm{H}} \tag{46}$$

$$\mathbf{R}_{\mathbf{YY}} = \mathbf{X}_p \mathbf{R}_{\mathbf{HH}} \mathbf{X}_p^{\mathrm{H}} + \mathbf{R}_{\mathbf{HH}} \circ \mathbf{R}_{\mathbf{X}_d\mathbf{X}_d} + \sigma_n^2 \mathbf{I}. \tag{47}$$

The calculation of the LMMSE estimator requires a-priori knowledge of the autocorrelation of the channel, i.e. the power delay profile needs to be known in advance. For example, this can be accomplished averaging over LS-estimates. The expected MSE for the LMMSE estimator is given by

$$\mathrm{MSE}(\widehat{\mathbf{H}}_{LM}) = \mathrm{tr}(\mathbf{R}_{\mathbf{HH}} - \mathbf{R}_{\widehat{\mathbf{H}}_{LM}\widehat{\mathbf{H}}_{LM}}) \tag{48}$$

$$= \mathrm{tr}(\mathbf{R}_{\mathbf{HH}} - \mathbf{R}_{\mathbf{HY}} \mathbf{R}_{\mathbf{YY}}^{-1} \mathbf{R}_{\mathbf{HY}}^{\mathrm{H}}) \tag{49}$$

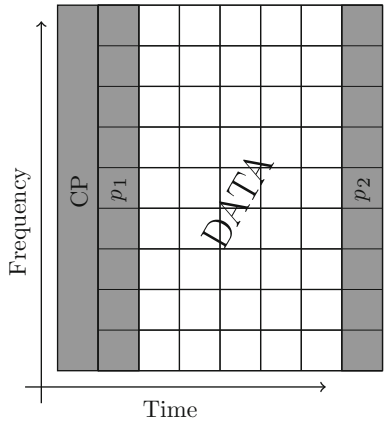


**Fig. 16** Proposed pilot arrangement

**Table 8** System parameters used for simulation

| Parameter | Symbol | Value |
|---|---|---|
| Number of subsymbols | $M$ | 7 |
| Number of subcarriers | $K$ | 64 |
| Cyclic prefix length | $N_{\mathrm{CP}}$ | 12 |
| Prototype filter | $g[n]$ | RC |
| Rolloff factor | $\alpha$ | 0.5 |
| Channel power delay profile | | Exponential |

### 3.2.3 Pilot Structure and Simulation Results

In this section, we compare the estimation accuracy of the three presented channel estimation algorithms. We use the pilot structure shown in Fig. 16. The system parameters for the presented evaluation are given in Table 8.

For the simulation either two ($p_1$, $p_2$) or only one ($p_1$) pilot subsymbols were used. Considering (42) and (44), the noise enhancement depends on the pseudo-inverse of the pilot signal. Accordingly, care needs to be taken when designing this signal. Values close to zero in the frequency domain of the pilot signal need to be avoided in order to not increase the noise. In the present simulation, Zadoff-Chu sequences are used for pilots on each subsymbol, where the sequence on the second pilot subsymbol is a circular shift of the first pilot subsymbol. This way it can be assured that both pilot sequences do not add up to zero in the frequency domain.

The estimation performance for all proposed estimators is shown in Fig. 17. Apparently, the naive LS estimation shows the worst estimation performance. The application of a-priori knowledge of the channel length significantly reduces the estimation MSE. Both LS estimators exhibit high error floors resulting from the interference part of the GFDM block, which is not accounted for in the estimation. On the other hand, the LMMSE estimator provides a significantly improved estimation performance, since it also takes self-interference into account. With this,

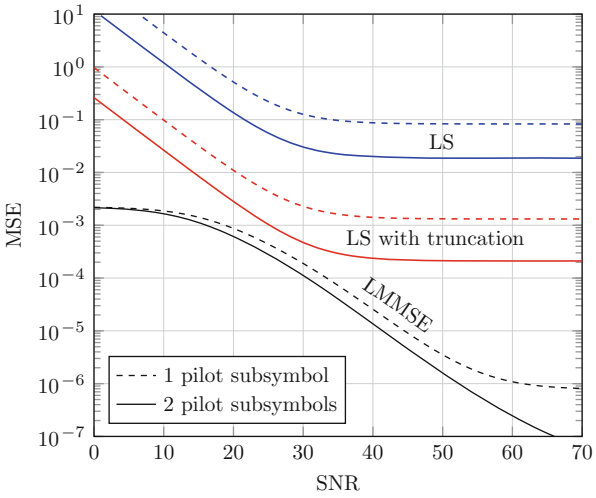


**Fig. 17** Mean squared errors of proposed estimators

the error floor is significantly reduced. In overall, the estimation performance is far superior compared to LS estimators. However, the LMMSE estimator requires knowledge of channel statistics.

In Fig. 17 the estimation performance is presented for both one and two used pilot subsymbols. According to the simulation, for LS estimation, the addition of the second pilot subsymbol reduces the error floor by roughly one order of magnitude and improves estimation by 3 dB SNR. With the LMMSE estimator, performance improves by roughly 1 dB, and the error floor is also reduced. However, for the LMMSE estimation, the improvement from addition of the second subsymbol is marginal compared to the improvement of LS estimators.

## 4 Multi-Antenna Systems

A 5G waveform must be compatible with MIMO techniques in order to achieve the necessary robustness and throughput in mobile environments. In this section, we introduce space time block code for GFDM to achieve transmit diversity. Furthermore, near-maximum likelihood (ML) detection algorithms are presented for MIMO systems using GFDM based spatial multiplexing (SM).

### 4.1 Space Time Block Code

This section describes a simple approach to achieve transmit diversity with GFDM, using time-reverse space time code (TR-STC), which has been proposed by Al-Dhahir [29] as a space time code (STC) for single carrier transmission over frequency-selective channels. The proposal takes two subsequent data blocks $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ of length $N$ which are separated by a CP. Let their corresponding discrete Fourier transforms be $\mathbf{X}_{(\cdot)} = \mathbf{F}\mathbf{x}_{(\cdot)}$, and $\mathbf{F}$ denotes the unitary Fourier transform matrix. Then, for two subsequent time slots the transmit signals on both antennas are given by

| | Antenna 1 | Antenna 2 |
|---|---|---|
| Block $i$ | $\mathbf{F}^{\mathrm{H}}\mathbf{X}_i$ | $-\mathbf{F}^{\mathrm{H}}\mathbf{X}^*_{i+1}$ |
| Block $i+1$ | $\mathbf{F}^{\mathrm{H}}\mathbf{X}_{i+1}$ | $\mathbf{F}^{\mathrm{H}}\mathbf{X}^*_i$, |

(50)

where $i$ is an even number and $(\cdot)^{\mathrm{H}}$ is the transpose-conjugate operator. Note that the name "time-reversal space-time coding" is reasoned by the property

$$(\mathbf{F}^{\mathrm{H}}\mathbf{X}^*_i)_n = x^*_i[\langle -n \rangle_N] \tag{51}$$

of the discrete Fourier transform.

At the receiver, after removing the CP, the CIR $\mathbf{h}_{j,l}$ between the $j$th transmit and $l$th receiving antenna circularly convolves the transmitted signals. Then, their frequency domain representation is calculated. Accordingly, assuming a block-fading channel during the transmission of two subsequent blocks, the frequency domain of the received blocks are given by

$$\mathbf{Y}_{i,l} = \mathbf{H}_{1,l}\mathbf{X}_i - \mathbf{H}_{2,l}\mathbf{X}^*_{i+1} + \mathbf{W}_{1,l} \tag{52}$$

$$\mathbf{Y}_{i+1,l} = \mathbf{H}_{1,l}\mathbf{X}_{i+1} + \mathbf{H}_{2,l}\mathbf{X}^*_i + \mathbf{W}_{2,l}, \tag{53}$$

where $\mathbf{H}_{j,l} = \mathrm{diag}(\mathbf{F}\mathbf{h}_{j,i})$. STC combining can then be carried out in the frequency domain according to

$$\begin{aligned} \widehat{\mathbf{X}}_i &= \mathbf{H}^{-1}_{eq} \sum_{l=1}^{L} \mathbf{H}^*_{1,l}\mathbf{Y}_{i,l} + \mathbf{H}_{2,l}\mathbf{Y}^*_{i+1,l} \\ \widehat{\mathbf{X}}_{i+1} &= \mathbf{H}^{-1}_{eq} \sum_{l=1}^{L} \mathbf{H}^*_{1,l}\mathbf{Y}_{i+1,l} - \mathbf{H}_{2,l}\mathbf{Y}^*_{i,l}, \end{aligned} \tag{54}$$

$$\text{where } \mathbf{H}_{eq} = \sum_{j=1}^{2} \sum_{l=1}^{L} \mathbf{H}^*_{j,l}\mathbf{H}_{j,l} \tag{55}$$

and $L$ is the number of receiving antennas. Finally, going back to the time domain with the inverse Fourier transform yields the estimates of the transmitted blocks

$$\hat{\mathbf{x}}_i = \mathbf{F}^{\mathrm{H}} \widehat{\mathbf{X}}_i. \tag{56}$$

When now identifying $\{\mathbf{x}_i\}_{i=0,1}$ with two subsequent GFDM blocks, we can apply the proposed STC technique to achieve transmit diversity. At the transmitter the GFDM modulated data blocks are space-time encoded with the technique above. At the receiver, the GFDM data blocks are recovered from the STC, yielding transmit diversity, and subsequently demodulated by the GFDM demodulator.

## 4.2 Spatial Multiplexing

Consider a $R \times T$ SM MIMO GFDM system with no transmit CSI and perfect CSI at the receiver side. The transmit signal of the $t$th transmit antenna is given by

$$\mathbf{x}_t = \mathbf{A}\mathbf{d}_t, \tag{57}$$

where $\mathbf{A}$ is the GFDM modulation matrix and $\mathbf{d}_t$ contains the transmit data of the $t$th antenna, taken from the set $\mathscr{S}$ of constellation symbols. The signal at the receive antennas can then be compactly written as

$$\underbrace{\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_R \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \mathbf{H}_{1,1}\mathbf{A} & \dots & \mathbf{H}_{1,T}\mathbf{A} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{R,1}\mathbf{A} & \dots & \mathbf{H}_{R,T}\mathbf{A} \end{bmatrix}}_{\widetilde{\mathbf{H}}} \underbrace{\begin{bmatrix} \mathbf{d}_1 \\ \vdots \\ \mathbf{d}_T \end{bmatrix}}_{\mathbf{d}} + \mathbf{n}, \tag{58}$$

where $\widetilde{\mathbf{H}}$ represents the equivalent channel of the MIMO system. $\mathbf{H}_{r,t}$ denotes the circulant channel matrix between the $t$th transmit and the $r$th receive antenna, $\mathbf{d}$ contains all $\mathbf{d}_t$ and $\mathbf{n}$ describes AWGN with covariance $N_0\mathbf{I}_{RN}$. The transmit symbols are energy-normalized such that $E[\mathbf{d}\mathbf{d}^{\mathrm{H}}] = \mathbf{I}$.

### 4.2.1 Near-ML Detection

At the receiver, the ML solution $\hat{\mathbf{d}}$ for (58) is

$$\hat{\mathbf{d}} = \arg \max_{\mathbf{d} \in \mathscr{S}^U} \|\mathbf{y} - \widetilde{\mathbf{H}}\mathbf{d}\|^2, \tag{59}$$

where $U = TM|\mathscr{K}|$ denotes the number of transmitted symbols per block and $\mathscr{K}$ is the set of allocated subcarriers on each antenna. Obviously, applying a

brute-force solution to (59) is prohibitively complex and hence, solutions with reduced complexity are of interest. In the orthogonal system OFDM $\widetilde{\mathbf{H}}$ can be block-diagonalized by the discrete Fourier transform (DFT), so that (59) reduces into $|\mathscr{K}|$ problems of dimension $T$. For each of these problems, a sphere decoder (SD) [30] can be applied to find the optimal ML solution. However, since the $\mathbf{H}_{r,t}\mathbf{A}$ cannot be simultaneously diagonalized, FD decoupling cannot be achieved in GFDM. To overcome the complexity issue, a combination of successive interference cancellation (SIC) with SD which achieves near-ML performance in high-SNR regimes is presented in the following.

Let the MMSE sorted QR-decomposition (SQRD) [31] of $\widetilde{\mathbf{H}}$ be

$$\begin{bmatrix} \widetilde{\mathbf{H}} \\ \sqrt{N_0}\mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix} \mathbf{R}\mathbf{P}^{\mathrm{T}}. \tag{60}$$

There, the matrix $\mathbf{P}$ permutes the columns of $\widetilde{\mathbf{H}}$ prior to decomposition. The received signal is multiplied with $\mathbf{Q}_1^{\mathrm{H}}$, yielding [31]

$$\tilde{\mathbf{y}} = \mathbf{Q}_1^{\mathrm{H}}\mathbf{y} = \mathbf{R}\mathbf{P}^{\mathrm{T}}\mathbf{d} - \sqrt{N_0}\mathbf{Q}_2^{\mathrm{H}}\mathbf{P}^{\mathrm{T}}\mathbf{d} + \mathbf{Q}_1^{\mathrm{H}}\mathbf{n}, \tag{61}$$

where the first, second and third term represent the useful signal, interference traded off for reduced noise enhancement and channel noise, respectively. The interference is ignored to yield

$$\tilde{\mathbf{y}} = \mathbf{R}\tilde{\mathbf{d}} + \tilde{\mathbf{n}} \tag{62}$$

with $\tilde{\mathbf{d}} = \mathbf{P}^{\mathrm{T}}\mathbf{d}$ and $\tilde{\mathbf{n}} = \mathbf{Q}_1^{\mathrm{H}}\mathbf{n}$. Still, the problem in (62) is of dimension $U \times U$ and direct application of SD requires prohibitive complexity. Instead, we introduce a factor $S$ that divides the problem of size $U$ into a chain of $U/S$ problems of size $S$ each. These subproblems are then subsequently solved. The proposed algorithm is given by

**while** $\mathrm{Dim}(\tilde{\mathbf{y}}) > 0$ **do**
  $\hat{\tilde{\mathbf{d}}}_S = \mathrm{SD}(\tilde{\mathbf{y}}_S, \mathbf{R}_{S,S})$ ▷ Jointly detect last $S$ streams.
  $\tilde{\mathbf{y}} \leftarrow \tilde{\mathbf{y}} - \mathbf{R}_{:,S}\hat{\tilde{\mathbf{d}}}_S$ ▷ Cancel interference.
  $\tilde{\mathbf{y}} \leftarrow \tilde{\mathbf{y}}_{\bar{S}}, \mathbf{R} \leftarrow \mathbf{R}_{\bar{S},\bar{S}}$ ▷ Reduce system size.
**end while**.

There, the subscripts $S, \bar{S}, :$ denote the last $S$, all but the last $S$, and all elements of the subscripted object, respectively. For matrices, two subscripts are used, where the first operates on rows and the second operates on columns. The proposed algorithm first performs joint detection of a subset of size $S$ of the complete problem. Afterwards, this detection is considered to be correct and the interference is cancelled from the remaining problems. Subsequently, it proceeds to the next subset until all subsets are processed. With the parameter $S$, detection complexity

**Table 9** System parameters used for simulation

| Parameter | | Symbol | Value |
|---|---|---|---|
| GFDM | # subsymbols, # subcarriers | $M, K$ | 5, 32 |
| | # allocated subcarriers | $\lvert\mathcal{K}\rvert$ | 16 |
| | Prototype filter | $g[n]$ | RC, $\alpha = 1$ |
| OFDM | Block length | $N$ | $160 = MK$ |
| | # allocated subcarriers | | $80 = M\lvert\mathcal{K}\rvert$ |
| General | CP length | $N_{\mathrm{CP}}$ | 16 |
| | Modulation | $\mathcal{S}$ | 64-QAM |
| | # transmit, receive antennas | $T, R$ | $2 \times 2$ |
| Channel PDP | Exponential, 0 to $-10$ dB, $N_{\mathrm{CP}}$ taps | | |

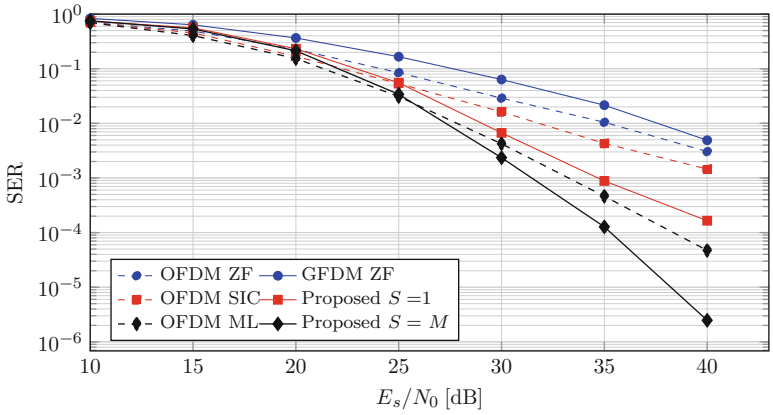


**Fig. 18** SER performance of proposed the detection algorithm in a fading multipath channel

and performance can be controlled, since with $S = 1$, standard SIC detection [31] with low complexity and with $S = U$, full ML detection with prohibitive complexity is performed.

In Table 9 the system parameters used for performance simulation of the proposed algorithm in independent Rayleigh-fading multipath channels are given. In Fig. 18, we compare the uncoded SER performance of the proposed algorithm with common detection methods for OFDM. For reference, also the SER performance for ZF equalization is provided, which shows that the SER of GFDM is 3 dB worse than OFDM due to the self-interference of the GFDM system. However, with the proposed algorithm, the SER of GFDM is found to be below that of OFDM both for pure SIC ($S = 1$) and SIC+ML ($S = M$) detection already at a SER of $\approx 5 \cdot 10^{-2}$. Note that for $S = M$, all subsymbols on one GFDM subcarrier are detected simultaneously. Since one subsymbol is transmitted on several frequency samples in GFDM, frequency diversity can be inherently exploited. This property cannot be achieved with OFDM, since for OFDM only one frequency sample is available for each symbol. Accordingly, GFDM can achieve a lower SER than OFDM for higher SNR values.

### 4.2.2 Soft MIMO Detection

After the invention of turbo codes, the concept of iterative decoding has been extended and has found tremendous applications in the transceiver design for wireless communications systems. Particularly, the detector as one key functional unit at the receiver shall not only estimate the transmitted symbols, i.e., hard output, but also provide the reliability information of them, i.e., soft output. In short, soft detection algorithms are necessary for iterative receiver designs.

Assume the data symbol vector $\mathbf{d} \in \mathscr{S}^U$ is generated from the codeword $\mathbf{c} \in \{0,1\}^L$ with $L \triangleq U \log_2 |\mathscr{S}|$. The optimal soft detector computes the a-posteriori log-likelihood ratio (LLR) of the $l$th code bit $c_l$ in the code bit vector $\mathbf{c}$ according to [32]

$$\lambda_l^{[\mathrm{p}]} = \log\left[\frac{\sum_{\mathbf{c}:c_l=1} P(\mathbf{c}|\mathbf{y},\widetilde{\mathbf{H}})}{\sum_{\mathbf{c}:c_l=0} P(\mathbf{c}|\mathbf{y},\widetilde{\mathbf{H}})}\right] = \log\left[\frac{\sum_{\mathbf{c}:c_l=1} P(\mathbf{y}|\mathbf{c},\widetilde{\mathbf{H}})P(\mathbf{c};\lambda_l^{[\mathrm{a}]})}{\sum_{\mathbf{c}:c_l=0} P(\mathbf{y}|\mathbf{c},\widetilde{\mathbf{H}})P(\mathbf{c};\lambda_l^{[\mathrm{a}]})}\right], \tag{63}$$

where the a-priori distribution of the codeword, i.e., $P(\mathbf{c};\lambda_l^{[\mathrm{a}]})$, is parameterized by the a-priori LLRs $\{\lambda_l^{[\mathrm{a}]}\}$ fed back from the decoder

$$P(\mathbf{c};\lambda_l^{[\mathrm{a}]}) = \prod_{l=0}^{L-1} P(c_l) = \prod_{l=0}^{L-1} \frac{e^{c_l\lambda_l^{[\mathrm{a}]}}}{1+e^{\lambda_l^{[\mathrm{a}]}}}. \tag{64}$$

In the case of MIMO-GFDM, the presence of ICI and ISI requires us to marginalize the distribution $P(\mathbf{c}|\mathbf{y},\widetilde{\mathbf{H}})$ over the complete codeword $\mathbf{c}$ except $c_l$ for obtaining $\lambda_l^{[\mathrm{p}]}$. For practical codeword lengths, approximations are evidently inevitable.

One possible approximation is to extend the above presented SIC-ML detection algorithm by applying the soft-input soft-output SD [30] instead of hard-input hard-output SD. As low complexity alternatives to SD for near-optimum soft MIMO detection, Markov chain Monte Carlo (MCMC) methods have been widely researched in the literature, e.g. [33–36]. Instead of using tree search to locate the most likely transmitted data symbols, they use Gibbs samplers to generate promising candidates based on a proposal distribution. While preserving near-optimum detection performance, it is possible to reduce exponential detection complexity to polynomial complexity. One issue with the MCMC technique discovered in the literature is the *stalling problem*, typically occurring at high signal-to-noise ratios (SNRs). The produced candidate set only consists of local optimums. To tackle this problem, we can activate multiple Gibbs sampling process, each of which generates samples independently [34]. We can also control the Gibbs sampler by a temperature parameter, e.g., in [35] and [37]. For more details, we refer readers to [38].

# 5 Proof-of-Concept: MIMO-GFDM Demonstrator

The development of a proof-of-concept GFDM transceiver with Software-Defined Radios is an important step from theory towards practice. Therefore, the described GFDM MIMO System is implemented on the National Instruments (NI) USRP-RIO 2953R platform. This section briefly describes the development of a $2 \times 2$ MIMO-GFDM demonstrator including the employment of TR-STC for achieving diversity gain, and preamble-based synchronization and channel estimation.

## *5.1 Implementation Model*

An overview of the processing blocks used to implement the transmitter and receiver can be found in Fig. 19.

At the transmitter, the inputs $\mathbf{d_1}$ and $\mathbf{d_2}$ are two data vectors that carry $N$ QAM symbols each. They are respectively filtered by a set of circular filters, yielding the modulated waveforms $\mathbf{x}_{\mathrm{d},1}$ and $\mathbf{x}_{\mathrm{d},2}$. Those two signals are fed to the TR-STC block, which applies the space-time block coding as described in Sect. 4.1. Next, the two predefined preamble signals $\mathbf{x}_{\mathrm{p},1}$ and $\mathbf{x}_{\mathrm{p},2}$ are acquired from a look-up table (LUT). From this point, all six signals undergo the same processing steps of CP appending and windowing. Lastly, the frame is assembled according to the structure displayed in Fig. 20. The samples of the two channels passed to the hardware are collected in $\mathbf{x}_1$ and $\mathbf{x}_2$.

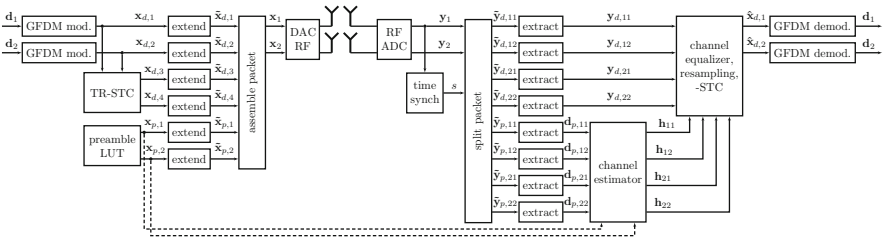


**Fig. 19** Transceiver block diagram

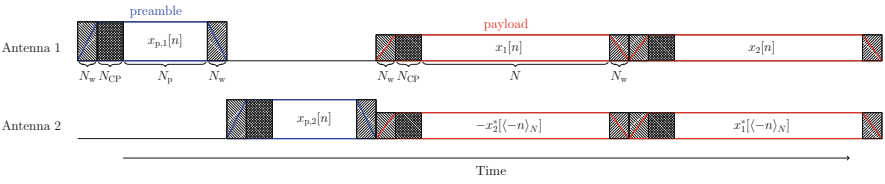


**Fig. 20** The TR-STC GFDM frame structure

On the receiver side, the software-defined radio (SDR) hardware provides the received signals $\mathbf{y}_1$ and $\mathbf{y}_2$, which each contain twice as many samples as in one transmitted frame. The signal on the first channel is used to obtain the start index of the frame $s$ through cross-correlation with the known preamble. In the subsequent extraction blocks, the received signals $\mathbf{y}_1$ and $\mathbf{y}_2$ are split up into four data blocks $\mathbf{y}_{\mathrm{d},11}$, $\mathbf{y}_{\mathrm{d},12}$ and $\mathbf{y}_{\mathrm{d},21}$, $\mathbf{y}_{\mathrm{d},22}$ and four preamble blocks $\mathbf{y}_{\mathrm{p},11}$, $\mathbf{y}_{\mathrm{p},21}$, $\mathbf{y}_{\mathrm{p},22}$ and $\mathbf{y}_{\mathrm{p},12}$, where the first index indicates the receive antenna and the second index corresponds to the time slot. The CP is removed in the same step. The channel estimation block uses the reference preamble signals $\mathbf{x}_{\mathrm{p},1}$, $\mathbf{x}_{\mathrm{p},2}$ to calculate the channel impulse responses $\mathbf{h}_{11}$, $\mathbf{h}_{12}$, $\mathbf{h}_{21}$, $\mathbf{h}_{22}$ based on the LS criterion. In the following block, the channel responses are resampled and then used by the space-time decoder to combine the received samples according to (54). The results are the estimated receive samples $\hat{\mathbf{x}}_{\mathrm{d},1}$ and $\hat{\mathbf{x}}_{\mathrm{d},2}$. These are then passed to the GFDM demodulator, which implements a linear receiver, i.e., ZF, to detect the data symbols $\hat{\mathbf{d}}_{\mathrm{d},1}$ and $\hat{\mathbf{d}}_{\mathrm{d},2}$. From the constellation points the error vector magnitude (EVM) and instantaneous SER can be used to evaluate the system performance later on.

## 5.2 *Demonstrator Setup*

The demonstrator consists of two terminals, one configured as transmitter and another set as receiver, as presented in Fig. 21. Each terminal is connected to a controlling PC and to a GPS reference clock to eliminate frequency offsets. On the transmitter host PC, data is generated, modulated with the GFDM waveform and TR-STC encoded. Table 10 presents the relevant parameters used in this experiment. The digital in phase and quadrature (IQ) samples are transferred via a PCI-Express connection to the transmitter front-end, which then performs digital to analog conversion and sends the signal over the air. At the receiver side, the IQ samples are captured and passed to the receiver host PC via the PCI-Express connection, where the waveform is demodulated.

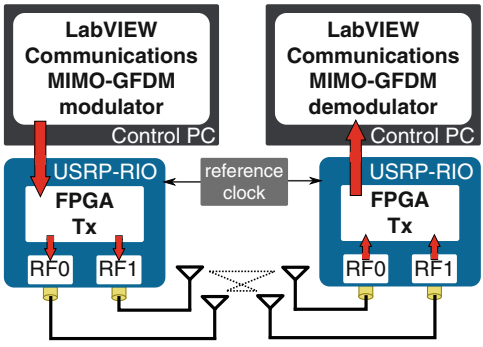


**Fig. 21** Schematic overview of the hardware setup

**Table 10** Parameters used in the demonstrator

| Parameter | Variable | Value |
|---|---|---|
| Number of subcarriers (data) | $K$, $K_p$ | 128 |
| Number of subcarriers (preamble) | $K_p$ | 128 |
| Number of subsymbols (data) | $M$ | 15 |
| Number of subsymbols (preamble) | $M_p$ | 5 |
| Number of transmit samples (data) | $N = KM$ | 1920 |
| Number of transmit samples (preamble) | $N_p = K_p M_p$ | 1920 |
| Set of active subcarriers (data) | $K_{\text{set}}$ | 11 to 54 |
| Set of active subcarriers (preamble) | $K_{\text{p, set}}$ | 10 to 55 |
| Set of active subsymbols (data) | $M_{\text{set}}$ | 1 to 15 |
| Set of active subsymbols (preamble) | $M_{\text{p, set}}$ | 1 to 5 |
| Prototype subcarrier filter | $\mathbf{g}$ | RRC |
| Prototype filter roll-off | $\alpha$ | 0.4767 |
| Window coefficients | $\mathbf{w}$ | |
| Window function roll-off | $\beta$ | 0.5 |
| Length of window flanks | $N_{\text{w}} = \beta K$ | 32 |
| Maximum channel delay spread | $N_{\text{ch}}$ | |
| Length of cyclic prefix | $N_{\text{CP}} = N_{\text{ch}}$ | 32 |
| Modulation order | $M_{\text{u}}$ | 4 |
| Number of channels | – | 2 |
| Center frequency | – | 1.99 GHz |
| Sampling frequency | – | 10 MHz |
| Transmitter gain | – | 0 dB |
| Receiver gain | – | 0 dB |

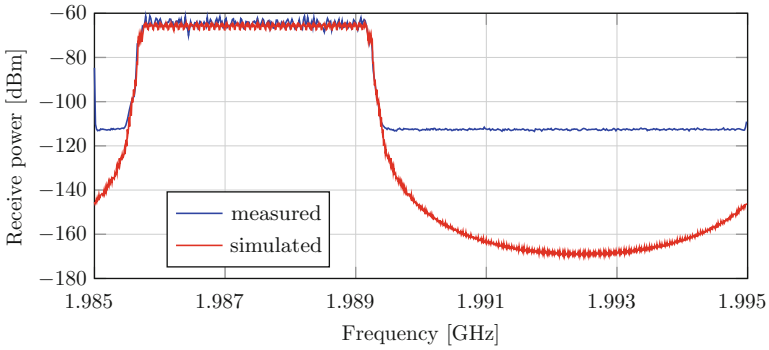


**Fig. 22** Comparison of measured and simulated spectrum

Figure 22 shows the measured spectrum from a spectrum analyzer and compares it with simulation results. The measured OOB radiation is at around −48 dB. This result is achieved due to the pulse shaping filter applied at each subcarrier. Therefore, GFDM is favorable in communication scenarios with highly fragmented

spectrum or in CRs. In [39], the authors have shown that GFDM has less influence and disturbance to legacy systems than OFDM. The simulated OOB level is considerably lower than the measured OOB level because it does not take into account the spectrum analyzer sensitivity and RF impairments, such as IQ-imbalance and nonlinearities, introduced by the USRP front-end.

## 6 Summary

The various scenarios and applications that must be covered by 5G networks will demand an unprecedented flexibility of the PHY layer. Although using a specific waveform to address the requirements of each scenario is one option, it is more desirable to have a single waveform that can be shaped to deal with the challenges imposed by the different applications. This chapter has proved GFDM to be a flexible multicarrier modulation that can be tuned to cover the major 4G and 5G waveforms. Namely, it has an efficient time-frequency resource grid for conveying information. This has been well exemplified in the context of Gabor transform. On top of it, OQAM can be applied in both time and frequency domains to obtain an orthogonal system. Precoding can be used to broaden the flexibility even more, for instance, precoding matrices can be used to achieve higher performance over frequency selective channels or lower OOB emissions. Roughly synchronized devices can share the time-frequency grid by using a single subcarrier as guard-band, while an efficient use of the CP can deal with long channel delay profiles and time-misalignments among the users. The overall GFDM performance can be further enhanced by MIMO techniques. All these features make GFDM a strong candidate for the 5G PHY layer.

This chapter has also addressed the receiver design for GFDM-based systems. It covers synchronization, channel estimation and MIMO detection. As an important step from theory to practice, the development of a MIMO-GFDM demonstrator has been presented at the end of this chapter.

## References

1. P. Pirinen, A brief overview of 5G research activities, in *1st International Conference on 5G for Ubiquitous Connectivity* (2014), pp. 17–22
2. N. Michailow, M. Matthé, I. Gaspar, A. Caldevilla, L. Mendes, A. Festag, G. Fettweis, Generalized frequency division multiplexing for 5th generation cellular networks. IEEE Trans. Commun. **62**(9), 3045–3061 (2014)
3. G. Fettweis, M. Krondorf, S. Bittner, GFDM - generalized frequency division multiplexing, in *VTC Spring 2009 - IEEE 69th Vehicular Technology Conference* (IEEE, Barcelona, 2009), pp. 1–4
4. M. Matthé, I.S. Gaspar, G. Fettweis, Short paper: reduced complexity calculation of LMMSE filter coefficients, in *Proceedings IEEE 82nd Vehicular Technology Conference*, Boston (2015)

5. I.S. Gaspar, N. Michailow, A.N. Caldevilla, E. Ohlmer, S. Krone, G. Fettweis, Low complexity GFDM receiver based on sparse frequency domain processing, in *IEEE 77th Vehicular Technology Conference, 2013. VTC Spring* (2013)
6. A. Farhang, N. Marchetti, L.E. Doyle, Low complexity modem design for GFDM. IEEE Trans. Signal Process. **99**, 1–1 (2015). Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7332798
7. N. Michailow, I. Gaspar, Generalized frequency division multiplexing: analysis of an alternative multi-carrier technique for next generation cellular systems, in … *Systems (ISWCS),* … (2012). Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6328352
8. M. Matthé, L. Mendes, N. Michailow, D. Zhang, G. Fettweis, Widely linear estimation for space-time-coded GFDM in low-latency applications. IEEE Trans. Commun. **63**(11), 4501–4509 (2015)
9. N. Michailow, L. Mendes, M. Matthé, I. Gaspar, A. Festag, G. Fettweis, Robust WHT-GFDM for the next generation of wireless networks. IEEE Commun. Lett. **19**(1), 106–109 (2015)
10. H. Bölcskei, Orthogonal frequency division multiplexing based on offset QAM, in *Advances in Gabor Analysis*, ed. by H. Feichtinger, T. Strohmer. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 2003), pp. 321–352
11. I. Gaspar, M. Matthé, N. Michailow, L. Mendes, D. Zhang, G. Fettweis, Frequency-shift Offset-QAM for GFDM. IEEE Commun. Lett. **99**, 1–1 (2015)
12. R.W. Chang, Synthesis of band-limited orthogonal signals for multichannel data transmission. Bell Syst. Tech. J. **45**(10), 1775–1796 (1966)
13. B. Saltzberg, Performance of an efficient parallel data transmission system. IEEE Trans. Commun. Technol. **15**(6), 805–811 (1967)
14. G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. Brink, I. Gaspar, N. Michailow, A. Festag, L. Mendes, N. Cassiau, D. Ktenas, M. Dryjanski, S. Pietrzyk, B. Eged, P. Vago, F. Wiedmann, 5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications. IEEE Commun. Mag. **52**(2), 97–105 (2014)
15. P. Banelli, S. Buzzi, G. Colavolpe, A. Modenini, F. Rusek, A. Ugolini, Modulation formats and waveforms for 5G networks: Who will be the heir of OFDM?: an overview of alternative modulation schemes for improved spectral efficiency. IEEE Signal Process. Mag. **31**(6), 80–93 (2014)
16. H. Lin, P. Siohan, Multi-carrier modulation analysis and WCP-COQAM proposal. EURASIP J. Adv. Signal Process. **2014**(1), 79 (2014). Available: http://asp.eurasipjournals.com/content/2014/1/79
17. A.M. Tonello, A novel multi-carrier scheme: cyclic block filtered multitone modulation, in *2013 IEEE International Conference on Communications (ICC)* (IEEE, Budapest, 2013), pp. 5263–5267
18. I. Kanaras, A. Chorti, M.R. Rodrigues, I. Darwazeh, Spectrally efficient fdm signals: bandwidth gain at the expense of receiver complexity, in *IEEE International Conference on Communications, 2009. ICC'09* (IEEE, Dresden, 2009)
19. M. Matthé, L. Mendes, I. Gaspar, N. Michailow, G. Fettweis, Multi-user time-reversal STC-GFDM for 5G networks. IEEE Trans. Wirel. Commun. (submitted, 2014). Available: http://jwcn.eurasipjournals.springeropen.com/articles/10.1186/s13638-015-0366-6
20. K. Gröchenig, *Foundations of Time-Frequency Analysis* (Birkhäuser, Boston, 2001)
21. J.J. Benedetto, C. Heil, D.F. Walnut, Gabor systems and the Balian-Low theorem, in *Gabor Analysis and Algorithms*, ed. by H.G. Feichtinger, T. Strohmer (Birkhäuser, Boston, 1998), Chap. 2, pp. 85–122
22. J. van de Beek, M. Sandell, P. Borjesson, ML Estimation of time and frequency offset in OFDM systems. IEEE Trans. Signal Process. **45**(7), 1800–1805 (1997)
23. T. Schmidl, D. Cox, Robust frequency and timing synchronization for OFDM. IEEE Trans. Commun. **45**(12), 1613–1621 (1997)
24. H. Minn, M. Zeng, V. Bhargava, On Timing Offset Estimation for OFDM Systems. IEEE Commun. Lett. **4**(7), 242–244 (2000)

25. A. Awoseyila, C. Kasparis, B.G. Evans, Improved preamble-aided timing estimation for OFDM systems. IEEE Commun. Lett. **12**(11), 825–827 (2008)
26. I. Gaspar, G. Fettweis, An embedded midamble synchronization approach for generalized frequency division multiplexing, in *IEEE Global Communications Conference (GLOBECOM)* (2015)
27. I. Gaspar, A. Festag, G. Fettweis, Synchronization using a pseudo-circular preamble for generalized frequency division multiplexing in vehicular communication, in *IEEE 82nd Vehicular Technology Conference (VTC Spring)* (2015)
28. R.A. Horn, C.R. Johnson, *Topics in Matrix Analysis* (Cambridge University Press, Cambridge, 1994). Available: https://books.google.com/books?id=LeuNXB2bl5EC&pgis=1
29. N. Al-Dhahir, Single-carrier frequency-domain equalization for space-time block-coded transmissions over frequency-selective fading channels. IEEE Commun. Lett. **5**(7), 304–306 (2001). Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=935750
30. C. Studer, H. Bölcskei, Soft-input soft-output single tree-search sphere decoding. IEEE Trans. Inf. Theory **56**(10), 4827–4842 (2010)
31. D. Wubben, R. Bohnke, V. Kühn, K.-D. Kammeyer, MMSE extension of V-BLAST based on sorted QR decomposition, in *2003 IEEE 58th Vehicular Technology Conference, 2003. VTC 2003-Fall*, vol. 1 (2003), pp. 508–512
32. A. Chindapol, J. Ritcey, Design, analysis, and performance evaluation for BICM-ID with square QAM constellations in Rayleigh fading channels. IEEE J. Sel. Areas Commun. **19**(5), 944–957 (2001)
33. B. Farhang-Boroujeny, H. Zhu, Z. Shi, Markov chain Monte Carlo algorithms for CDMA and MIMO communication systems. IEEE Trans. Signal Process. **54**(5), 1896–1909 (2006)
34. X. Mao, P. Amini, B. Farhang-Boroujeny, Markov chain Monte Carlo MIMO detection methods for high signal-to-noise ratio regimes, in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM)*, Washington, DC (2007)
35. M. Senst, G. Ascheid, A Rao-Blackwellized Markov chain Monte Carlo algorithm for efficient MIMO detection, in *Proc. IEEE Int. Conf. on Commun. (ICC)*, Tokyo (2011)
36. R. Datta, N. Michailow, M. Lentmaier, G. Fettweis, GFDM interference cancellation for flexible cognitive radio PHY design, in *2012 IEEE Vehicular Technology Conference (VTC Fall)* (IEEE, Quebec City, 2012), pp. 1–5
37. T. Datta, N. Ashok Kumar, A. Chockalingam, B. Rajan, A novel MCMC algorithm for near-optimal detection in large-scale uplink mulituser MIMO systems, in *Proc. IEEE Inf. Theory and Appl. Workshop (ITA)*, San Diego (2012), pp. 69–77
38. D. Zhang, M. Matthé, L. Mendes, G. Fettweis, A Markov Chain Monte Carlo algorithm for near-optimum detection of MIMO-GFDM signals, in *Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Hong Kong (2015)
39. M. Danneberg, R. Datta, A. Festag, G. Fettweis, Experimental testbed for 5g cognitive radio access in 4g lte cellular systems, in *2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM)* (2014), pp. 321–324

# Spectrally Efficient Frequency Division Multiplexing for 5G

**Izzat Darwazeh, Ryan C. Grammenos, and Tongyang Xu**

**Abstract** The focus of this chapter is on novel multi-carrier communication techniques, which share the common goal of increasing spectrum efficiency in future communication systems. In particular, a technology termed Spectrally Efficient Frequency Division Multiplexing (SEFDM) is described in detail outlining its benefits, challenges and trade-offs when compared to the current state-of-the-art. A decade of research has been devoted to examining SEFDM from different angles; mathematical modelling, algorithm optimisation, hardware implementation and system experimentation. The aim of this chapter is to therefore give a taste of this technology and in doing so, the chapter is organised as follows; first, it is explained how SEFDM fits within the remit of future 5th Generation (5G) communication systems; second, the design principles and implementation trade-offs associated with SEFDM systems are described; third, a number of linear and more sophisticated polynomial detection schemes are compared in terms of performance and complexity; finally, the chapter concludes by outlining a number of experimental testbeds which have been developed for the purpose of evaluating the performance of SEFDM in practical scenarios.

## 1 Introduction

### *1.1 Transition from 4G to 5G*

The beginning of the twenty-first century has been marked by a pressing need for more spectrum [9]. Networks are confronted with an ever-growing number of users who run bandwidth-hungry applications using a wide range of devices. Yet, spectrum is a scarce and an expensive resource. This fact coupled with users' demand for higher data rates, better Quality of Service (QoS) and lower costs, has led telecommunication operators and engineers to recognise that the best means for increasing capacity in future communication systems is via improved spectrum efficiency.

I. Darwazeh (✉) • R.C. Grammenos • T. Xu
University College London, London, UK
e-mail: i.darwazeh@ucl.ac.uk; r.grammenos@ucl.ac.uk; tongyang.xu.11@ucl.ac.uk



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_10

To this end, academia and industry are working together to define the characteristics of the next communication standard, commonly referred to nowadays as 5G systems [4]. Emerging technologies and disruptive ecosystems, such as machine-to-machine (M2M) communications and the Internet of Things (IoT), are also considered to be important pieces of the 5G puzzle and come under the umbrella of the recently coined term *tactile* Internet [18]. Two major European projects addressing 5G are METIS [20] and 5GNOW [76]. In particular, 5GNOW explores novel physical layer schemes which use non-orthogonal waveforms to overcome the strict orthogonality and synchronisation criteria specified in existing systems.

In line with these innovative projects, this chapter describes a novel Multi-carrier Modulation (MCM) scheme termed SEFDM, which promises to deliver bandwidth savings at the expense of receiver complexity [50].

## *1.2 Multi-Carrier Schemes for 5G*

Modulation schemes for high-speed data transmission have been of continuing interest for well over half a century [34]. The advancement of modern communication systems, however, can be attributed to the use of MCM techniques, which provide better immunity against multipath fading compared to single carrier schemes.

While Orthogonal Frequency Division Multiplexing (OFDM) is regarded as one of the key technology enablers in 4th Generation (4G) mobile telecommunication networks [73], it has certain disadvantages which make it inapt for spectrum optimisation, as depicted in Fig. 1. Thereby, MCM techniques aimed at improving spectrum utilisation are gaining increasingly more attention by the academic community and the business world across the globe [68]. These techniques propose the packing of more data within a conventional time-frequency plane by:

- Deliberately and counter-intuitively violating the orthogonality principle defined for OFDM systems, hence reducing the frequency spacing between the sub-carriers to save bandwidth or increasing the data rate within the same bandwidth.
- Transmitting non-orthogonal pulses with special properties.
- Applying advanced detection algorithms at the receiver, such as Sphere Decoding (SD), which has been made feasible due to advances in silicon technology.

One of the first systems designed to increase the data rate relative to an OFDM system was OFDM/Offset Quadrature Amplitude Modulation (OQAM) [33]. This was achieved by discarding the guard interval. As a result, alternative prototype functions providing good localisation in the time domain had to be employed to mitigate the effects of Inter-symbol Interference (ISI). The use of non-orthogonal functions offered new degrees of freedom favouring the design of improved pulses which were better-suited for minimising the Inter-carrier Interference (ICI) and ISI effects present in frequency-selective and time-dispersive channels [52]. An example of a prototype function offering quasi-optimum localisation is the Isotropic Orthogonal Transform Algorithm (IOTA) function which is generated by applying an IOTA filter to a Gaussian function [43].

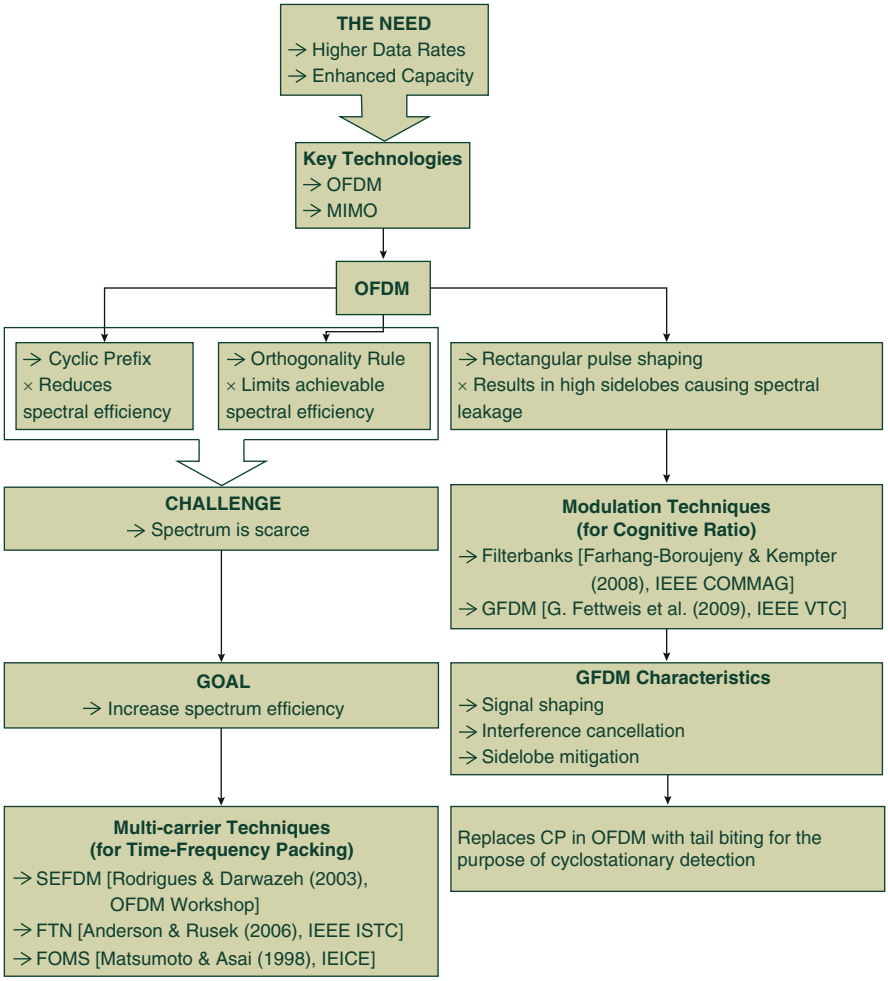


**Fig. 1** Transition from 4G to 5G: challenges and promising solutions

In retrospect, a number of systems have been explored with the aim of increasing spectrum efficiency through bandwidth savings and/or flexible frequency allocation [70]. Examples of systems which reduce spectrum utilisation by half but constrained to one-dimensional (1D) modulation schemes include Fast Orthogonal Frequency Division Multiplexing (FOFDM) [65] and M-ary Amplitude Shift Keying (ASK) OFDM (MASK-OFDM) [77]. Systems designed to operate with two-dimensional (2D) modulation methods include High Compaction MCM (HC-MCM) [29], Overlapped FDM (Ov-FDM) [44], Faster than Nyquist (FTN) signalling [3], Frequency Overlapped Multi-carrier System (FOMS) [54], as well as SEFDM [66].

FTN systems increase the data rate within the Nyquist signal bandwidth at the expense of introducing additional ISI [53]. This was proved in 1975 by Mazo [55],

who showed that for a reduction in transmission bandwidth by a factor $\rho = 0.8$ with reference to conventional Nyquist pulses, the signalling rate can be increased by $(\frac{1-\rho}{\rho}) \cdot 100 = 25\,\%$ without a penalty in error performance [55]. This concept, however, was validated only for 1D modulation schemes, such as Pulse Amplitude Modulation (PAM), ASK and Binary Phase Shift Keying (BPSK). In 2005, Rusek and Anderson [3] extended the concept to both a 2D and multi-carrier space, for example 16-Quadrature Amplitude Modulation (QAM) modulation [67] and multi-stream FTN signalling [3].

A different system that offers sidelobe mitigation, pulse shaping and interference cancellation, is Generalized Frequency Division Multiplexing (GFDM) [19]. GFDM reduces out-of-band radiation, thanks to Raised Cosine (RC) pulse shaping, and improves signal detection with the aid of tail biting techniques. Tail biting allows a circular structure to be maintained in each transmitted block, thus inducing cyclostationarity. OFDM is in fact a special case of GFDM with the number of blocks set equal to one [61]. It has been demonstrated that GFDM is well-suited for allowing a secondary system employing this multi-carrier scheme to be overlaid on top of an OFDM primary system [56] for use in Cognitive Radio (CR) applications. GFDM may be classified as a filterbank multi-carrier technique [17], the latter regarded as a promising method for allowing *agile waveforming* over fragmented spectrum [58] and for controlling spectral leakage.

The non-orthogonal concept may also be applied in multiple access scenarios such as sparse code multiple access (SCMA), non-orthogonal multiple access (NOMA) and multi-user shared access (MUSA). These techniques can superimpose signals from multiple users in the code-domain or the power-domain to enhance system access performance.

NOMA [69] is a non-orthogonal multiple access scheme in the power domain. Multiple users are superimposed with different power gains and separated via a Successive Interference Cancellation (SIC) detector at the receiver. MUSA [30] is an advanced multi-carrier Code Division Multiple Access (CDMA) system, which allows multiple users to transmit at the same time and frequency. Unlike the typical CDMA spreading sequence, MUSA employs non-orthogonal complex spreading sequences at the transmitter to modulate signals of each user leading to a denser deployment of existing resources. At the receiver, an advanced SIC detection algorithm is operated to remove the interference and recover each user's data. In SCMA [59], symbols from multiple users are sparsely spread in multiple layers which overlap in the frequency domain to enable non-orthogonal multiple access. All these non-orthogonal multiple access schemes aim to non-orthogonally superimpose multiple existing orthogonal waveforms from different users.

## *1.3 Spectrally Efficient Frequency Division Multiplexing (SEFDM)*

SEFDM is a non-orthogonal multi-carrier technique which plays an important role in the area of non-orthogonal waveform design. It has similar aims and characteristics to other non-orthogonal multi-carrier systems, for example FTN,

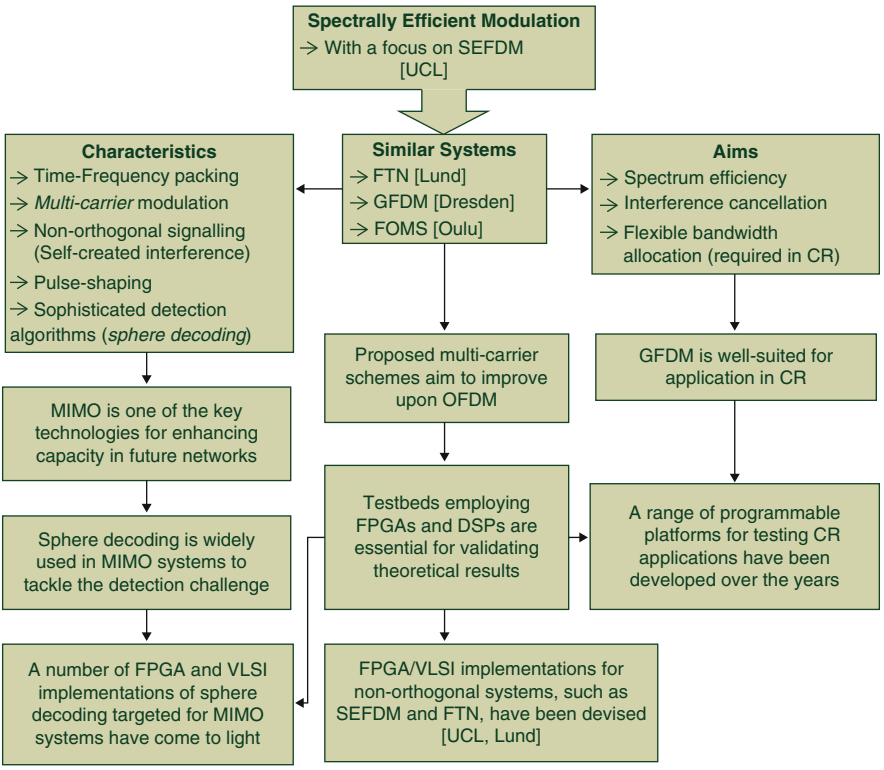


**Fig. 2** Aims and developments of non-orthogonal multi-carrier schemes

GFDM and FOMS, as illustrated in Fig. 2. The legend keys correspond to the following descriptions:

- **UCL**: This refers to the team working on SEFDM. The team is led by Izzat Darwazeh who is affiliated with University College London (UCL), United Kingdom. Notable researchers (current and alumni) working on SEFDM include: M.R.D. Rodrigues, I. Kanaras, A. Chorti, R.G. Clegg, S. Isam, M.R. Perrett, R.C. Grammenos, P.N. Whatmough and T. Xu.
- **Lund**: This refers to the team working on FTN. The team is led by John B. Anderson who is affiliated with Lund University, Sweden. Notable researchers include: F. Rusek, V. Öwall and D. Dasalukunte.
- **Dresden**: This refers to the team working on GFDM. The team is led by Gerhard Fettweis who is affiliated with Dresden University, Germany.
- **Oulu**: This refers to the team working on FOMS. The team is led by Tadashi Matsumoto who is affiliated with Oulu University, Finland.

The history of SEFDM can be traced back to 2002 when an OFDM based technique termed FOFDM [65] which doubled the data rate was proposed. Subsequently,

**Table 1** Timeline of SEFDM research and development

| Year | Literature |
|---|---|
| *SEFDM principles* | |
| 2002 | – Fast OFDM proposal [65] |
| 2003 | – SEFDM proposal [66] |
| 2012 | – SEFDM: a near-decade review [12] |
| *SEFDM transmission* | |
| 2010 | – SEFDM signal generation using IDFT based techniques [36]<br>– A precoded SEFDM system [35] |
| 2012 | – Characterisation of the self-created interference in SEFDM systems [39] |
| *SEFDM detection* | |
| 2008 | – A combined MMSE-ML technique for the detection of SEFDM signals [46]<br>– Near optimum detection for SEFDM systems [47] |
| 2009 | – A pruned SD algorithm for use in SEFDM systems [48]<br>– Semi definite programming detection for SEFDM systems [49]<br>– Evaluating the trade-off between performance and complexity for different SEFDM receivers [50] |
| 2010 | – A fast constrained sphere decoder for use in SEFDM systems [51]<br>– Joint channel equalisation and detection of SEFDM signals [11] |
| 2011 | – Detection of SEFDM signals using a truncated SVD approach [41]<br>– Evaluation of fixed complexity SEFDM receivers [37]<br>– Reducing the PAPR in SEFDM systems [38] |
| 2012 | – Robust channel estimation in SEFDM systems [40] |
| 2013 | – DSP evaluation of SEFDM detection techniques [26]<br>– Improving the detection of SEFDM signals through soft decision [84] |
| 2014 | – Detection of SEFDM signals for higher order modulation schemes [79]<br>– Block SEFDM: Reducing the complexity of SEFDM using a multi-band architecture [80]<br>– Bandwidth saving in a multi-path fading channel using turbo equalisation [78] |
| *SEFDM implementation* | |
| 2011 | – FPGA implementation of a reconfigurable SEFDM transmitter [62]<br>– VLSI architecture for a reconfigurable SEFDM transmitter [74]<br>– FPGA design of a truncated SVD based SEFDM receiver [27] |
| 2012 | – Verification and evaluation of an FPGA based SEFDM transmitter [63]<br>– Hardware implementation of a reconfigurable SEFDM receiver [25] |
| 2013 | – FPGA implementation of a real-time FSD detector for use in SEFDM systems [83] |
| *Other SEFDM systems* | |
| 2009 | – Application of SEFDM in physical layer security [10] |
| 2014 | – The first experimental optical SEFDM system [13] |
| 2015 | – Practical implementation of a beyond LTE-Advanced SEFDM system [82]<br>– First demonstration of a radio-over-fibre SEFDM system [57]<br>– Transmission of a 24 Gb/s Dual Polarization Coherent Optical SEFDM signal [60] |

the principle of SEFDM was proposed in 2003 [66] to overcome the limitations encountered with FOFDM, as discussed in Sect. 2.1. Since then, a number of researchers have worked on different aspects of SEFDM signals ranging from novel algorithms and techniques to generate and detect such signals, through to their practical implementation using state-of-the-art hardware devices. Table 1 gives an account of the key milestones achieved during the study of SEFDM signals over the last decade.

## 2 SEFDM Design Issues

The ever-increasing demand for higher data rates and improved coverage has called for a change of regime with the ultimate goal of optimising spectrum utilisation. SEFDM was developed to address the limits imposed upon spectral efficiency in OFDM systems and offers trade-offs in complexity and performance for diverse bandwidth compression rates.

### *2.1 Preliminary Concepts*

Taking into account that the Cyclic Prefix (CP) in OFDM systems reduces bandwidth efficiency by a factor of $\frac{N}{N+N_g}$, where $N$ is the number of sub-carriers and $N_g$ is the length of the CP in terms of discrete time samples, the first attempts to improve spectral efficiency focused on removing the CP and replacing rectangular pulses by pulses which were better localised in the time and frequency domains to avoid ISI. This alternative scheme termed OFDM/OQAM, which offsets the data timing by half a symbol, can guarantee orthogonality between sub-carriers only for real-valued symbols [43].

One of the key publications examining the relationship between time, frequency, orthogonality and bandwidth efficiency appeared in 1945 by Gabor [21]. This work along with the work by Mazo [55] inspired the pursuit of systems like Multi-stream FTN [3], GFDM [19], Filterbank Multi-carrier (FBMC) [16] and SEFDM. Furthermore, the use of FTN and SEFDM has been extended beyond wireless systems to optical links, as they appear in the work by Colavolpe [6] and Darwazeh [13], respectively. These techniques may be subsumed under one term as *time-frequency packing*. A timeline of the fundamental concepts, methods and systems developed over the years for time-frequency packing is illustrated in Fig. 3.

In a conventional Frequency Division Multiplexing (FDM) system, the frequency spacing between the sub-carriers is such that it obeys Nyquist's first criterion, also known as the Nyquist ISI criterion, which is the condition for zero ISI between transmitted symbols [64]. As the name suggests, FTN results in the transmission of more data within the minimum bandwidth specified by the Nyquist criterion.

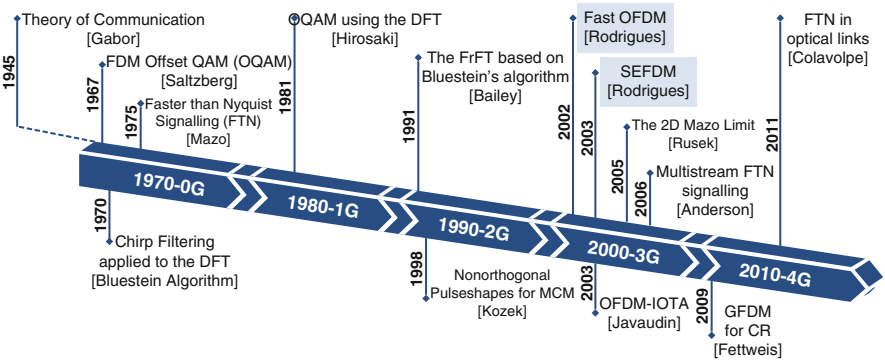


**Fig. 3** Timeline of spectrally efficient concepts and non-orthogonal communication techniques

This causes ISI which for a multi-carrier system also leads to ICI. Pulse shaping filters are therefore employed whose bandwidth and spectral shape determine how closely adjacent sub-carriers overlap with each other. Achievable information rates for FTN have been investigated recently [85]. While the focus here is on SEFDM, as proposed by Rodrigues and Darwazeh in 2003 [66], FTN, GFDM and SEFDM have a number of characteristics in common, such as:

- Aiming to increase spectral efficiency or favouring flexible bandwidth allocation;
- Using multi-carrier modulation similar to OFDM;
- Exploiting non-orthogonal signalling at the expense of self-created interference;
- Making use of pulse shaping filters to mitigate ISI (in FTN and GFDM);
- Employing iterative, tree-search and soft-decoding detection algorithms to recover the transmitted data successfully.

FOFDM [65] and MASK-OFDM [77] are special cases of spectrally efficient multi-carrier systems. These techniques operate only for non-complex modulation schemes, in other words 1D signals, such as BPSK and ASK, allowing the spacing between the sub-carriers to be reduced by half relative to a conventional OFDM system. This limitation motivated the invention of SEFDM [66], which allows the use of complex modulation schemes, such as M-QAM with $M > 2$. SEFDM offers an additional degree of freedom over OFDM by relaxing the stringent orthogonality constraint defined for OFDM. The attractiveness of SEFDM compared to other spectrally efficient techniques is the fact that it does not require special pulse shaping functions. Instead, it uses rectangular pulse shaping as in OFDM.

## 2.2 *SEFDM Signals*

SEFDM is based on similar key principles to OFDM in that a high data rate serial input stream is split into $N$ parallel, overlapping sub-carriers. Consequently, the conventional transceiver for SEFDM, as originally proposed in [66], resembles

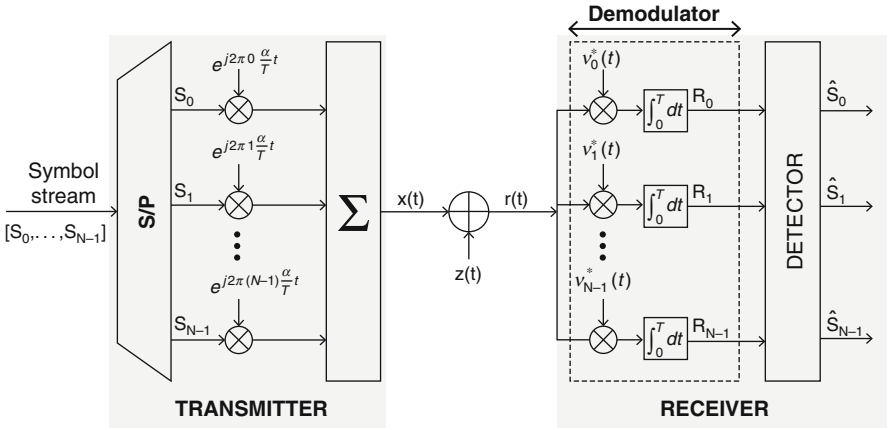


**Fig. 4** Conceptual architecture for an SEFDM transceiver using a bank of modems

that for OFDM, as illustrated in Fig. 4. Figure 4 shows that the SEFDM receiver is different from that of OFDM in that it incorporates two stages. The first stage involves a demodulation of the incoming received signal whose output is then fed to a second stage being a detector to generate estimates of the transmitted data symbols. It is assumed that no pulse shaping filters are employed, in other words, rectangular pulses are used, such as in OFDM.

As depicted in Fig. 4, $N$ complex input symbols, denoted by $s_n = s_{n_{\Re e}} + j s_{n_{\Im m}}$, modulate $N$ sub-carriers generating a continuous time domain SEFDM signal expressed as[1]

$$x(t) = \sum_{l=-\infty}^{+\infty} \sum_{n=0}^{N-1} s_{l,n} e^{j2\pi n \Delta f \alpha t}, \tag{1}$$

where $\alpha$ is a key system parameter in SEFDM systems and dictates the level of bandwidth compression. Contrary to OFDM, SEFDM deliberately and counter-intuitively violates the orthogonality rule defined for OFDM by reducing the spacing between the sub-carriers. Hence, in SEFDM the sub-carrier spacing is equal to $\Delta f = \frac{\alpha}{T}$, where $T$ is the SEFDM symbol duration with $\alpha < 1$ corresponding to an SEFDM signal and $\alpha = 1$ to an OFDM signal.

A general comparison between SEFDM and OFDM is illustrated in Fig. 5 to demonstrate the benefit of bandwidth saving in SEFDM. Figure 5a shows a typical OFDM spectrum, where the sub-carrier spacing $\Delta f_1$ equals $\frac{1}{T}$. Figure 5b depicts an SEFDM system with sub-carrier spacing $\Delta f_2 = \frac{\alpha}{T}$. Figure 5c presents

[1]The notations $\Re e$ and $\Im m$ represent the real and imaginary parts of a complex number, respectively, $l$ denotes the SEFDM symbol index, $n$ denotes the sub-carrier index and $s_{l,n}$ represents the information symbol conveyed by sub-carrier with index $n$ during the time stamp with index $l$.

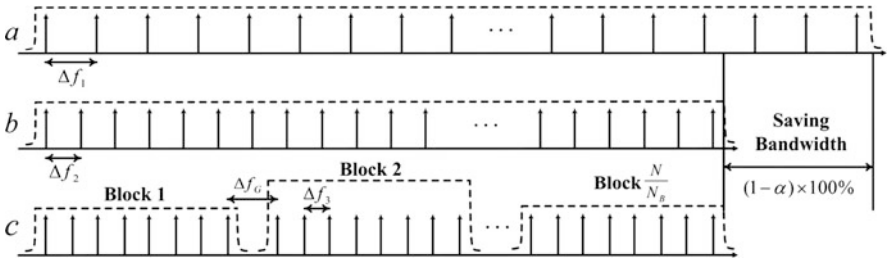


**Fig. 5** Spectra illustration for (**a**) OFDM, (**b**) SEFDM, and (**c**) Block-SEFDM. Each impulse represents a single sub-carrier and there are N sub-carriers overall for each system, respectively

an improved SEFDM system termed block-spectrally efficient frequency division multiplexing (B-SEFDM). It is evident that both the conventional SEFDM and B-SEFDM systems can save $(1-\alpha)\times 100\,\%$ of bandwidth compared to an OFDM system.

In OFDM, the frequency spacing $\Delta f$ between the overlapping sub-carriers has to equal the reciprocal of the OFDM symbol period $T$ to comply with the orthogonality rule. Mathematically, the orthogonality rule guarantees that the sub-carriers are mutually independent, in other words their cross-correlation is zero. In SEFDM, however, this is not the case, since the sub-carriers are packed closer together in a non-orthogonal manner. This gives rise to cross-talk between the sub-carriers, in other words self-created ICI, which presents a key challenge in SEFDM systems.

The self-created ICI present in SEFDM systems stems from the fact that the cross-correlation between the sub-carriers is no longer zero. This non-zero cross-correlation may in turn be quantified using the so-called correlation matrix **C**, which describes the self-created ICI. The magnitude of the non-zero cross-correlation determines the ill-conditioning of the matrix **C**. Ill-conditioning means that small changes in the entries of the input matrix yield very large changes in the solution matrix. These small changes are typically caused by round-off errors during matrix computations. As a result of this ill-conditioning, the correlation matrix **C** may tend to singularity,[2] thus becoming non-invertible and complicating the design of receiver architectures.

The elements of the correlation matrix **C** are given by

$$\begin{aligned} c_{m,n} &= \frac{1}{Q}\sum_{k=0}^{Q-1} e^{\frac{j2\pi mk\alpha}{Q}} e^{-\frac{j2\pi nk\alpha}{Q}} \\ &= \frac{1}{Q}\times\left\{\begin{array}{l} Q \\ \frac{1-e^{j2\pi\alpha(m-n)}}{1-e^{\frac{j2\pi\alpha(m-n)}{Q}}} \end{array}\begin{array}{l} ,\ m=n \\ ,\ m\neq n \end{array}\right\}. \end{aligned} \tag{2}$$

[2] A matrix is singular if its determinant is equal to zero [71].

where $m$,$n$ are indices of any two sub-carriers in the system with $m, n = 1, \ldots, N$. The diagonal terms ($m = n$) correspond to the auto-correlation of each sub-carrier with itself, which according to Eq. (2) is equal to one. The off-diagonal terms ($m \neq n$) measure the cross-correlation between any two overlapping sub-carriers, which for an SEFDM system would be non-zero. From this description and with reference to Eq. (2), it should be evident that for an OFDM system, the correlation matrix $\mathbf{C}$ is the identity matrix of size $N \times N$.

In SEFDM, the conditioning of the correlation matrix $\mathbf{C}$ depends on the system size, in terms of the number of sub-carriers $N$, and the level of bandwidth compression $\alpha$. The matrix becomes more ill-conditioned as $N$ is increased and/or the value of $\alpha$ is decreased leading to performance degradation. These concepts, as well as solutions to address these challenges are examined in further detail in Sect. 3.

## 3 Generation and Detection of SEFDM Signals

This section examines the transceiver architecture for an SEFDM system assuming an Additive White Gaussian Noise (AWGN) channel. Different detection techniques are compared and contrasted in terms of their error performance, complexity and ease of implementation.

### *3.1 SEFDM Signal Transmission*

The conventional method of generating SEFDM signals using a bank of analogue modulators does not scale well with the increase in the number of sub-carriers, taking into account that the complexity of this method is of the order $\mathcal{O}(N^2)$. Hence, the potential of employing Discrete Fourier Transform (DFT) operations to address this issue appears particularly attractive.

Since SEFDM is inherently a non-orthogonal system this means that standard DFT operations cannot be used directly. To this end, a first example of an algorithm employing the DFT in a modified configuration to generate non-orthogonal signals is the Inverse Fractional Fourier Transform (IFrFT) as proposed by Bailey [5], which considers fractional roots of unity. In analogy to an Inverse Discrete Fourier Transform (IDFT) operation, the output of the IFrFT may be expressed as

$$x_{Fr}(m) = \frac{1}{N} \sum_{n=0}^{N-1} s_n e^{j2\pi m\alpha n/N}, \tag{3}$$

which converges to the IDFT formula for $\alpha = 1$. As noted in [5], the IFrFT can be computed in a fast and efficient manner using the algorithm proposed by

Bluestein [7]. Using this latter algorithm the computation of the IFrFT requires $20N \log_2 N$ floating-point operations. While this results in the algorithm being 20 times more computationally intensive compared to the conventional IDFT algorithm, this technique still yields a significantly lower complexity compared to the analogue counterpart.

Notwithstanding, the ultimate goal is to generate non-orthogonal signals using standard DFT operations. The first attempts date back to the 1980s when Hirosaki used the DFT for an OQAM OFDM system [33]. A similar approach was adopted later on in FOFDM systems where the value of $\alpha$ is fixed at 0.5. More recently, however, standard DFT operations have been employed in non-orthogonal systems where the value of $\alpha$ is arbitrary. A first example appears in the HC-MCM system proposed by Hamamura [29]. Similar techniques were developed later on by Isam and Darwazeh [36] for use in SEFDM systems. These latter methods allow SEFDM signals to be generated using IDFT operations, thus offering the same advantages in terms of complexity as OFDM.

The equation for the continuous time SEFDM signal was given in Eq. (1). Let $T_s$ be the sampling period introduced to discretise the time and let $\rho$ be an oversampling factor such that $Q = \rho N$ with $\rho \geqslant 1$. Then the discrete time axis will be given by $mT_s$ while the symbol period will be given by $T = QT_s$. Consequently, the discrete time expression for a single SEFDM symbol may be expressed as

$$x(m) = \frac{1}{\sqrt{Q}} \sum_{n=0}^{N-1} s_n e^{j2\pi m\alpha n/Q}, \tag{4}$$

where $x(m)$ represents the time samples indexed by the integer $m$, $Q$ represents the total number of time samples contained in a single SEFDM symbol, $\mathbf{s} = [s_0, \ldots, s_{N-1}]'$ is a vector of complex data symbols modulated using Phase Shift Keying (PSK) or QAM modulation and $[\cdot]'$ denotes a vector or matrix transpose operation. Specifically, $\mathbf{s}$ represents the input data vector with $\mathbf{s} \in \mathfrak{M}^N$ where $\mathfrak{M}$ represents a discrete alphabet, in other words the set of constellation points. The size of the alphabet, known as the constellation cardinality, is denoted by $M$. Where QAM is employed, it is assumed that the constellation is rectangular with $M$ equal to even powers of two and that the data symbols are encoded using Gray encoding.

Three different, nevertheless equivalent, SEFDM transmitter types were proposed by Isam, which are illustrated in Fig. 6 with details provided in [36]. In these architectures, the bandwidth compression $\alpha$ is expressed as a ratio of integers, in other words $\alpha = \frac{b}{c}$ where $b, c \in \mathbb{Z}_{>0}$ with $b < c$. To simplify the forthcoming discussion and without loss of generality, let it be assumed that no oversampling is used ($\rho = 1$) yielding $Q = N$. The transmitters shown in Fig. 6 are then designed to offer diverse trade-offs.

Type 1, known as the *proportional inputs* transmitter, is the simplest method in that it employs a standard IDFT block of length $\frac{N}{\alpha}$. This means that the size of this block compared to a conventional OFDM system is $\frac{1}{\alpha}$ times larger. The number of input data symbols remains the same, in other words $N$ with the remaining $\frac{N}{\alpha} - N$

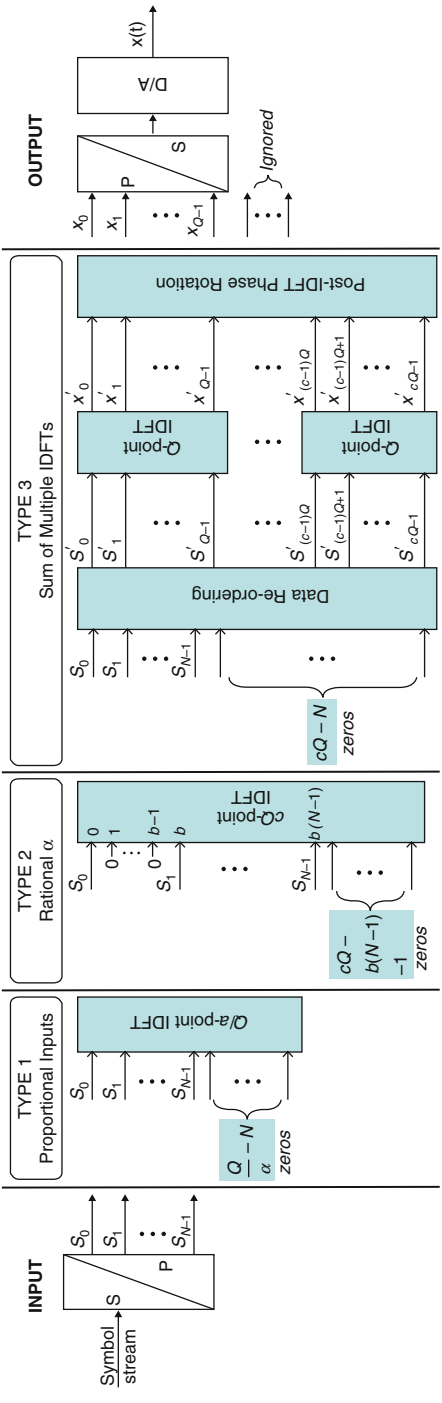


**Fig. 6** DFT based SEFDM transmitter types

IDFT inputs padded with zeros. The output of the IDFT is truncated, hence only $N$ samples are taken forward to construct the SEFDM signal while the remaining outputs are ignored. This results in the SEFDM signal transmitting the same amount of information over the same number of sub-carriers as OFDM but with the benefit of reduced bandwidth utilisation. The key drawback of the Type 1 transmitter is that it requires the ratio $\frac{N}{\alpha}$ to be an integer number to prevent performance degradation. Consequently, this limits the flexibility in choosing the value of $\alpha$.

For this reason, two alternative transmitter schemes were also proposed, namely Type 2 and Type 3, shown in Fig. 6. Type 2, known as the *rational* $\alpha$ transmitter, uses a $cN$ length IDFT block. Only the inputs of the IDFT whose indices are integer multiples of $b$, in other words for $n = 0, b, 2b, \ldots, (N-1)b$ are linked to the input data symbols while the remaining IDFT inputs are padded with zeros. This is achieved via the following condition

$$s_n = \begin{cases} s_{\frac{n}{b}} & \text{if } n \; mod \; b = 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where *mod* is defined as the modulus of the remainder after division.

Complying with the condition of Eq. (5), the generation process for the Type 2 transmitter may be expressed as

$$x(m) = \frac{1}{\sqrt{N}} \sum_{n=0}^{cN-1} s_n e^{j2\pi \frac{mn}{cN}}. \tag{6}$$

The benefit of this transmitter technique is that $b$ and $c$ can be chosen independently to give any desired value of $\alpha$ for any number of sub-carriers $N$. The length of the IDFT block in this case is $c$ times larger compared to an equivalent OFDM block and $b$ times larger compared to the Type 1 SEFDM transmitter.

Thereby, the Type 3 transmitter, known as the *sum of multiple IDFTs*, was proposed. This transmitter uses $c$ separate IDFT blocks each of length $N$ instead of one large $cN$ IDFT block. The advantage of this transmitter over the Type 2 architecture is two-fold; first, it allows multiple IDFT blocks to be configured in parallel, a feature attractive for hardware implementation employing Field Programmable Gate Arrays (FPGAs); second, it reduces the computation time required for the IDFT processing stage provided that all $c$ IDFT blocks can operate concurrently in parallel. It is important to note that the Type 3 transmitter requires two additional operations, namely a reordering stage before the IDFT block, as well as a post-IDFT phase rotation stage. The expression for the Type 3 transmitter is therefore given by

$$x(m) = \frac{1}{\sqrt{N}} \sum_{u=0}^{c-1} e^{j2\pi \frac{um}{cN}} \sum_{n=0}^{N-1} s_{u+nc} e^{j2\pi \frac{mn}{N}}. \tag{7}$$

### 3.2 SEFDM Signal Demodulation

As illustrated in Fig. 4, the SEFDM signal $x(t)$ is transmitted through an AWGN channel and arrives at the receiver as

$$r(t) = x(t) + z(t), \tag{8}$$

where $z(t)$ represents the noise impairment introduced by the channel. The output of the receiver's demodulator is acquired by correlating the received signal with a conjugate set of correlation functions $v_n(t)$, thus giving

$$R_n = \int_0^T r(t)v_n^*(t)dt, n = 0, \ldots, N-1. \tag{9}$$

The receiver model expressed mathematically by Eq. (9) is valid given two key assumptions; first, the demodulator is a linear filter to prevent noise colouring; second, the noise is Gaussian, in other words random and not burst, the latter being the case in a multipath propagation scenario due to fading and Doppler effects.

Assuming ideal sampling and ideal timing, the linear statistical model of Eq. (9) may be represented in matrix form as

$$\mathbf{R} = (\mathbf{X} + \mathbf{Z})\mathbf{V}^H, \tag{10}$$

where $\mathbf{R} = [R_n]$ is the $N \times 1$ vector of observation statistics. Recalling from Eq. (4), each discrete SEFDM symbol consists of $Q$ time samples, hence the transmitted signal may be expressed in matrix form as

$$\mathbf{X} = \mathbf{\Phi}\mathbf{s} \tag{11}$$

where $\mathbf{X} = [X_m]$ is a $Q$-dimensional vector of time samples, $\mathbf{s} = [s_n]$ is an $N$-dimensional vector of complex transmitted data symbols and $\mathbf{\Phi}$ is a $Q \times N$ sub-carrier matrix with elements equal to $e^{\frac{j2\pi mn\alpha}{Q}}$. The vector $\mathbf{Z} = [Z_m]$ represents $Q$ noise samples, which are complex independent and identically distributed (i.i.d.) Gaussian samples with mean $\mu = 0$ and variance $\sigma^2 = \frac{\mathcal{N}_0}{2}$, where $\mathcal{N}_0$ is the noise spectral density. Finally, $\mathbf{V}$ is the matrix representation of the basis functions where $\mathbf{V}^H$ denotes the Hermitian, or otherwise complex conjugate transpose, of matrix $\mathbf{V}$.

The functions $v_n(t)$ must be orthonormal or orthogonal [45] to satisfy the linear filtering requirement pertaining to Eq. (9). Two methods have been proposed to generate the basis functions $v_n(t)$. The first method uses *orthonormalisation procedures*, such as Gram-Schmidt (GS) or Löwdin processes, to generate an orthonormal base from a set of linearly independent but non-orthogonal sub-carriers [23, 45]. The second method uses *matched filtering*, as defined in [1]. The two methods are equivalent, as established in [1].

In matched filtering, the conjugate basis functions $v_n^*(t)$ equal the complex conjugate sub-carriers $\mathbf{\Phi}^H$, in other words

$$R_n = \int_0^T r(t)v_n^*(t)dt = \int_0^T r(t)e^{-j2\pi n\alpha\Delta ft}dt, n = 0, \ldots, N-1. \tag{12}$$

Subsequently, Eq. (10) becomes

$$\mathbf{R} = \mathbf{\Phi}^H\mathbf{X} + \mathbf{\Phi}^H\mathbf{Z} = \mathbf{\Phi}^H\mathbf{\Phi}\mathbf{s} + \mathbf{\Phi}^H\mathbf{Z} = \mathbf{C}\mathbf{s} + \mathbf{Z}_{\mathbf{\Phi}^H} \tag{13}$$

where $\mathbf{C}$ is the sub-carriers correlation matrix and describes the self-created interference between adjacent sub-carriers, as explained in Sect. 2.2, while $Z_{\mathbf{\Phi}^H}$ is the AWGN correlated with the conjugate sub-carriers.

If instead of matched filtering we used orthonormalisation procedures, the sub-carriers correlation matrix $\mathbf{C}$ would be replaced with the projections matrix $\mathbf{D}$, where $\mathbf{D}$ is given by

$$\mathbf{D} = \mathbf{B}^H\mathbf{\Phi}, \tag{14}$$

and $\mathbf{B}$ is the matrix representation of the orthonormal basis functions.

In conclusion, the linear statistical model for the receiver may be expressed in general form as

$$\mathbf{R} = \mathbf{H}\mathbf{s} + \mathbf{Z}, \tag{15}$$

where $\mathbf{H}$ is termed the interference matrix and corresponds to either the projections matrix $\mathbf{D}$ or the sub-carriers correlation matrix $\mathbf{C}$ depending on the basis functions $v_n(t)$ employed at the receiver.

### *3.3 SEFDM Signal Detection*

Figure 7 depicts a practical transceiver comprising three different stages for the demodulation and detection of SEFDM signals. The first stage is a demodulator which, as described in Sect. 3.2, may be implemented using orthonormalisation processes or matched filtering. Matched filtering is preferred since it is the optimum receiving filter and can be realised using DFT operations. The second stage involves the use of linear or iterative techniques while the third stage employs more sophisticated algorithms, such as Maximum Likelihood (ML) or SD. A snapshot of the demodulation and detection methods employed in SEFDM is depicted in Fig. 8.

Section 3.2 explained that the SEFDM demodulator is responsible for generating the statistics vector $\mathbf{R}$. Each element of this vector $\mathbf{R}$ is merely a discrete sample which has a value proportional to the energy of the received symbol with the added

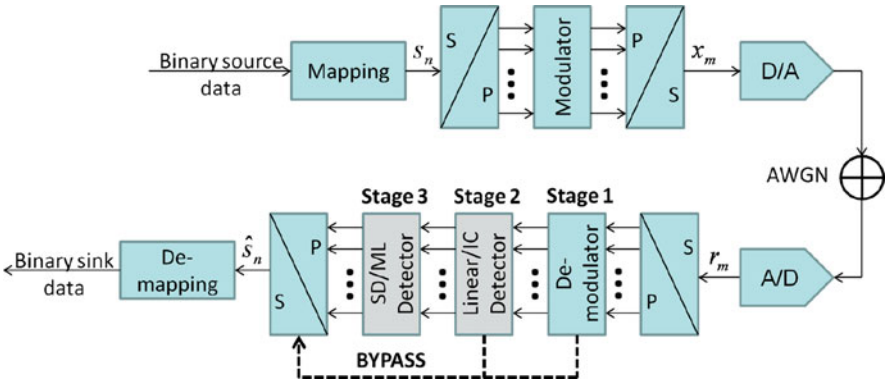


**Fig. 7** Practical SEFDM transceiver

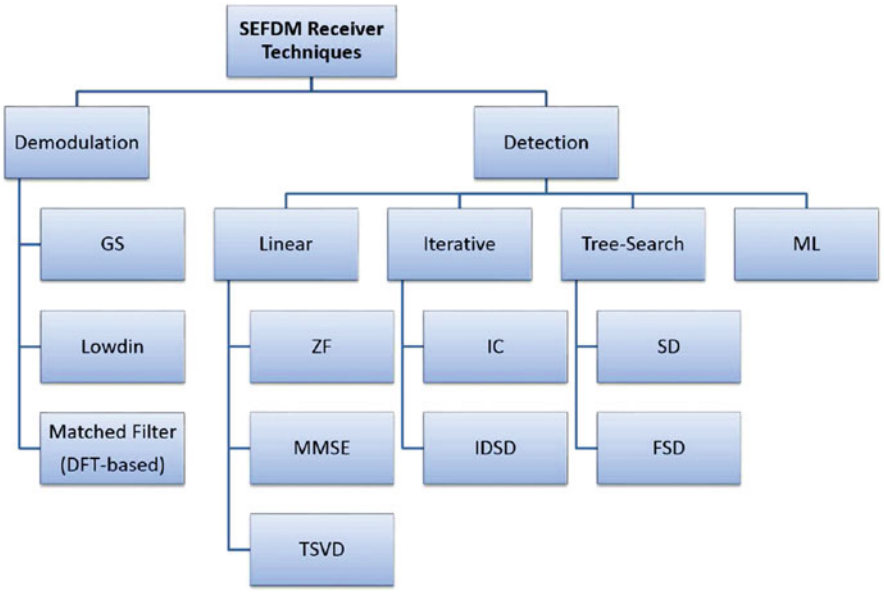


**Fig. 8** SEFDM demodulation and detection techniques

noise. The task of the detector is to obtain the best possible estimate $\hat{\mathbf{s}}$ of each transmitted data symbol, where $\hat{\mathbf{s}}$ denotes the estimate of the transmitted signal vector $\mathbf{s}$ after slicing, based on the values of $\mathbf{R}$ and the properties of the sub-carriers correlation matrix $\mathbf{C}$.

Recalling from Eq. (15), the linear statistical model for the received SEFDM signal is given by $\mathbf{R} = \mathbf{Hs} + \mathbf{Z}$. Linear detection methods treat this relation as an unconstrained linear estimation problem and thus aim to recover the transmitted symbols by cancelling out the self-created interference. The most popular methods

include Zero Forcing (ZF) and Minimum Mean Squared Error (MMSE) with the former yielding the Least Squares (LS) solution. If the estimator matrix used to generate the symbol estimates is $\mathbf{G}$, then the solution to the unconstrained problem with respect to Eq. (15) is given by

$$\tilde{\mathbf{s}} = \mathbf{GR}, \tag{16}$$

with $\tilde{\mathbf{s}} \in \mathbb{C}^N$, where $\tilde{\mathbf{s}}$ denotes the unconstrained estimate of the transmitted signal vector $\mathbf{s}$ before slicing and $\mathbb{C}^N$ is the set of all complex $N$-tuples[3] assuming a complex constellation scheme is employed, for example 4-QAM. If the constellation scheme contains only real values, such as BPSK, the set $\mathbb{C}^N$ may be reduced to the set $\mathbb{R}^N$.

The solution of Eq. (16) is referred to as the unconstrained estimate since it does not take into account that the elements of $\mathbf{s}$ can in fact take only the values pertaining to a limited set of constellation points defined in a discrete alphabet $\mathfrak{M}^N$. To this end, the problem is constrained by applying a slicing operation to each of the entries of $\tilde{\mathbf{s}}$ yielding

$$\hat{\mathbf{s}} = \lfloor \tilde{\mathbf{s}} \rceil , \tag{17}$$

with $\hat{\mathbf{s}} \in \mathbb{Z}^N$ where $\mathbb{Z}^N$ denotes the set of integer $N$-tuples and $\lfloor \cdot \rceil$ denotes a slicing operator, which rounds the value to the nearest constellation point.

### 3.3.1 Linear Detection

The ZF detector generates symbol estimates by forcing the interference terms to zero. Even though ZF completely eliminates interference, it could potentially amplify the noise to a great extent, especially at points where spectral nulls occur [22], leading to severe performance degradation. MMSE improves performance over ZF by taking the presence of noise into account with the aim of minimising the total expected error.

Previous work has shown that the ill-conditioning of the interference matrix significantly degrades the estimates generated using the ZF or MMSE detectors [1, 45]. Moreover, when the interference matrix is singular it becomes non-invertible, thus the ZF technique cannot be employed. A technique termed Truncated Singular Value Decomposition (TSVD) was therefore proposed to address this challenge [41].

The TSVD approach generates an approximate pseudoinverse $\mathbf{H}_\xi$ of the interference matrix $\mathbf{H}$. The truncation of the small singular values renders $\mathbf{H}_\xi$ less sensitive to perturbations yielding a solution with higher confidence levels.

[3] An $N$-tuple is a sequence or ordered set of $N$ elements.

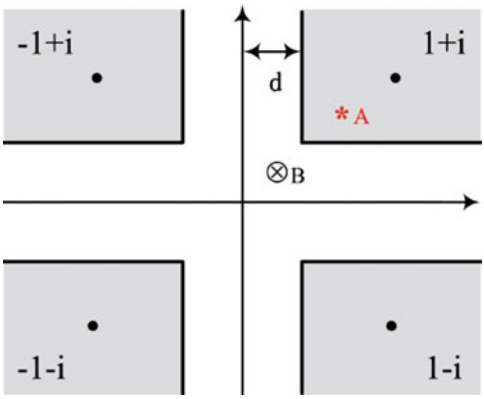


**Fig. 9** Soft mapping principle

### 3.3.2 Iterative Detection

Iterative Cancellation (IC) techniques were applied by Kanaras in [45] and by Grammenos in [24] to the projections matrix **D** and to the sub-carriers correlation matrix **C**, respectively, to generate symbol estimates. Having generated an upper triangular matrix, the symbol estimates are obtained via back-substitution. IC results in improved Bit Error Rate (BER) performance over ZF and TSVD. This is due to the fact that IC constrains the symbol estimate on each iteration, in other words it takes into account that each entry of **s** pertains to a value from a finite set of constellation points $\mathscr{M}$. For this reason, such IC techniques are classified as *hard mapping* detectors.

An improved iterative detector which uses *soft mapping* was applied to SEFDM first in [32] and subsequently in [84]. The soft mapping principle is illustrated in Fig. 9. The white uncertainty interval is defined by $d = 1 - \frac{m}{\upsilon}$, where $m$ is the $m_{th}$ iteration and $\upsilon$ denotes the number of iterations. Only points that fall in the grey area can be mapped to the corresponding constellation point. Other points remain unchanged and left to the next iteration. It should be noted that the uncertainty area is reduced after each iteration. Hence, the algorithm always converges to a solution so long as a sufficient number of iterations takes place.

### 3.3.3 Maximum Likelihood and Sphere Decoding

It is well known that ML is the optimum detector for achieving the theoretical error performance. The detection problem in this case may be expressed as follows

$$\hat{\mathbf{s}}_{ML} = \arg \min_{\mathbf{s} \in \mathscr{M}^N} \|\mathbf{R} - \mathbf{H}\mathbf{s}\|^2, \tag{18}$$

which translates into finding the vector symbol $\hat{\mathbf{s}}$ from all possible transmitted vector symbols $\mathbf{s} \in \mathscr{M}^N$ that minimises the Euclidean distance, in other words the vector symbol that has most likely been transmitted. Assuming that complex Gaussian noise has been added to each symbol and that all transmitted vector symbols

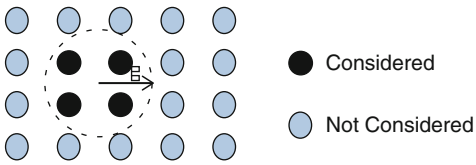


**Fig. 10** Sphere search constrained by the radius $g$

have equal probability of occurring, the ML algorithm searches over the entire set $\mathbf{s} \in \mathscr{M}^N$ to find $\hat{\mathbf{s}}$. Hence, it is evident that this hinders a practical implementation of ML.

SD was examined [45] as a more viable solution for recovering SEFDM signals successfully with tolerable error penalties since it reduces the ML solution to a problem of polynomial complexity. SD solves the same problem as ML, yet in this case, the search plane is limited to an N-dimensional hypersphere, whereby only the vector symbols that lie within this hypersphere are considered as possible solutions. Thereby, Eq. (18) becomes

$$\hat{\mathbf{s}}_{SD} = \arg \min_{\mathbf{s} \in \mathscr{M}^N} \|\mathbf{R} - \mathbf{H}\mathbf{s}\|^2 \leqslant g, \tag{19}$$

where $g$ is a scalar denoting the radius of the sphere which is centred around the statistics vector $\mathbf{R}$, as illustrated in Fig. 10. SD is classified as a tree-search algorithm because the constellation points refer to the nodes of the tree and the algorithm is executed by traversing the levels and branches of the tree until a complete path is found.

The conventional SD algorithm has a variable computational complexity which changes according to the noise in the channel and the properties of the interference matrix. For this reason, the Fixed Sphere Decoding (FSD) was proposed in [41] and its algorithm is outlined in Fig. 11. The FSD guarantees a constant throughput by fixing the complexity of the conventional SD algorithm. This complexity is independent of the noise present in the system, as well as the ill-conditioning of the interference matrix.

The complexity of the algorithm is user-defined by setting the parameter $\zeta$, which defines the complexity order of the algorithm, with $W = (\sqrt{M}^{\zeta})$ where $W$ denotes the tree width. During the initialisation phase, the first $\zeta$ levels are fully expanded. In the subsequent levels, only the best $W$ candidate nodes are expanded. The tree width determines the complexity of the FSD, as it defines the number of node visits per level. The larger the tree width, the better the performance at the expense of increased complexity. From Fig. 11, it should be evident that the update interval centres $\theta_{ij}$ at each level can be computed independently in parallel rendering the FSD algorithm an attractive choice for hardware implementation using FPGAs.

The sorting stage which performs comparisons amongst all candidate nodes $B_i$ at each level constitutes one of the main bottlenecks. The larger the tree width, the higher the number of comparisons. This leads to an increase in the overall latency of the algorithm and subsequently reduction in throughput. On the other hand, reducing

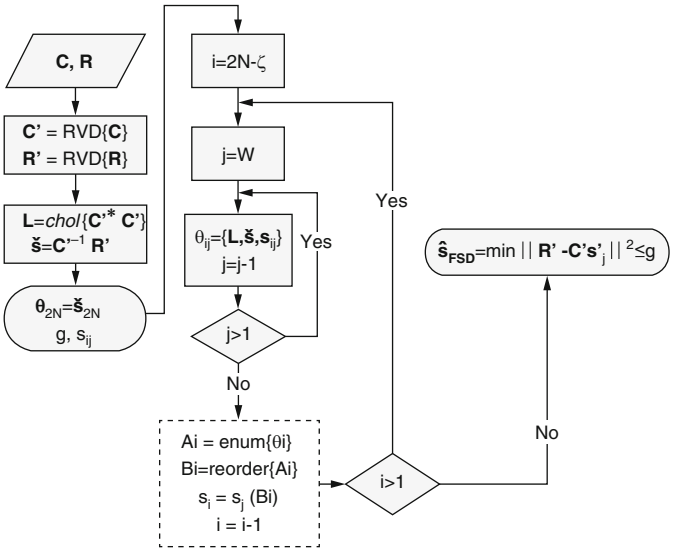


**Fig. 11** Flowchart of the FSD algorithm for the detection of SEFDM signals

the tree width or removing the sorting stage affects the order in which a parent node branches out to a child node, thus degrading BER performance. Hence, to strike a balance between the conflicting objectives of good BER performance and tangible computational complexity, two techniques were applied; the first, uses Sort Free (SF) detection which reduces latency thanks to the simplified sorting stage implemented in the algorithm; the second technique, involves the utilisation of Modified Real Valued Decomposition (MRVD), which offers improved error rates. These techniques have been shown to deliver performance gains in Multiple Input Multiple Output (MIMO) systems [2].

Figure 12 compares the spanning tree for two variants of the FSD with different sorting strategies, namely the Shortest Path (SP), entailed in the original FSD algorithm, and the novel SF technique, demonstrating the simplified structure of the latter version. On the other hand, the improved performance offered by MRVD lies in the fact that, at each level, the algorithm flips between the real and imaginary parts of the next symbol in line. Therefore, MRVD can be regarded as providing 'more complete' solutions compared to the conventional Real-valued Decomposition (RVD).

### 3.3.4 Numerical Results

In this section, we compare different detection schemes in terms of error performance for varying system sizes and levels of bandwidth compression. Table 2 summarises the system parameters employed in the simulations.

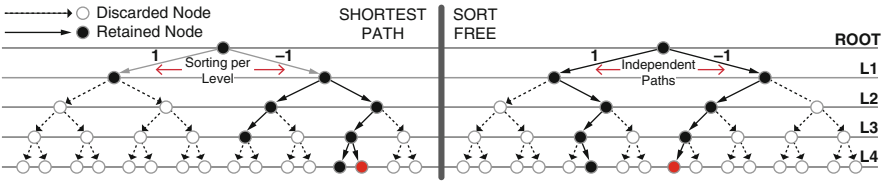


**Fig. 12** Tree of Shortest-Path FSD (*left*) vs. Sort-Free FSD (*right*) ($N = 2$, $W = 2$)

**Table 2** Simulation parameters for different detection schemes

| Parameter | Value |
|---|---|
| Total number of sub-carriers $N_{IFFT}$ | 32 / 64 |
| Used sub-carriers $N$ | 16 |
| Bandwidth compression factor $\alpha$ | 0.7 / 0.8 |
| Modulation scheme | BPSK / 4-QAM |
| Oversampling factor $\rho$ | 1 / 2 |
| $E_b/N_o$ range | 1–20 dB |
| Number of runs | 10,000 |

The simulation methodology adopted has the following characteristics:

- The source data is generated using a uniform random number generator.
- All results are averaged over 10,000 randomly generated SEFDM symbols unless otherwise stated.
- The bits are mapped to symbols using either BPSK or 4-QAM unless otherwise stated. In the case of 4-QAM, it is assumed that Gray coding is employed.
- The number of bits processed is approximately equal to $\frac{10}{DesiredBER}$.
- The channel is linear, discrete-time and time-invariant.
- The only source of error is due to AWGN.[4]
- It is assumed that the receiver has perfect knowledge of the interference matrix $\mathbf{H}$ at all times under all conditions and that decisions yielding symbol estimates are made strictly based on the received amplitude of the noisy signal.

The legend keys that appear in the figures correspond to the following descriptions:

- *Matched Filtering (MF)*: The symbol estimates are given by $\hat{\mathbf{s}}_{MF} = \lfloor \mathbf{R} \rceil = \left\lfloor \mathbf{\Phi}^H \mathbf{r} \right\rceil$ where $\mathbf{r}$ is a vector representing discrete time samples of the received signal having propagated through an AWGN channel.
- *Matched Filtering followed by Zero Forcing (MF-ZF)*: The symbol estimates are given by $\hat{\mathbf{s}}_{ZF} = \left\lfloor \mathbf{C}^{-1}\mathbf{R} \right\rceil$.

[4]Results showing the performance of SEFDM in fading channels have been published in the work of Chorti et al. [11] and Isam et al. [40].

- *Demodulation using Iterative Modified Gram Schmidt (IMGS) processes [45] followed by Zero Forcing (IMGS-ZF)*: The symbol estimates are given by $\hat{\mathbf{s}}_{ZF_{IMGS}} = \lfloor \mathbf{D}^{-1}\mathbf{R} \rceil$.
- *Matched Filtering followed by Truncated Singular Value Decomposition (MF-TSVD)*: The symbol estimates are given by $\hat{\mathbf{s}}_{TSVD} = \lfloor \mathbf{C}_{\blacksquare}\mathbf{R} \rceil$.
- *Matched Filtering followed by Iterative Cancellation (MF-IC)*: The symbol estimates are given by $\hat{\mathbf{s}}_{IC} = \hat{s}_m$ with $m = 1, \ldots, N$ and $\hat{s}_m = \left\lfloor \frac{1}{c_{m,m}}(R_m - \sum_{n=m+1}^{N} c_{m,n}\hat{s}_n) \right\rceil$.
- *Demodulation using Iterative Modified Gram Schmidt processes followed by Iterative Cancellation (IMGS-IC)*: The symbol estimates are given by $\hat{\mathbf{s}}_{IC_{IMGS}} = \hat{s}_m$ with $m = 1, \ldots, N$ and $\hat{s}_m = \left\lfloor \frac{1}{d_{m,m}}(R_m - \sum_{n=m+1}^{N} d_{m,n}\hat{s}_n) \right\rceil$.
- *Demodulation using Löwdin processes [45] followed by Iterative Cancellation (Lowdin-IC)*: The symbol estimates are given by $\hat{\mathbf{s}}_{IC_{Lowdin}} = \hat{s}_m$ where $\hat{s}_m = \left\lfloor \frac{1}{l_{m,m}}(R_m - \sum_{n=m+1}^{N} l_{m,n}\hat{s}_n) \right\rceil$, with $m = 1, \ldots, N$ and where $l_{m,n}$ denotes the elements of the Hermitian matrix generated using the Löwdin orthonormalisation process.
- *Matched Filtering followed by Minimum Mean Squared Error (MF-MMSE)*: The symbol estimates are given by $\hat{\mathbf{s}}_{MMSE} = \left\lfloor \mathbf{C}\left(\mathbf{C}\mathbf{C}^H + \frac{\sigma^2}{\sigma_s^2}\mathbf{I}\right)^{-1}\mathbf{R} \right\rceil$.
- *Iterative Detection with soft decision (ID)*: The symbol estimates are given by $\hat{\mathbf{s}}_{ID} = S_n$ with $S_n = \lambda\mathbf{R} + (e - \lambda\mathbf{C})S_{n-1}$ where $S_n$ is an N-dimensional vector of recovered symbols after $n$ iterations, $S_{n-1}$ is an N-dimensional vector of estimated symbols after $n-1$ iterations, $e$ is an $N \times N$ identity matrix and $\lambda$ is a convergence factor that determines how many iterations are used in the detection.

Figure 13 shows that for BPSK, the IC schemes using MF or Löwdin orthonormalisation offer the best and commensurate error performance. For Signal-to-Noise Ratio (SNR) regimes where $\frac{E_b}{N_0} \geqslant 14\,\text{dB}$, the IC method employing IMGS approaches the performance of TSVD and MMSE. A key observation from Fig. 13 is that MF on its own can provide the same BER performance than both MF-TSVD and MF-MMSE. This is an important result as it allows the linear stage in Fig. 7 to be bypassed, should one require the lowest complexity receiver (a straightforward demodulator) at the expense of error performance when compared to an IC receiver.

Figure 14 compares the BER of the original FSD detector against its variants. It is clear that TSVD and MRVD when separately applied to the standalone SF yield significant performance gains, especially in the higher SNR regimes. When both techniques are combined and applied to SF, the resultant TSVD-MRVD-SF FSD detector approaches the performance of the more optimum, shortest-path TSVD-FSD detector, yet with the benefit of reduced computational complexity.

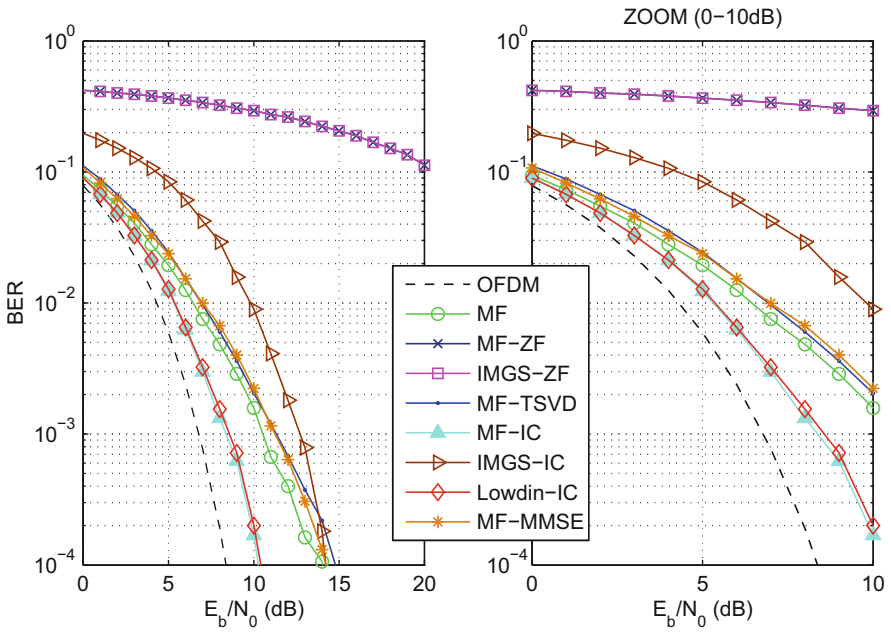


**Fig. 13** BER of different detection schemes for BPSK with $N = 16$, $\alpha = 0.8$ and $\rho = 2$

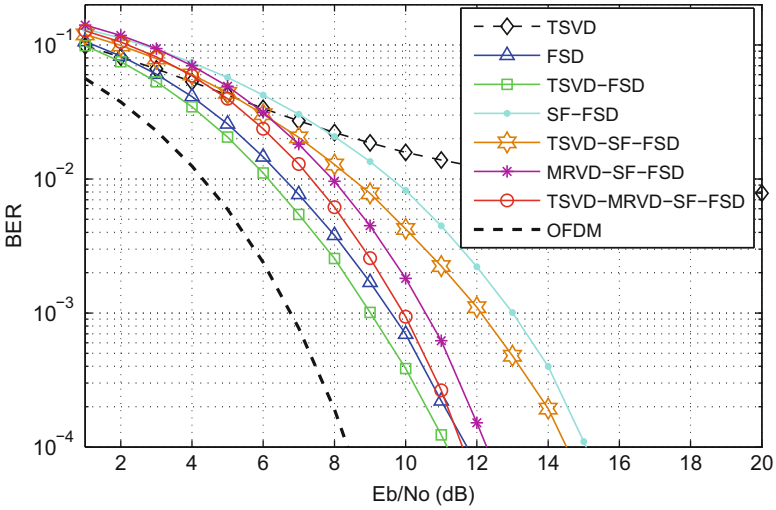


**Fig. 14** BER performance for the different FSD variants (4-QAM, $\alpha = 0.8$, $N = 16$ and $W = 16$)

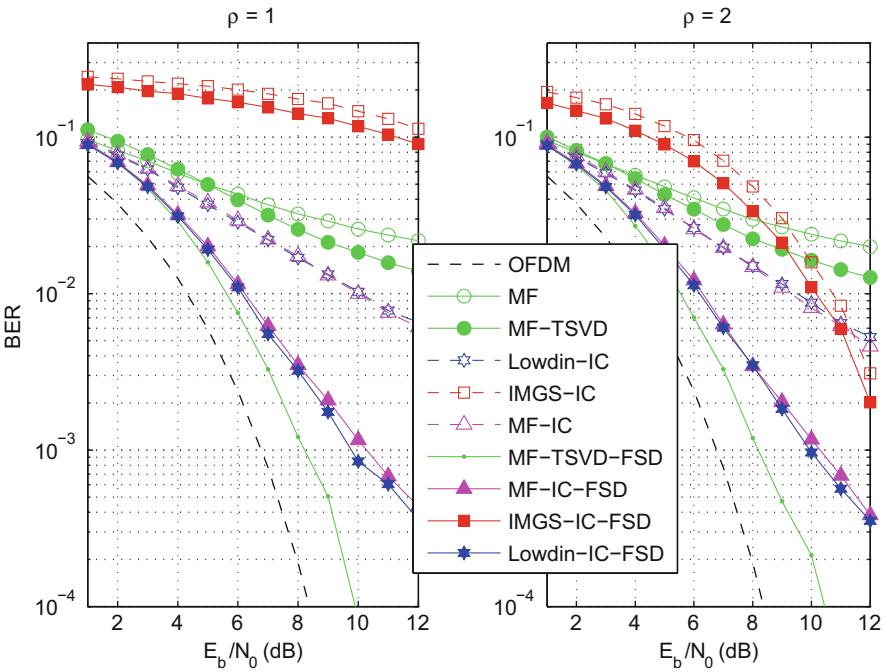


**Fig. 15** BER of hybrid and non-hybrid detection schemes for $\rho = 1$ (*left sub-plot*) and $\rho = 2$ (*right sub-plot*) ($N = 16$, $\alpha = 0.8$, FSD tree width $W = 16$)

Figure 15 shows that the hybrid TSVD-FSD detector outperforms all other hybrid and non-hybrid detectors (which employ hard mapping). Figure 16, however, demonstrates that the soft iterative detector combined with FSD (ID-FSD) is superior to TSVD-FSD for different bandwidth compression factors. More importantly, when $\alpha = 0.8$, ID-FSD offers near-optimum performance.

## *3.4 Coded SEFDM*

Section 3.3 demonstrated that iterative techniques, such as iterative cancellation and iterative detection with soft decision, offer good BER performance. As the system size increases, however, giving rise to higher interference power, even these methods fail to recover SEFDM signals successfully.

To address this challenge, Turbo equalisation was introduced in the SEFDM receiver [78], as an improved means for mitigating the self-created interference. A detailed description of the Turbo equalisation principle may be found in [28]. A functional block diagram of the Turbo-SEFDM system is illustrated in Fig. 17. This system architecture was conceptually described in [78, 81] and experimentally

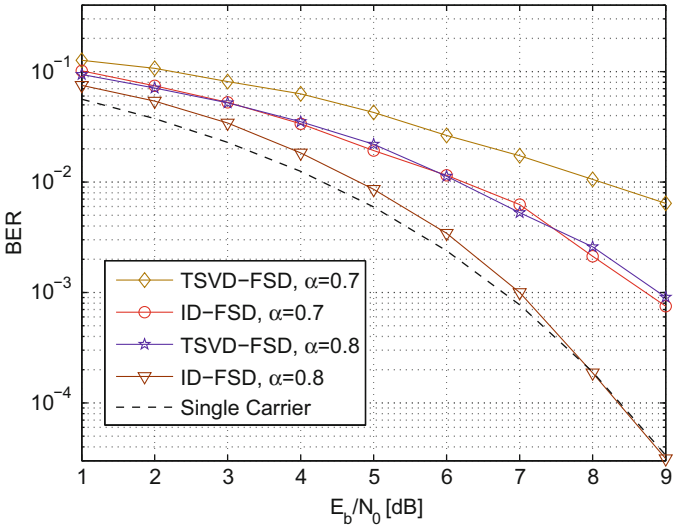


**Fig. 16** BER performance for ID-FSD and TSVD-FSD detectors carrying 4QAM symbols for $N = 16$, $\upsilon = 20$ and $W = 16$

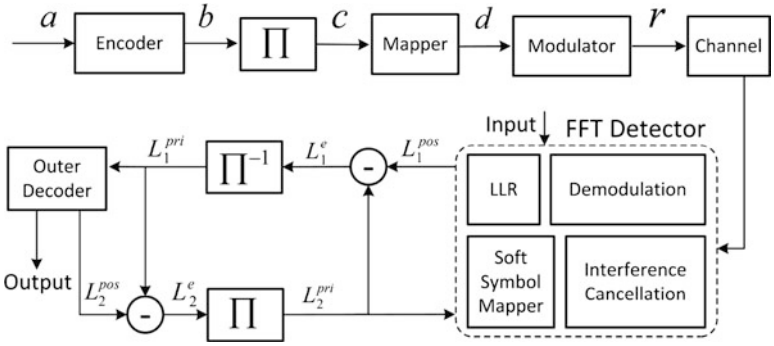


**Fig. 17** Functional block diagram of the Turbo-SEFDM system

evaluated in [82]. Here, we present a summary of the proposed system and interested readers may resort to the aforementioned references for a detailed description.

At the transmitter side, the bit stream labelled **a** is encoded using either convolutional codes or Turbo codes. Subsequently, the coded bit stream labelled **b** is permuted within the interleaver module **Π**. The interleaved bit stream **c** is then mapped to complex symbols **d** and finally modulated **r** using an SEFDM modulator [75].

At the receiver side, the signal is demodulated using an SEFDM FFT based demodulator [75]. The log-likelihood ratio (LLR) unit generates soft information, referred to as *a posteriori* information and denoted by $L_1^{pos}$, for each bit. The sign of the LLR value reflects the sign of the bit while its magnitude determines

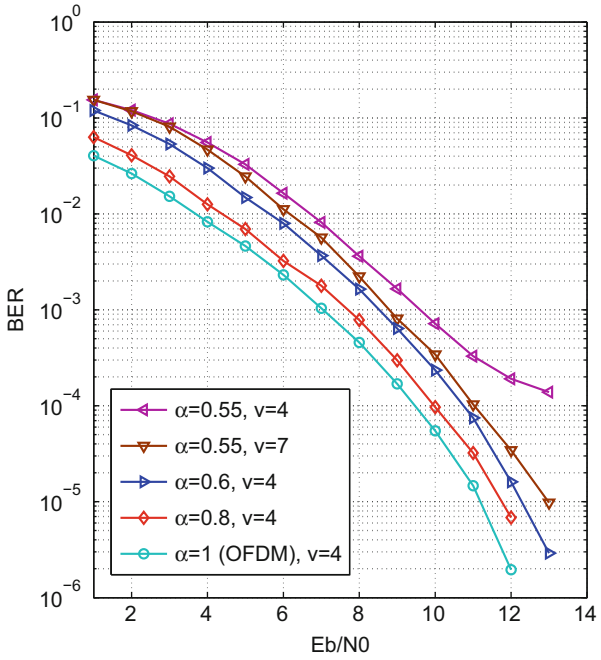


**Fig. 18** Performance of Turbo-SEFDM in a multipath fading channel with $N = 1024$ at various $\alpha$

the reliability of the sign of the bit. This soft information is then transformed to extrinsic information, $L_1^e$ by subtracting the *a priori* information, $L_2^{pri}$. This extrinsic information is exchanged between the detector and the outer decoder in an iterative fashion. On each iteration, the interference represented by the updated soft symbols, is subtracted from the original demodulated symbols. Hence, the receiver maximises the *a posteriori* probability (APP) for a given bit.

Figures 18 and 19 demonstrate the performance of Turbo-SEFDM in a frequency selective fading channel where $h(t) = 0.8765\delta(t) - 0.2279\delta(t - T_s) + 0.1315\delta(t - 4T_s) - 0.4032e^{\frac{i\pi}{2}}\delta(t - 7T_s)$. It is assumed that perfect channel state information (CSI) is known at the receiver. Figure 18 illustrates that the bandwidth can be compressed up to 45 % with an error penalty of up to 2 dB using only four to seven iterations. Figure 19 shows that the best performance can be achieved after just four iterations for a bandwidth saving of 40 %.

## 4 SEFDM Implementation Aspects

The dimension of the SEFDM problem presents significant challenges compared to the application of similar detection methods in MIMO systems. In this section, we discuss the complexity and performance trade-offs for the different detectors examined in the previous sections. We also explore testbeds that were devised in

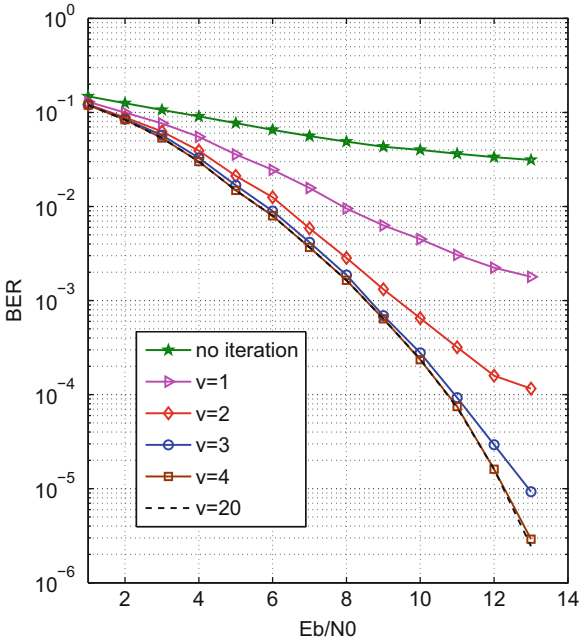


**Fig. 19** Performance of Turbo-SEFDM in a multipath fading channel with $\alpha = 0.6$ for a different number of iterations

recent years for the implementation of SEFDM, as well as the recent application and experimental demonstration of SEFDM in other systems, notably in Long Term Evolution (LTE), optical communications and radio-over-fibre (RoF).

## *4.1 Performance and Complexity Trade-offs*

SEFDM increases multiplexing gain with respect to an OFDM system by permitting the use of more sub-carriers in the default OFDM bandwidth. This gain, however, comes at the expense of complex signal processing at the receiver which is required in order to separate the sub-carriers suffering from self-created ICI. A linear increase in the number of sub-carriers, however, results in a more than linear increase in the complexity order of both the demodulation and detection stages.

### 4.1.1 In Demodulation, Linear and Iterative Detection

A demodulator employing GS or Löwdin orthonormalisation has the same complexity order as a conventional analogue multi-carrier system equal to $\mathcal{O}(N^2)$. The matched filter technique addressed this issue by substituting these processes with DFT operations, thus allowing the use of the Fast Fourier Transform (FFT) algorithm reducing the complexity order to $\mathcal{O}(Nlog_2N)$. The linear detection

methods explored in Sect. 3.3 involve matrix inversion or decomposition. Analytical tools like MATLAB, generally apply Gaussian elimination to compute the inverse of a matrix, such as the computation required in the LS solution of the ZF estimate. The TSVD detector uses the standard Singular Value Decomposition (SVD) matrix decomposition method to generate an approximate inverse. In both cases, the order of complexity is assumed to be cubic $\mathcal{O}(N^3)$ [8, 45]. The inverted matrix is then multiplied by the statistics vector $\mathbf{R}$, which implies that the total number of multiplications amounts to $N^2$. The slicing operator acts as a simple threshold detector, therefore its complexity is considered to be negligible compared to the complexity of the prerequisite matrix operations.

Conversely, iterative detection techniques operate on upper triangular matrices, where the total number of multiplications equals $\frac{N}{2}(N+1) \approx \frac{N^2}{2}$, which is less than the number of multiplications required for the ZF or TSVD techniques. Iterative detection requires the inversion of only the diagonal elements of the matrix.

In summary, the benefits and drawbacks of using linear or iterative detection are as follows:

- **Linear detection**: The fundamental difficulty in this case lies in the inversion of the sub-carriers correlation matrix $\mathbf{C}$ which has a large dimension equal to $N \times N$ and is also complex-valued which instantly doubles the dimension of the problem to $2N \times N$. The computational complexity associated with the inversion or decomposition of matrices is often prohibitive in wireless systems [22]. $\mathbf{C}$ in SEFDM systems is deterministic and has constant values for a given configuration of $N$ and $\alpha$. Thereby, the elements of $\mathbf{C}^{-1}$ only need to be generated once, a process which can take place offline, and subsequently stored in block memories. The computation of the symbol estimates is then a straightforward complex-valued matrix multiplication between $\mathbf{R}$ and $\mathbf{C}^{-1}$. The default number of words required for the storage of the elements pertaining to $\mathbf{C}^{-1}$ is $N^2$. Matrix multiplication can be implemented efficiently through the concurrent execution of multiple multiply-accumulate operations using FPGAs.
- **Iterative detection**: Iterative detection alleviates the need for matrix inversion which means that the results suffer less from round-off errors. The performance of iterative detection largely depends on the correctness of the first symbol decoded since any errors during the initial stages of the algorithm's execution will propagate through the iterative process and could potentially lead to severe BER degradation. It does, however, require a cascaded architecture which hinders parallel processing. One possible solution to this issue is to decompose the entire iteration process into a single step. In FPGA terms, this decomposition would necessitate combinatorial logic, which is generally avoided as it increases the load on the clock severely limiting the maximum achievable clock frequency.

Hence, based on the above, it is evident that linear and iterative schemes present different challenges with varying trade-offs. The choice depends on the specific

technology adopted to implement these algorithms in practice. FPGAs offer significant performance gains over conventional Digital Signal Processors (DSPs) provided that their parallel processing capabilities are leveraged.

### 4.1.2 In Sphere Decoding

The complexity of an algorithm is determined by two aspects; first, the number of arithmetic operations required to execute the algorithm, such as multiplications, additions and subtractions; second, the number of compare-select operations that take place during the sorting stage.

While the conventional SD provides a significant reduction in complexity compared to ML, its execution is highly sensitive to noise and the interference properties of SEFDM systems. An authoritative account of the complexity associated with SD is given by Hassibi and Vikalo [31], as well as Jalden and Ottersten [42]. There is also rich literature discussing these issues in MIMO systems, a notable example being the work presented by Viterbo and Boutros [72].

This variable complexity associated with the conventional SD renders it inapt for hardware realisation. In addition, the sequential nature of the algorithm does not fit well with the parallel processing capabilities of FPGAs. For this reason, the FSD was explored which fixes the algorithm's execution time and has an architecture which is better tailored for FPGA implementation. A key enhancement of the FSD was the simplification of the sorting stage, which was found to be the main bottleneck in the algorithm's execution.

Table 3 compares the total number of operations required at each tree level for the SP and SF detection schemes. The depth of the tree, commensurate with the system dimension, is denoted by $D$ which is equal to $2N$ as a result of using RVD. From this table, it is evident that the number of comparisons at each level is fixed and independent of the system dimension. For the SF method, a single comparator is required for each node to compute the minimum distances to the centre point. The number of comparisons is equal to the tree width. On the contrary, the SP technique requires multiple cascaded minimum finders to perform the reordering of the candidate nodes with respect to the minimum distance from the centre point. In both cases, the number of comparisons is dictated by the tree width.

**Table 3** Total number of operations per level for the Shortest-Path and Sort-Free FSD detectors

| Metric | Shortest-path (SP) | Sort-free (SF) |
|---|---|---|
| Multiplications | $W(D-k)$ | $W(D-k)$ |
| Additions | $W(2(D-k)+1+\sqrt{M})$ | $W(2(D-k)+1)$ |
| Comparisons | $\frac{1}{2}(MW^2+\sqrt{M}W+4W)$ | $W$ |

### *4.2 Experimental Testbeds*

Much work has been done with respect to the hardware development of an SEFDM transceiver. Independent FPGA realisations of a reconfigurable SEFDM transmitter and ZF\TSVD receiver have appeared in [62] and [25], respectively. In [62], the Type 3 transmitter, discussed in Sect. 3.1, was implemented using multiple Inverse Fast Fourier Transform (IFFT) blocks configured in parallel, thus achieving the lowest latency implementation possible.

Subsequently, a route to a Very Large Scale Integration (VLSI) implementation of a reconfigurable SEFDM transmitter in 32-nm CMOS was published in [75] in which a pruned IFFT architecture was introduced. Running at a clock frequency of 100 MHz, the throughput can reach 17.4 Mbps with QPSK modulation, and up to 52.2 Mbps with 64QAM modulation, with a bandwidth saving of up to 50 % relative to OFDM. Here, it is also worth mentioning that VLSI implementations of FTN transceivers have also come to light [15].

A joint hardware implementation of the hybrid TSVD-FSD receiver with the aid of FPGAs and DSPs was presented in [26]. A high-level representation of the complete hardware system is depicted in Fig. 20. The linear detector, details of which are published in [25], was realised on an FPGA device with a flexible and reconfigurable design supporting different system sizes, oversampling factors, modulation orders and levels of bandwidth compression while providing a theoretical data rate of up to 136.8 Mbps. On the DSP side, a commercially available DSP development environment was used to implement the SP and SF FSD detectors on the device and to handle the data transfer between the DSP and the computer terminal.

The SEFDM concept has been expanded and evaluated in other experimental systems. In [81], a multi-band architecture termed B-SEFDM was tested in a practical, LTE like environment. A new optical SEFDM (O-SEFDM) system was proposed in [14] and subsequently demonstrated in [13]. Results [13] showed that 4-QAM O-SEFDM can outperform 8-QAM optical OFDM by 1.6 dB. This latter work constitutes the first experimental verification of 25 % optical faster than Nyquist rate. More recently, a dual polarisation coherent optical fibre system employing O-SEFDM was developed [60]. This system is capable of operating at 24 Gb/s while occupying a total bandwidth ranging from 6 GHz (for $\alpha = 1$) to 4.02 GHz (for $\alpha = 0.67$).

Finally, the use of SEFDM in RoF was also demonstrated, for the first time, for beyond 4G cellular networking [57]. In this work, a proof-of-concept SEFDM-LTE type radio over multi-mode fibre transmission was shown to operate successfully at 36 Mb/s.

## 5 Conclusions

This chapter considered the design, implementation and performance evaluation of a novel multi-carrier modulation technique termed SEFDM, which improves spectrum efficiency at the expense of receiver complexity. The key challenge that

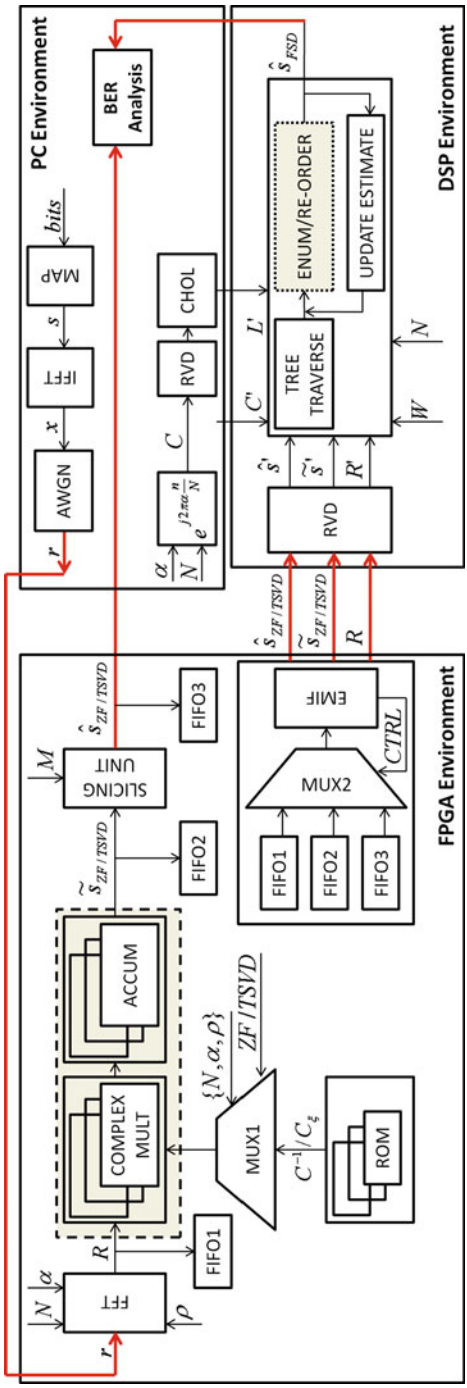


**Fig. 20** System block diagram for the hybrid TSVD-FSD hardware detector implementation

arises in SEFDM systems is the self-created interference which complicates the design of receiver architectures. To this end, sophisticated detection algorithms have been developed to address this challenge. Subsequently, these algorithms are optimised to make them better-suited for application in the real world. Finally, a number of hardware prototypes and experimental testbeds have been realised to assess the feasibility of employing SEFDM in practical systems including wireless, optical and RoF systems.

## References


1. S.I. Ahmed, Spectrally Efficient FDM Communication Signals and Transceivers : Design, Mathematical Modeling and System Optimization, PhD thesis, University College London (UCL), 2011
2. K. Amiri, C. Dick, R. Rao, J.R. Cavallaro, Flex-Sphere: an FPGA configurable sort-free sphere detector for multi-user MIMO wireless system, in *Proceedings of the Software Defined Radio Technology Conference*, Washington, DC, SDR, 2008
3. J.B. Anderson, F. Rusek, Improving OFDM : multistream faster-than-Nyquist signaling, in *4th International Symposium Turbo Codes and Related Topics, ISTC*, Munich, 2006
4. J.G. Andrews, S. Buzzi, W. Choi, S.V. Hanly, A. Lozano, A.C.K. Soong, J.C. Zhang, What will 5G be? IEEE J. Sel. Areas Commun. **32**(6), 1065–1082 (2014)
5. D.H. Bailey, The fractional Fourier transform and applications. SIAM Rev. **33**(3), 389–404 (1991)
6. A Barbieri, D. Fertonani, G. Colavolpe, Improving the spectral efficiency of linear modulations through time-frequency packing, in *IEEE International Symposium on Information Theory, ISIT*, Toronto, vol. 2, 2008, pp. 2742–2746
7. L. Bluestein, A linear filtering approach to the computation of discrete Fourier transform. IEEE Trans. Audio Electroacoust. **18**(4), 451–455 (1970)
8. A. Burg, VLSI Circuits for MIMO Communication Systems, PhD thesis, Swiss Federal Institute of Technology (ETH), Zurich, 2006
9. J.W. Burns, Measuring spectrum efficiency - The art of spectrum utilisation metrics. Tech. rep., Aegis Systems Ltd., 2002
10. A. Chorti, I. Kanaras, Masked M-QAM OFDM : a simple approach for enhancing the security of OFDM systems, in *IEEE 20th International Symposium Personal, Indoor and Mobile Radio Communication, PIMRC*, Tokyo, 2009, pp. 1682–1686
11. A. Chorti, I. Kanaras, M.R.D. Rodrigues, I. Darwazeh, Joint channel equalization and detection of Spectrally Efficient {FDM} signals, in *2010 IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2010, pp. 177–182
12. R.G. Clegg, S. Isam, I. Kanaras, I. Darwazeh, A practical system for improved efficiency in frequency division multiplexed wireless networks. IET Commun. **6**(4), 449 (2012)
13. I. Darwazeh, T. Xu, T. Gui, Y. Bao, Z. Li, Optical SEFDM system; bandwidth saving using non-orthogonal sub-carriers. Photon. Technol. Lett. IEEE **26**(4), 352–355 (2014)
14. I. Darwazeh, T. Xu, T. Gui, Y. Bao, Z. Li, Optical spectrally efficient FDM system for electrical and optical bandwidth saving, in *2014 IEEE International Conference on Communications (ICC)*, 2014, pp. 3432–3437
15. D. Dasalukunte, F. Rusek, V. Öwall, Multicarrier faster-than-Nyquist transceivers: hardware architecture and performance analysis. IEEE Trans. Circ. Syst. I: Reg. Papers **58**, 1–12 (2010)
16. B. Farhang-Boroujeny, Filter bank multicarrier modulation : a waveform candidate for 5G and beyond. Adv. Electr. Eng. **2014**, 25 (2014)

17. B. Farhang-Boroujeny, R. Kempter, Multicarrier communication techniques for spectrum sensing and communication in cognitive radios. IEEE Commun. Mag. **46**(4), 80–85 (2008)
18. G.P. Fettweis, The tactile internet: applications and challenges. IEEE Veh. Technol. Mag. **9**(1), 64–70 (2014)
19. G. Fettweis, M. Krondorf, S. Bittner, GFDM - Generalized Frequency Division Multiplexing, in *IEEE 69th Vehicular Technology Conference, VTC*, Barcelona, 2009
20. FP 7 European Project 317669, METIS (Mobile and wireless communications Enablers for Twenty-twenty (2020) Information Society) (2012). https://www.metis2020.com/
21. D. Gabor, Theory of communication. J IEE: Part III: Radio Commun. Eng. **93**(26), 429–441 (1946)
22. A. Goldsmith, *Wireless Communications* (Cambridge University Press, Cambridge, 2005)
23. G.H. Golub, C.F.V.V. Loan, *Matrix Computations*, 3rd edn. (The Johns Hopkins University Press, Baltimore, MD, 1996)
24. R.C. Grammenos, Spectrum Optimisation in Wireless Communication Systems: Technology Evaluation, System Design and Practical Implementation. PhD thesis, University College London - Department of Electronic and Electrical Engineering, 2013
25. R.C. Grammenos, I. Darwazeh, Hardware implementation of a practical complexity spectrally efficient FDM reconfigurable receiver, in *IEEE 23rd International Symposium Personal, Indoor and Mobile Radio Communication, PIMRC*, Sydney, 2012
26. R.C. Grammenos, I. Darwazeh, Performance trade-offs and DSP evaluation of spectrally efficient FDM detection techniques, in *IEEE International Conference Communication, ICC*, Budapest (2013, to appear)
27. R.C. Grammenos, S. Isam, I. Darwazeh, FPGA design of a truncated SVD based receiver for the detection of SEFDM signals, in *IEEE 22nd International Symposium Personal, Indoor and Mobile Radio Communication, PIMRC*, Toronto, 2011
28. J. Hagenauer, The turbo principle: tutorial introduction and state of the art, in *Proceedings of the International Symposium Turbo Codes*, 1997, pp. 1–11
29. M. Hamamura, S. Tachikawa Bandwidth efficiency improvement for multi-carrier systems, in *IEEE 15th International Symposium Personal, Indoor and Mobile Radio Communication, PIMRC*, Barcelona, 2004, pp. 48–52
30. S. Hara, R. Prasad, Overview of multicarrier {CDMA}. Commun. Mag. IEEE **35**(12), 126–133 (1997)
31. B. Hassibi, H. Vikalo, On the sphere-decoding algorithm I. Expected complexity. IEEE Trans. Signal Process **53**(8), 2806–2818 (2005)
32. S.J. Heydari, M. Ferdosizadeh, F.A. Marvasti, Iterative detection with soft decision in spectrally efficient FDM systems. Tech. Rep. (2013). arXiv:1304.4003v1
33. B. Hirosaki, An orthogonally multiplexed QAM system using the discrete Fourier transform. IEEE Trans. Commun. **29**(7), 982–989 (1981)
34. E. Hopner, An experimental modulation-demodulation scheme for high-speed data transmission. IBM J. Res. Dev. **3**, 74–84 (1959)
35. S. Isam, I. Darwazeh, Precoded spectrally efficient FDM system, in *IEEE 21st International Symposium Personal Indoor and Mobile Radio Communication, PIMRC*, Istanbul, 2010, pp. 99–104
36. S. Isam, I. Darwazeh, Simple DSP-IDFT techniques for generating spectrally efficient FDM system signals, in *7th IEEE, IET International Symposium Communication Systems Networks and Digital Signal Processing, CSNDSP*, Newcastle, 2010
37. S. Isam, I. Darwazeh, Design and performance assessment of fixed complexity spectrally efficient FDM receivers, in *IEEE 73rd Vehicular Technology Conference, VTC*, Budapest, 2011
38. S. Isam, I. Darwazeh, Peak to average power ratio reduction in spectrally efficient FDM systems, in *IEEE International Conference Telecommunication, ICT*, Ayia Napa, 2011
39. S. Isam, I. Darwazeh, Characterizing the intercarrier interference of non-orthogonal spectrally efficient FDM system, in *8th IEEE, IET International Symposium Communication Systems Networks and Digital Signal Processing, CSNDSP*, Poznan, 2012

40. S. Isam, I. Darwazeh, Robust channel estimation for spectrally efficient FDM system, in *IEEE International Conference Telecommunication, ICT*, Jounieh, 2012
41. S. Isam, I. Kanaras, I. Darwazeh A truncated SVD approach for fixed complexity spectrally efficient FDM receivers, in *IEEE Wireless Communication and Networking Conference, WCNC*, Cancun, 2011
42. J. Jalden, B. Ottersten, On the complexity of sphere decoding in digital communications. IEEE Trans. Signal Process. **53**(4), 1474–1484 (2005)
43. J.P. Javaudin, D. Lacroix, Technical description of the OFDM/IOTA modulation. Tech. rep., France Telecom R&D, Tech. Rep. R1-03-168, 2003
44. W. Jian, Y. Xun, Z. Xi-lin, D. Li, The prefix design and performance analysis of DFT-based overlapped frequency division multiplexing (OvFDM-DFT) system, in *3rd International Workshop on Signal Design and Its Applications in Communication, IWSDA*, Chengdu, 2007, pp. 361–364
45. I. Kanaras, Spectrally Efficient Multicarrier Communication Systems: Signal Detection, Mathematical Modelling and Optimisation. PhD thesis, University College London (UCL), 2010
46. I. Kanaras, A. Chorti, M.R.D. Rodrigues, I. Darwazeh, A combined MMSE-ML detection for a spectrally efficient non orthogonal FDM signal, in *5th International Conference on Broadband Communications, Networks, and Systems, BROADNETS*, London, 2008
47. I. Kanaras, A. Chorti, M.R.D. Rodrigues, I. Darwazeh, An optimum detection for a spectrally efficient non-orthogonal FDM system, in *Proceedings of the 13th International OFDM-Workshop*, Hamburg, 2008
48. I. Kanaras, A. Chorti, M.R.D. Rodrigues, I. Darwazeh, A new quasi-optimal detection algorithm for a non orthogonal Spectrally Efficient FDM, in *9th International Symposium Communications and Information Technologies, ISCIT*, Incheon, 2009
49. I. Kanaras, A. Chorti, M.R.D. Rodrigues, I. Darwazeh, Investigation of a semidefinite programming detection for a spectrally efficient FDM system, in *IEEE 20th International Symposium Personal, Indoor and Mobile Radio Communication, PIMRC*, Tokyo, 2009
50. I. Kanaras, A. Chorti, M.R.D. Rodrigues, I. Darwazeh, Spectrally efficient FDM signals: bandwidth gain at the expense of receiver complexity, in *IEEE International Conference on Communications, ICC*, Dresden, 2009
51. I. Kanaras, A. Chorti, M.R.D. Rodrigues, A fast constrained sphere decoder for ill conditioned communication systems. IEEE Commun. Lett. **14**(11), 999–1001 (2010)
52. W. Kozek, A.F. Molisch, Nonorthogonal pulseshapes for multicarrier communications in doubly dispersive channels. IEEE J. Sel. Areas Commun. **16**(8), 1579–1589 (1998)
53. A.D. Liveris, C.N. Georghiades, Exploiting faster-than-Nyquist signaling. IEEE Trans. Commun. **51**(9), 1502–1511 (2003)
54. T. Matsumoto, T. Asai Frequency-overlapped multicarrier signaling with inter-carrier interference cancellation (FOLMSIC). IEICE SAT **97**(486), 69–76 (1998)
55. J.E. Mazo Faster-than-Nyquist signaling. Bell Syst. Tech. J. **54**(8), 1451–1462 (1975)
56. N. Michailow, M. Lentmaier Integration of a GFDM secondary system in an OFDM primary system, in *Future Network and Mobile Summit*, Warsaw, 2011
57. S. Mikroulis, T. Xu, J.E. Mitchell, I. Darwazeh, First demonstration of a spectrally efficient {FDM} radio over fiber system topology for beyond {4G} cellular networking, in *European Conference on Network and Optical Communications, NOC 2015*, 2015
58. M. Nekovee, Cognitive radio access to TV white spaces: spectrum opportunities, commercial applications and remaining technology challenges, in *IEEE International Symposium on New Frontiers Dynamic Spectrum Access Networks, DySPAN* , Singapore, 2010
59. H. Nikopour, H. Baligh, Sparse code multiple access, in *2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2013, pp. 332–336
60. D. Nopchinda, T. Xu, R. Maher, B. C. Thomsen, I. Darwazeh, Dual Polarization Coherent Optical Spectrally Efficient Frequency Division Multiplexing. IEEE Photon. Technol. Lett. **28**(1), pp. 83–86, (2016)

61. D. Panaitopol, R. Datta, G. Fettweis, Cyclostationary detection of cognitive radio systems using GFDM modulation, in *IEEE Wireless Communication and Networking Conference, WCNC*, Paris, 2012, pp. 930–934
62. M.R. Perrett, I. Darwazeh, Flexible hardware architecture of SEFDM transmitters with real-time non-orthogonal adjustment, in *IEEE International Conference on Telecommunications, ICT*, Ayia Napa, 2011
63. M.R. Perrett, R.C. Grammenos, I. Darwazeh, A verification methodology for the detection of spectrally efficient FDM signals generated using reconfigurable hardware, in *IEEE International Conference on Communications, ICC*, Ottawa, 2012
64. J. Proakis, *Digital Communications*, 4th edn. (McGraw-Hill, New York, 2000)
65. M.R.D. Rodrigues, I. Darwazeh, Fast OFDM: a proposal for doubling the data rate of OFDM schemes, in *IEEE International Conference on Telecommunication, ICT*, Beijing, 2002, pp. 484–487
66. M.R.D. Rodrigues, I. Darwazeh, A spectrally efficient frequency division multiplexing based communications system, in *Proceedings of the 8th International OFDM Workshop*, Hamburg, 2003, pp. 48–49
67. F. Rusek, J.B. Anderson, The two dimensional Mazo limit, in *Proceedings of the International Symposium Information Theory, ISIT*, Adelaide, 2005, pp. 970–974
68. A. Sahin, I. Guvenc, H. Arslan, A survey on multicarrier communications: prototype filters, lattice structures, and implementation aspects. *IEEE Commun Surv. Tutor.* **16**(3), 1312–1338 (2014)
69. Y. Saito, A. Benjebbour, Y. Kishiyama, T. Nakamura, System-level performance evaluation of downlink non-orthogonal multiple access {(NOMA)}, in *2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2013, pp. 611–615
70. F. Schaich, T. Wild, Waveform contenders for 5G - OFDM vs. FBMC vs. UFMC, in *6th International Symposium Communication, Control, and Signal Processing*, 2014, pp. 457–460
71. G. Strang, *Introduction to Linear Algebra*, 4th edn. (Wellesley Cambridge Press, Wellesley, 2009)
72. E. Viterbo, J. Boutros A universal lattice code decoder for fading channels. IEEE Trans. Inf. Theory **45**(5), 1639–1642 (1999)
73. R. Westwick, Mobile WiMAX versus LTE: a comparison of next generation mobile broadband technologies. J. Telecommun. Manage. **1**(1), 79–85 (2008)
74. P.N. Whatmough, M.R. Perrett, S. Isam, I. Darwazeh, VLSI architecture for a reconfigurable spectrally efficient FDM baseband transmitter. IEEE Trans. Circuits Syst. I: Reg. Papers **59**(5), 1–12 (2012)
75. P.N. Whatmough, M.R. Perrett, S. Isam, I. Darwazeh, {VLSI} architecture for a reconfigurable spectrally efficient {FDM} baseband transmitter. IEEE Trans. Circuits Syst. I: Reg. Papers **59**(5), 1107–1118 (2012)
76. G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. Brink, I. Gaspar, N. Michailow, A. Festag, L. Mendes, N. Cassiau, D. Ktenas, M. Dryjanski, S. Pietrzyk, B. Eged, P. Vago, F. Wiedmann, {5GNOW}: non-orthogonal, asynchronous waveforms for future mobile applications. Commun. Mag. IEEE **52**(2), 97–105 (2014)
77. F. Xiong, M-ary amplitude shift keying OFDM system. IEEE Trans. Commun. **51**(10), 1638–1642 (2003)
78. T. Xu, I. Darwazeh, A soft detector for spectrally efficient systems with non-orthogonal overlapped sub-carriers. Commun. Lett. IEEE **18**(10), 1847–1850 (2014)
79. T. Xu, I. Darwazeh, M-QAM signal detection for a non-orthogonal system using an improved fixed sphere decoder, in *9th IEEE/IET International Symposium on Communication Systems, Networks & Digital Signal Processing 2014 (CSNDSP14)*, Manchester, 2014, pp. 623–627
80. T. Xu, I. Darwazeh, Multi-band reduced complexity spectrally efficient FDM systems, in *9th IEEE/IET International Symposium on Communication Systems, Networks & Digital Signal Processing 2014 (CSNDSP14)*, Manchester, 2014, pp. 904–909
81. T. Xu, I. Darwazeh, Spectrally efficient FDM: spectrum saving technique for 5G?, in *2014 1st International Conference on 5G for Ubiquitous Connectivity (5GU)*, 2014, pp. 273–278

82. T. Xu, I. Darwazeh, Bandwidth compressed carrier aggregation, in *IEEE ICC 2015 - Workshop on 5G & Beyond - Enabling Technologies and Applications (ICC'15 - Workshops 23)*, London, 2015
83. T. Xu, R.C. Grammenos, I. Darwazeh, FPGA implementations of real-time detectors for a spectrally efficient FDM system, in *IEEE International Conference on Telecommunications, ICT*, Casablanca, 2013
84. T. Xu, R.C. Grammenos, F. Marvasti, I. Darwazeh, An improved fixed sphere decoder employing soft decision for the detection of non-orthogonal signals. IEEE Commun. Lett. **17**(10), 1964–1967 (2013)
85. Y.G. Yoo, J.H. Cho, Asymptotic optimality of binary faster-than-Nyquist signaling. IEEE Commun. Lett. **14**(9), 788–790 (2010)

# Full-Duplex Wireless Communications for 5G

**Mingxin Zhou, Yun Liao, and Lingyang Song**

**Abstract** In this chapter, we will introduce Full-Duplex (FD) wireless communications for 5G, which enables simultaneous transmission and reception over the same frequency band. In this way, the spectral efficiency can be improved significantly compared with half-duplex (HD). However, there exists severe self interference (SI), signal leakage from the local transmitter to its own receiver. Three different classes of SI mitigation techniques are presented in this chapter: propagation-domain SI suppression, analog-domain SI cancelation, and digital-domain SI cancelation. Furthermore, the system performance of several FD schemes in several different application scenarios is presented. Theoretically, the spectral efficiency of FD bidirectional and cooperative communications can be doubled, while for cognitive radio networks, the FD-based protocol can achieve much better sensing performance than the traditional HD-based cognitive radio schemes.

## 1 Introduction

The enormous increase in demand for data traffic drives the quest for efficient use of the available and valuable resources. However, most of the existing wireless communication devices can only operate in half-duplex, which dissipates the precious resources by exploiting either time-division or frequency-division. A full-duplex (FD) system simultaneously transmits and receives signals over the same frequency band [1], providing the potential of doubling the capacity compared to conventional half-duplex systems [2] and sustaining the evolution of the fifth generation (5G) technologies within limited spectrum. Beyond improvement of spectral efficiency, FD can also be beneficial to high layers, such as medium access (MAC) layer. From the perspective of MAC layer, enabling frame level FD transmissions, where a terminal is able to reliably receive an incoming frame while simultaneously transmitting an outgoing frame, can provide terminals with

M. Zhou • Y. Liao • L. Song (✉)
School of Electronics Engineering and Computer Science, Peking University,
Beijing 100871, China
e-mail: lingyang.song@gmail.com



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_11

new capabilities. For example, in FD cognitive radio networks, secondary users can sense and access the vacant spectrum simultaneously, leading to significant sensing performance and data transmission efficiency enhancement [2, 3].

Despite the attractive features of FD, it is greatly affected by the signal leakage from the local output to input, referred to as self-interference (SI). Because it is locally generated, the power of the SI is much stronger than that of the desired signal. Apart from the direct SI, reflected interference signals from nearby obstacles also exist. Recent years, there has been a large number of SI cancelation techniques [2] making FD communications more and more feasible, which will be introduce in the next subsection.

Overall, FD is a promising technology that can be employed in the next-generation wireless networks. There are two reasons as follows: First, there is a widespread consensus that most conventional approaches to increasing spectral efficiency, such as advances in modulation, coding and MIMO techniques, have now been exhausted, and the FD system tends to be a non-conventional approach that can further increase the spectral efficiency. In addition, Indoor dense networks to be deployed in 5G, such as small-cell systems, are making the SI cancelation problem more manageable than previous long-coverage macro-cell networks. This is because in small-cell systems, the transmit power is smaller, and the path loss is reduced due to its short coverage.

## *1.1 Self-interference Cancelation Techniques*

First, an example of small-cell cellular systems is provided to demonstrate the impact of SI [2]. The maximum allowable transmit powers of the base stations and mobile handsets are 24 and 21 dBm in small-cell systems, respectively, and the noise floors are both $-100$ dBm. If we assume 15 dB isolation between the base station's transmit and receive signal paths, then the SI of the base station is $24 - 15 - (-100) = 109$ dB above the noise floor. This means a FD base station must cancel self interference by about 109 dB to achieve the link SINR equivalent to that of a half duplex system.

Why is self-interference hard to subtract? After all, the FD transceiver knows the signal being transmitted, it can just subtract it from the received signal. However, this is not the whole picture. Although the transceiver knows the original version of the clean transmitted digital baseband signal, once it is converted to analog and up-converted to the carrier frequency and then transmitted through a fading channel, the received version is quite different from the original one. Actually, after all these processes, the received version is a complicated distorted function of the original clean signal along with noise.

Thus, sophisticated approaches to SI cancelation is needed to obtain the desired signal. In the following, these approaches are described, which can be typically classified into three types, namely propagation-domain, analog-circuit-domain, and digital-domain approaches [2].

### 1.1.1 Propagation-Domain Self-interference Suppression

Wireless-propagation-domain isolation techniques aim to electromagnetically isolate the transmit chain from the receive chain, i.e., to suppress the self-interference before it manifests in the receive chain circuitry, so that the downstream receiver hardware does not need to accurately process signals with a huge dynamic range.

In separate-antenna systems, the path loss between the FD terminals' transmit and receive antennas (or antenna arrays) can be increased by spacing them apart and/or by placing absorptive shielding between them, as quantified in [4–6]. Although the simplicity makes the pass-loss-based approaches attractive, their effectiveness is greatly limited by the size of device: the smaller the device, the less room there is to implement such techniques. Cross-polarization technique offers an additional mechanism to electromagnetically isolate the transmit and receive antennas. For example, an FD terminal can be sophistically designed so that it transmits only horizontally polarized signals and receives only vertically polarized signals, and the interference between them can be avoided [5, 7]. Similarly, with directional transmit and/or receive antennas (i.e., antennas with non-uniform radiation/sensing patterns), one may align their null directions to achieve the same goal [8]. In fact, by carefully placing a single receive antenna at precisely a location where the carrier waveforms are exactly 180° out of phase, one can near-perfectly null the received signal at the receive antenna, and the self-interference can be near-perfectly canceled [1, 7].

### 1.1.2 Analog-Circuit-Domain Self-interference Cancellation

Analog-circuit-domain cancellation focuses on cancelation of SI in the analog receive-chain circuitry before the ADC. The basic idea is to model and predict the distortions caused by SI and compensate the received signal accordingly in the analog domain. This cancelation may occur either before or after the downconverter and the (low-noise amplifier) LNA. In these techniques, a signal that resembles the SI at the receive-chain is generated, and utilized to cancel SI by signal subtraction. The generated signal can either be tapped at the transmit antenna, or obtained from the digital domain, and apply the necessary gain/phase/delay adjustments digitally (where it is much easier to do so), and then convert it to the analog-circuit domain for use in self-interference cancellation [9, 10].

For the cancelation schemes before the downconverter, i.e., at RF, the canceling signal also needs to be upconverted to RF. The analog cancelers can be classified based on whether the canceling signal is generated by processing the self-interference signal prior to or after upconversion. Those cancelers where the canceling signal is generated by processing prior to upconversion are called pre-mixer cancelers, whereas cancelers where the canceling signal is generated by processing afterwards are called post-mixer cancelers [11]. Figure 1a shows the structure of a pre-mixer canceler with processing function $f(\cdot)$, in which $x[n]$ is the

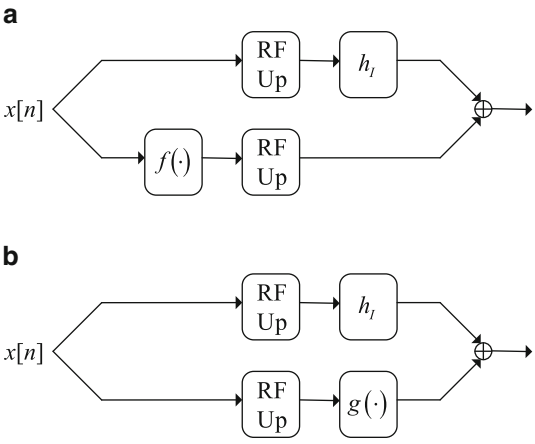


**Fig. 1** Schematics of the pre-mixer (**a**) and post-mixer (**b**) analog cancelers

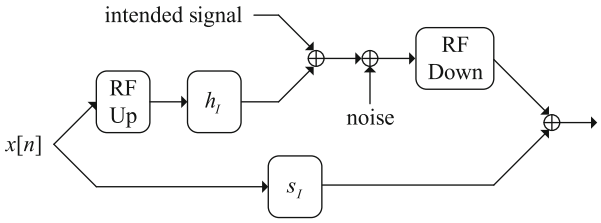


**Fig. 2** Structure of a baseband analog canceler

original signal, and $h_I$ is the self-interference channel from the transmit-chain to the receive-chain. Figure 1b is the schematic of a post-mixer canceler with processing function $g(\cdot)$. Functions $f(\cdot)$ and $g(\cdot)$ are ideal if they can completely eliminate self-interference from the received signal.

An analog canceler where the canceling signal is generated in baseband and the cancelation occurs in the analog baseband is called a baseband analog canceler. Figure 2 shows a representation of the baseband analog canceler. In baseband analog cancelers, transmitted signal $x[n]$ is processed by function $s(\cdot)$, and added directly to the received signal to perform the cancelation.

### 1.1.3 Digital-Domain Self-interference Cancellation

Digital-domain SI cancelation works in the digital domain after the received signal has been quantized by the ADC by applying sophisticated DSP techniques to the received signal [2]. The advantage of using digital-domain SI cancellation techniques is the reduction of circuit complexity and power consumption. However, the dynamic range of the ADC imposes strict limit on the maximum amount of SI suppression, which means that to implement digital-domain methods, a sufficient amount of the SI suppression must be done before the ADC using other

techniques like the propagation-domain and/or analog-circuit-domain methods described above. In this sense, the digital-domain cancellation can be regarded as the last step of defense against self-interference, where the goal is to cancel the self-interference left over from the propagation-domain and analog-circuit-domain approaches.

## *1.2 Key Application Scenarios*

FD technology, with its key feature of enabling simultaneous transmission and reception and its potential of doubling the spectral efficiency, has ignited great interest in both academia and industry. We introduce the following three key application scenarios: FD bidirectional communications, FD cooperative communications, FD cognitive radio networks.

Figure 3 depicts FD bidirectional communication between a pair of FD transceivers. The two FD nodes transmit and receive data to and from each other over the same frequency band simultaneously. The two nodes $A$ and $B$ are equipped with $N_A$ and $N_B$ antennas, respectively. Let $link_{AB}$ denote the transmission link from node $A$ to $B$, and $link_{BA}$ to denote the link from $B$ to $A$. Each node receives a combination of its desired signal from the other node and SI from its own transmit antennas. In this scenario, if the SI can be reduced completely, then the spectral efficiency will be doubled by using FD rather than half duplex (HD) communications.

The next application of FD to wireless communication came in the context of cooperative communications illustrated in Fig. 4, which generally consist of three types of nodes, a source node, a relay node, and a destination node. The source

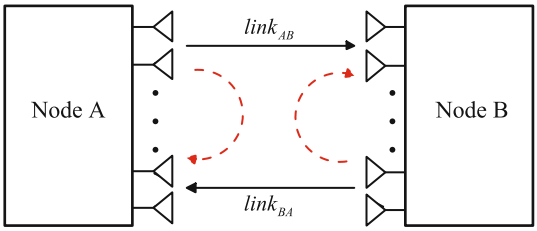


**Fig. 3** Full-duplex bidirectional communications

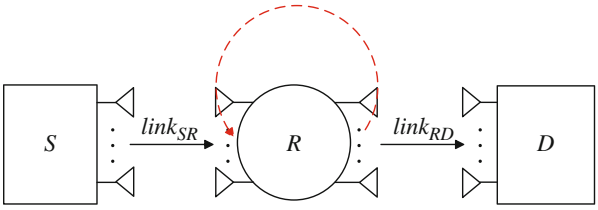


**Fig. 4** Full-duplex cooperative communications

node wants to transmit its data signal to the destination node. The relay node is used for increase the coverage of the source node, i.e., it receives, amplifies and retransmits the wireless signal. Let us denote the transmission links from S to R as $link_{SR}$ and R to D as $link_{RD}$, respectively. When FD transmission is applied to the relay node, however, the source and relay nodes can simultaneously transmit their signals via $link_{SR}$ and $link_{RD}$ in the same frequency band. Similar to the FD bidirectional scenario, the FD cooperative communications can theoretically double the spectral efficiency.

Finally, the FD cognitive radio networks (CRNs) is considered. In CRNs, the secondary (unlicensed) users (SUs) are allowed to share the licensed spectrum with the primary (licensed) users (PUs) as long as SUs do not cause harmful interference at the PUs. Spectrum sensing is essential in identifying spectrum holes that can be utilized by SUs. Conventionally, SUs need to periodically suspend their ongoing transmissions to perform spectrum sensing, which leads to spectrum waste and tardy reaction to the PU's state change. When SUs have FD capability, they can perform simultaneous sensing and transmission, as shown in Fig. 5, and thus to bypass the intrinsic problems in conventional HD CRNs.

Apart from the aforementioned application scenarios, full-duplex wireless can be applied to more communication scenarios, such as cellular networks, ad-hoc networks, local area networks, satellite communications, vehicle communications, etc. Taking FD cellular network as an example, the base station can operate in FD, then it can support simultaneous uplink and downlink communications, potentially doubling the spectral efficiency.

The rest of the chapter is organised as follows: Section 2 introduces FD bidirectional communications, including the FD/HD mode switching and transmit-receive antenna pair selection schemes. In Sect. 3, the FD cooperative communications are described. The one-way joint relay and antenna mode selection and two-way

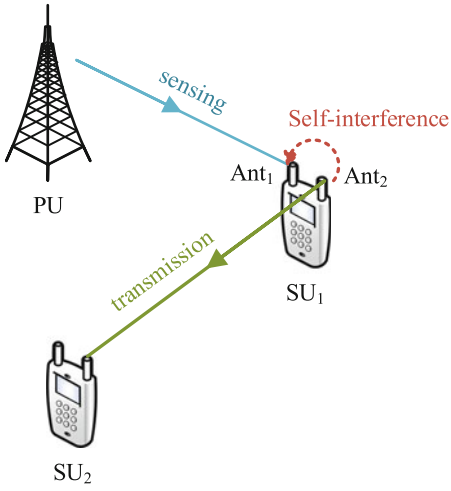


**Fig. 5** Full-duplex cognitive radio network

relay selection schemes are briefly introduced in this section, and the performance analysis of the schemes is provided. Finally, the "listen-and-talk" protocol of FD CRNs is presented in the last section.

## 2 Full-Duplex Bidirectional Communications

Thanks to the various state-of-the-art approaches for self-IC schemes, SI is no longer a critical bottleneck to implementing a practical FD system. In this section, a few signal processing techniques of FD bidirectional communications are elaborated, and the analysis of their performance is given.

### *2.1 Mode Switching Between Full-Duplex and Half-Duplex*

Typically, there are two kinds of baseline for FD and HD mode switching: fixed number of antennas, and fixed number of RFs. When the number of antennas is fixed, the system can be adaptively switched between FD and half-duplex spatial-multiplexing (HD-SM), where both modes have the same degree of freedom. On the other hand, when the number of transmit and receive RFs is fixed, the number of active RFs in FD can be twice of that in HD,[1] thus the degree of freedom of FD is potentially doubled. In this section, only the first kind is considered, and details of the second kind can be found in [12].

The system model is illustrated in Fig. 6, where each node is equipped with two antennas. In FD, one antenna is used for transmission and the other is used for reception at each node, while, in HD-SM, both antennas at one node are used for transmission and those at the other node are used for reception. In FD, the SI

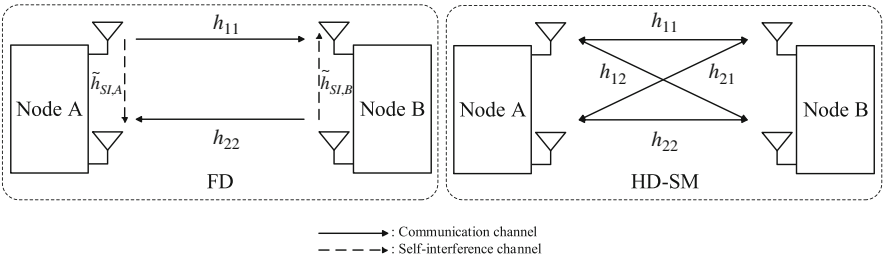


**Fig. 6** Mode switching between FD and HD-SM

[1] In FD, all of the RFs are active, while only transmit RFs at the transmitter and receive RFs at the receiver in HD are active.

can be partially cancelled by the aforementioned SI cancelation techniques, and the residual interference can generally be modeled to follow Rayleigh distribution. On the other hand, the HD-SM system can use all of the communication links and avoid the self-interference problem, however, it may suffer from the sum-rate loss caused by the spatial correlation. As depicted in Fig. 6, there are overall 6 links, containing 4 communication links and 2 self-interference links. The communication links between the two nodes can be denoted as $\mathbf{H} = \left[h_{ij}\right]$, $i,j = 1,2$, the entry of which follows independent identically distributed (i.i.d.) Rayleigh distribution with zero mean and unit variance, while the residual self-interference channel is $\tilde{h}_{SIk}$, $k = A, B$, with zero mean and unit variance.

In FD, the two nodes employ 2 communication links to exchange information. After self interference cancellation, both nodes receive a combination of the signal transmitted by the other node, the residual self interference, and noise

$$\begin{aligned} y_A &= \sqrt{\rho} h_{11} x_B + \sqrt{\eta} \tilde{h}_{SI,A} x_A + n_A \\ y_B &= \sqrt{\rho} h_{22} x_A + \sqrt{\eta} \tilde{h}_{SI,B} x_B + n_B, \end{aligned} \tag{1}$$

where $\rho$ and $\eta$ denote the average SNR and INR, respectively. $x_A$, $x_B$ denote the signal transmitted from nodes $A$ and $B$, and $n_A, n_B$ denote the AWGN with zero mean and unit variance at the corresponding nodes.

At the same time, when the HD-SM system is employed, assuming time division duplex in this paper, each node processes one half of the time resources for transmission through a $2 \times 2$ channel $\mathbf{H}$. We first consider received signal at node B, which is given by

$$\mathbf{y}_{B,SM} = \sqrt{\rho} \mathbf{H} \mathbf{x}_A + \mathbf{n}_B, \tag{2}$$

where $\mathbf{n}_B$ is an AWGN vector with unit variance. $\mathbf{x}_A$ is the signal vector transmitted from node $A$.

The HD-SM system can avoid the self-interference problem, however, it suffers from the sum-rate loss caused by the spatial correlation. Under the separable correlation model [13], it is possible to express the channel matrix as $\mathbf{H} = \mathbf{\Phi}_R^{1/2} \mathbf{H}_0 \mathbf{\Phi}_T^{1/2}$, where $\mathbf{H}_0$ is an i.i.d. zero mean complex Gaussian random matrix with unit variance, and $\mathbf{\Phi}_R$ and $\mathbf{\Phi}_T$ represent the normalized correlation matrixes at receiver and transmitter, respectively. Assuming a block-fading reciprocal channel, where channel varies independently at different time slots, and the forward channel is reciprocal with the reverse channel. Because the two nodes process one half of time resource, transmit/receive correlation matrices at a node are receive/transmit correlation matrices at the other node. Thus, the received signal at node $A$ is symmetric, and the average achievable rates for the two nodes are equal. For simplicity, in the following sections, only node $B$ is considered as the receiver.

### 2.1.1 Average Sum Rates

Due to symmetry, the average rates of the two directions are equal, thus the average sum rate is a doubled rate of one direction

$$\bar{R}_{FD} = 2\mathrm{E}\left[\log_2\left(1 + \frac{\rho|h_{22}|^2}{\eta\left|\tilde{h}_{SI,A}\right|^2 + 1}\right)\right], \tag{3}$$

After some manipulations, the closed form expression can be expressed as [14]

$$\bar{R}_{FD}\frac{2\exp\left(\frac{1}{\rho}\right)\log_2 e}{\left(\frac{\eta}{\rho} - 1\right)}\left[\exp\left(\frac{1}{\eta} - \frac{1}{\rho}\right)\mathrm{E}_1\left(\frac{1}{\eta}\right) - \mathrm{E}_1\left(\frac{1}{\rho}\right)\right]. \tag{4}$$

where $\mathrm{E}_1(\cdot)$ is the exponential integral function of the first order.

Next, the performance of HD-SM is given. The case of HD-SM with power constraint $\rho$ and equal power allocation at the transmitter is considered. The eigenvalues of the transmit and receive correlation matrixes, $\mathbf{\Phi}_R$ and $\mathbf{\Phi}_T$ are represented as $\phi_{r,1} > \phi_{r,2}$ and $\phi_{t,1} > \phi_{t,2}$, respectively. Therefore, the sum rate of the correlated HD-SM system can be represented as [13]

$$\begin{aligned}\bar{R}_{SM} &= \mathrm{E}\left\{\log_2 \det\left(\mathbf{I} + \frac{\rho}{2}\mathbf{H}_0^H\mathbf{\Phi}_R^H\mathbf{H}_0\mathbf{\Phi}_T\right)\right\} \\ &= \frac{2\sum_{l=1}^{2}\det\left(\mathbf{\Xi}(l)\right)}{\ln(2)\rho(\phi_{t,2} - \phi_{t,1})(\phi_{r,2} - \phi_{r,1})},\end{aligned} \tag{5}$$

where $\Xi(l)$, $l = 1, 2$, are $2 \times 2$ matrixes with entries given by

$$\{\Xi(l)\}_{i,j} = \begin{cases} \frac{\rho}{2}\phi_{t,i}\phi_{r,j}\exp\left(\frac{2}{\rho\phi_{t,i}\phi_{r,j}}\right)\mathrm{E}_1\left(\frac{2}{\rho\phi_{t,i}\phi_{r,j}}\right), & i = l \\ \frac{\rho}{2}\phi_{t,i}\phi_{r,j} + 1, & i \neq l \end{cases}. \tag{6}$$

As a result, combining the exact sum rates of the FD mode (4) and the HD-SM mode (5), we can define

$$\begin{aligned}\Delta R^{EX} &\triangleq \bar{R}_{FD} - \bar{R}_{SM}. \\ &= \frac{2\exp\left(\frac{1}{\rho}\right)\log_2 e}{\left(\frac{\eta}{\rho} - 1\right)}\left[\exp\left(\frac{1}{\eta} - \frac{1}{\rho}\right)\mathrm{E}_1\left(\frac{1}{\eta}\right) - \mathrm{E}_1\left(\frac{1}{\rho}\right)\right] \\ &\quad - \frac{2\sum_{l=1}^{2}\det\left(\mathbf{\Xi}(l)\right)}{\ln(2)\rho(\phi_{t,2} - \phi_{t,1})(\phi_{r,2} - \phi_{r,1})}.\end{aligned} \tag{7}$$

The system can adaptively select between FD and HD-SM modes depending on $\Delta R^{EX}$. That is, the system selects HD-SM, if $\Delta R^{EX} < 0$; otherwise, it selects FD. The criterion is determined by three statistical factors: the average INR $\eta$ in FD, the spatial correlation $\mathbf{\Phi}_T$, $\mathbf{\Phi}_R$ in HD-SM, and the average SNR $\rho$. However, it is rather complicated to determine the exact criterion through these three factors. In the following, the approximate results are provided to facilitate the mode switching.

### 2.1.2 Approximate Results

First the approximate average sum rate of the FD system is calculated at high SNR. Using the approximation $\mathrm{E}_1(\varepsilon) \approx -\gamma - \ln(\varepsilon)$ and $\exp(\varepsilon) \approx 1 + \varepsilon$ for $\varepsilon = \frac{1}{\rho} \to 0$ [15], the approximate sum rate of the FD system in (4) can be rewritten as

$$\begin{aligned} \bar{R}_{FD}^{AP} =& 2\frac{\rho+1}{\rho-\eta}\log_2\rho - 2\gamma\frac{\rho+1}{\rho-\eta}\log_2 e \\ & - \frac{2\left(\rho^2-1\right)\log_2 e}{(\rho-\eta)\,\rho}\exp\left(\frac{1}{\eta}\right)\mathrm{E}_1\left(\frac{1}{\eta}\right), \end{aligned} \tag{8}$$

where $\gamma \approx 0.5772$ is Euler's constant [15]. Furthermore, at low or moderate INR, i.e., $\rho \gg \eta$, (8) can be well approximated by

$$\bar{R}_{FD}^{AP} = [2\log_2\rho - 2\gamma\log_2 e] - \exp\left(\frac{1}{\eta}\right)\mathrm{E}_1\left(\frac{1}{\eta}\right)\log_2 e. \tag{9}$$

The first term is the high-SNR sum rate of the FD system without residual self interference, and the second term denotes the rate degradation caused by the residual self interference.

On the other hand, at high SNR, the average sum rate of the HD-SM system can be well approximated by Paulraj et al. [16]

$$\begin{aligned} \bar{R}_{SM}^{AP} &= \mathrm{E}\left\{\log_2\det\left(\frac{\rho}{2}\mathbf{H}_0\mathbf{H}_0^H\right)\right\} + \log_2\det\left(\mathbf{\Phi}_T\right) \\ &\qquad + \log_2\det\left(\mathbf{\Phi}_R\right) \\ &\approx \left[2\log_2(\frac{\rho}{2}) + (1-2\gamma)\log_2 e\right] + \log_2\det\left(\mathbf{\Phi}_T\right) \\ &\qquad + \log_2\det\left(\mathbf{\Phi}_R\right). \end{aligned} \tag{10}$$

The first term is the high-SNR sum rate for independent fading HD-SM. The second and third terms represent the rate degradation due to the spatial correlations at transmit and receive sides, respectively. It is worth noting that $\det\left(\mathbf{\Phi}_T\right)$ and $\det\left(\mathbf{\Phi}_R\right)$ are less than one, and a larger correlation $\mathbf{\Phi}$ implies a smaller $\det(\mathbf{\Phi})$.

Similarly, an approximate selection criterion can be derived by calculating the difference between the approximate average sum rates in (10) and (23)

$$\begin{aligned}\Delta R^{AP} &\triangleq \bar{R}_{FD}^{AP} - \bar{R}_{SM}^{AP} \\ &= [2 - \log_2 e] - \log_2 \left[\det\left(\boldsymbol{\Phi}_T\right) \det\left(\boldsymbol{\Phi}_R\right)\right] \\ &\quad - 2\exp\left(\frac{1}{\eta}\right) \mathrm{E}_1\left(\frac{1}{\eta}\right) \log_2 e,\end{aligned} \tag{11}$$

Given the high-SNR assumption, $\Delta R^{AP}$ is just determined by two factors: monotonically increasing with spatial correlation in the HD-SM system, and monotonically decreasing with the average INR $\eta$, but independent of SNR $\rho$.

Let $\Delta R^{AP} = 0$, and the approximate selection threshold can be calculated. We consider the exponential correlation model [13] with equal correlation coefficients at both sides

$$\boldsymbol{\Phi}_T = \boldsymbol{\Phi}_R = \begin{bmatrix} 1 & \beta^* \\ \beta & 1 \end{bmatrix}, \tag{12}$$

Then, the selection threshold of spatial correlation $\beta$th can be given by

$$|\beta \mathrm{th}| = \begin{cases} \sqrt{1 - \frac{2}{\exp\left[0.5 + \exp\left(\frac{1}{\eta}\right)\mathrm{E}_1\left(\frac{1}{\eta}\right)\right]}}, & \eta > \eta_0 \\ 0, & \eta \le \eta_0 \end{cases}. \tag{13}$$

Given $\eta \ge \eta_0$, the system selects HD-SM, if $|\beta| > |\beta \mathrm{th}|$; otherwise, it selects FD. If $\eta \le \eta_0$, $\Delta R_{AP}$ is always positive, i.e., FD is always superior to HD-SM even if the MIMO channel is independent. $\eta_0$ is the value that satisfies

$$\begin{aligned}1 - \frac{2}{\exp\left[0.5 + \exp\left(\frac{1}{\eta_0}\right)\mathrm{E}_1\left(\frac{1}{\eta_0}\right)\right]} &= 0 \\ \exp\left(\frac{1}{\eta_0}\right)\mathrm{E}_1\left(\frac{1}{\eta_0}\right) &= \ln 2 - 0.5\end{aligned} \tag{14}$$

By using $\frac{1}{2}\ln\left(1 + \frac{2}{x}\right) < \exp(x)\mathrm{E}_1(x) < \ln\left(1 + \frac{1}{x}\right)$ [15] for $x = \frac{1}{\eta_0}$, we can obtain

$$\frac{2}{\sqrt{e}} - 1 < \eta_0 < \frac{2}{e} - \frac{1}{2}. \tag{15}$$

Or equivalently, we have $-6.71\,\mathrm{dB} < \eta_0 < -6.28\,\mathrm{dB}$.

The derivative of $|\beta\text{th}|$ can be written as

$$\frac{d|\beta\text{th}|}{d\eta} = \frac{d|\beta\text{th}|}{dt}\frac{dt}{d\eta}, \tag{16}$$

where $t = \exp\left(\frac{1}{\eta}\right)\mathrm{E}_1\left(\frac{1}{\eta}\right)$. It is obvious that $\frac{d|\beta\text{th}|}{dt} > 0$, and $\frac{dt}{d\eta}$ can be given by

$$\begin{aligned}\frac{dt}{d\eta} &= \frac{1}{\eta} - \frac{1}{\eta^2}\exp\left(\frac{1}{\eta}\right)\mathrm{E}_1\left(\frac{1}{\eta}\right) \\ &> \frac{1}{\eta} - \frac{1}{\eta}\frac{\eta+1}{2\eta+1} \\ &= \frac{1}{2\eta+1} > 0\end{aligned} \tag{17}$$

where the first inequality is obtained by using $x\exp(x)\mathrm{E}_1(x) < \frac{x+1}{x+2}$ [15] for $x = \frac{1}{\eta}$. Thus, we can obtain $\frac{d|\beta\text{th}|}{d\eta} > 0$, which implies that the spatial correlation threshold in (13) monotonically increases with self interference.

Figure 7 depicts the exact and approximate switching thresholds between FD and HD-SM, where we assume $\beta_r = \beta_t = \beta$ without loss of generality. The exact threshold is obtained by comparing the exact sum rates of the FD and HD-SM systems in (4) and (5). The approximate threshold is given in (13). Above the threshold is the area where the FD system outperforms the HD-SM system, whereas the HD-SM system achieves a better performance in the region below the curve.

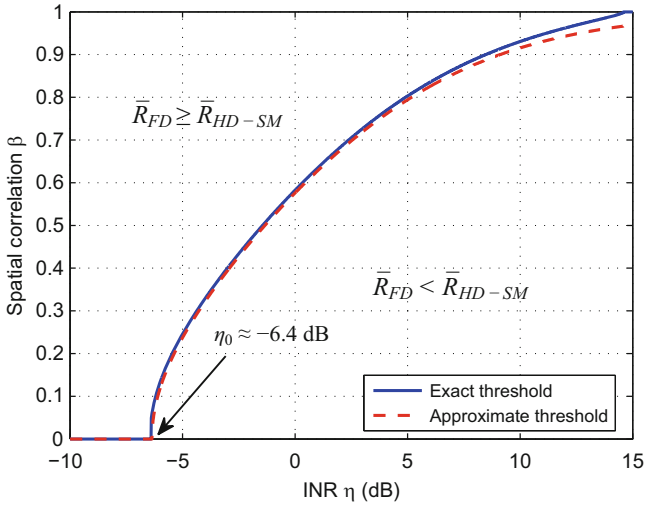


**Fig. 7** Exact and approximate switching thresholds, with SNR = 30 dB

Moreover, through simulation, we can obtain $\eta_0 \approx -6.4\,\text{dB}$, which accords with the aforementioned analysis $-6.71\,\text{dB} < \eta_0 < -6.28\,\text{dB}$. Through this figure, it can be shown that at high SNR, the approximate threshold is effective to determine whether the FD system or the HD-SM system should be employed in a specific condition.

## 2.2 Transmit-Receive Antenna Pair Selection

In this subsection, a transmit-receive antenna pair selection (TRAPS) strategy [17, 18] is elaborated, where the functions of the antennas are not predefined, and the system can adaptively configure the antennas for transmission and reception based on the given criteria. In the following, the maximum sum rate and minimum symbol error rate (SER) criteria are introduced, and their performance is analyzed.

As shown in Fig. 8, a bidirectional communication scenario with two nodes, $N_1$ and $N_2$, is considered , where each node is equipped with two antennas, which can be dynamically configured for transmission and reception. There exist four kinds of possible combinations (patterns), as depicted in Fig. 8. The channels between the two nodes exhibit independent Rayleigh fading. Let $h_{i\bar{i}}^{(j)}$ denote the channel of pattern $(j)$ from node $N_{\bar{i}}$ to $N_i$, $i = 1, 2$, $j = 1, 2, 3, 4$, $\bar{i} = 3 - i$, in Fig. 8.

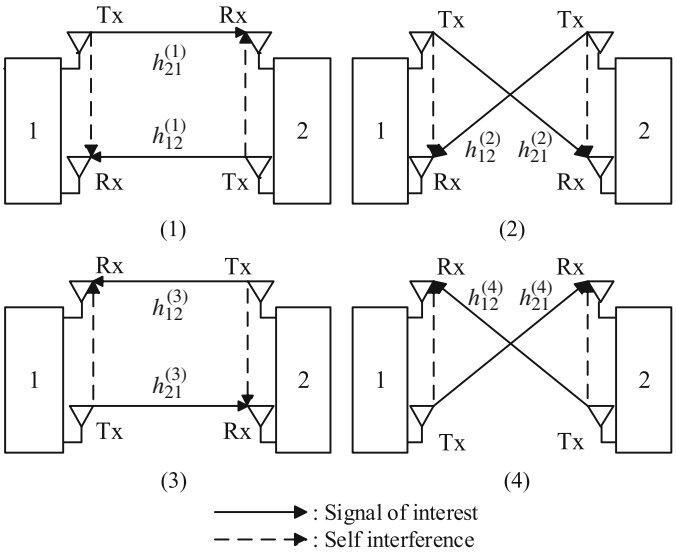


**Fig. 8** System model of TRAPS scheme

We assume that they are all complex-Gaussian random variables with zero mean and unit variance, i.e., $h_{\bar{i}i}^{(j)} \sim \mathcal{CN}(0,1)$. The transmit powers of the two nodes are equal, denoted by $P_s$.

The received signal at node $N_i$ can be expressed as

$$y_i^{(j)} = \sqrt{P_s} h_{\bar{i}i}^{(j)} x_{\bar{i}}^{(j)} + \sqrt{\eta P_s} w_i^{(j)} + n_i^{(j)}, \tag{18}$$

where $x_{\bar{i}}^{(j)}$ denotes the transmitted signal from node $\bar{i}$ in pattern $j$. The second term represents the residual interference after cancelation. $w_i$ follows $\mathcal{CN}(0,1)$, and $\eta$ denotes the interference cancelation factor, indicating the cancelation capability. A small $\eta$ indicates a good performance of interference cancelation. $n_i$ denotes the additive white Gaussian noise (AWGN) at node $i$, with zero mean and unit variance. Therefore, the instantaneous received signal-to-interference-plus-noise ratio (SINR) of node $N_i$ in pattern $j$ can be expressed as

$$\rho_i^{(j)} = \frac{P_s}{\eta P_s + 1} |h_{\bar{i}i}^{(j)}|^2. \tag{19}$$

The probability density function (PDF) of $\rho_i^{(j)}$ is given by

$$f_{\rho_i^{(j)}}(x) = \frac{\eta P_s + 1}{P_s} e^{-\frac{\eta P_s + 1}{P_s} x}. \tag{20}$$

Furthermore, we have $|h_{\bar{i}i}^{(1)}|^2 = |h_{\bar{i}i}^{(3)}|^2$, $|h_{\bar{i}i}^{(2)}|^2 = |h_{\bar{i}i}^{(4)}|^2$. Thus, pattern (1), (3) and pattern (2), (4) have the same instantaneous SINR at each node, respectively. For simplicity, only patterns (1) and (2) are considered for the rest of this subsection.

### 2.2.1 Maximum Sum Rate Criterion

In the Max-SR approach, where the pattern is selected with a higher sum rate of the two nodes

$$\mathcal{P} = \arg\max_j \left( \sum_{i=1}^{2} R_i^{(j)} \right), \tag{21}$$

where $R_i^{(j)} = \log_2 \left( 1 + \frac{P_s}{\eta P_s + 1} |h_{\bar{i}i}^{(j)}|^2 \right)$ represents the rate of node $N_i$ in pattern $(j)$. Thus, the sum rate of the Max-SR approach can be given by

$$\bar{R}_{max} = \mathbb{E} \left[ \max_j \left( \sum_{i=1}^{2} R_i^{(j)} \right) \right], \tag{22}$$

By using $\max(A+B, C+D) \leq \max(A,C) + \max(B,D)$, for $A, B, C, D \geq 0$,[2] An analytical upper bound is derived as

$$\bar{R}_{ub} = \mathbb{E}\left[\sum_{i=1}^{2} \max_{j}\left(R_i^{(j)}\right)\right]. \tag{23}$$

After some manipulations, (23) can be rewritten as

$$\bar{R}_{ub} = 2\left[2e^{\frac{\eta P_s+1}{P_s}}\mathrm{E}_1\left(\frac{\eta P_s+1}{P_s}\right) - e^{\frac{2\eta P_s+2}{P_s}}\mathrm{E}_1\left(\frac{2\eta P_s+2}{P_s}\right)\right]\log_2 e. \tag{24}$$

In the high-SNR regime, the average sum rate is obtained as

$$\lim_{P_s\to\infty} \bar{R}_{ub} = 2e^{\eta}\mathrm{E}_1(\eta)\log_2 e + 2\left[e^{\eta}\mathrm{E}_1(\eta) - e^{2\eta}\mathrm{E}_1(2\eta)\right]\log_2 e. \tag{25}$$

The first term of (25) is the sum rate of traditional FD in the high-SNR regime, while the second term, which is easily proved to be positive, represents the improvement of the sum rate ceiling by using Max-SR approach.

We define $\Delta R = e^{\eta}\mathrm{E}_1(\eta) - e^{2\eta}\mathrm{E}_1(2\eta)$, and we have $d\Delta R/d\eta < 0$. It indicates that when the residual self interference gets smaller, the improvement of the sum rate increases. Furthermore, when the self interference is canceled properly, i.e., $\eta \to 0$, an upper bound of $\Delta R$ can be obtained, by using $\mathrm{E}_1(\eta) \approx -\gamma - \ln\eta$, $e^{\eta} \approx 1+\eta$

$$\Delta R \leq \Delta R_{\eta\to 0} \approx (1+\eta)\ln 2 + \eta(\gamma + \ln 2\eta) \approx \ln 2. \tag{26}$$

Substituting (26) into (25), it can be shown that the improvement of the sum rate increases with the improvement of the interference cancelation ability, i.e., the decrease of $\eta$, and it is upper bounded by 2.

Figure 9 plots the sum rate versus the transmit power $P_s$ for the FD with or without TRAPS. The exact sum rate curves are given by Monte Carlo simulations with the interference cancelation factor $\eta = \{-3, -10, -20\,\mathrm{dB}\}$. The upper bound of the Max-SR is depicted by (24). Figure 9 shows that the Max-SR achieves a sum rate gain, and converges to a higher sum rate ceiling than that of the traditional FD. In addition, with the improvement of interference cancelation, i.e., the decrease of $\eta$, the sum rate gain increases correspondingly.

[2]This inequality can be proved by classified discussion on the different relationships of A, B, C and D.

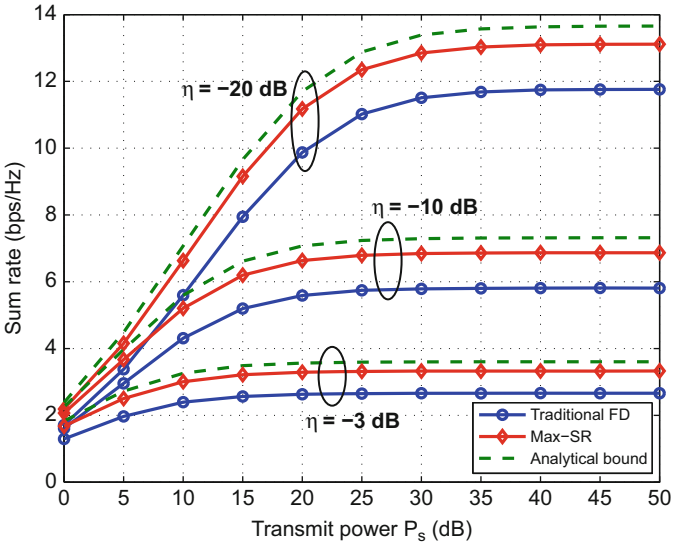


**Fig. 9** Analytical and simulated sum rate versus transmit power $P_s$ of Max-SR and traditional FD, with $\eta = \{-3, -10, -20\,\text{dB}\}$

### 2.2.2 Minimum SER Antenna Criterion

In the Min-SER criterion, the pattern is selected with a lower sum SER of the two nodes

$$\mathcal{P} = \arg\min_{j} \left( \frac{1}{2} \sum_{i=1}^{2} \text{SER}_i^{(j)} \right), \tag{27}$$

where $\text{SER}_i^{(j)}$ denotes the SER at $N_i$ from the other node in pattern $(j)$. The average SER of the Min-SER approach is conditioned on the instantaneous SINR and can be written as

$$\overline{\text{SER}}_{min} = \frac{1}{2}\text{E}\left[ \min_{j} \left( \sum_{i=1}^{2} Q\left( \sqrt{b\rho_i^{(j)}} \right) \right) \right], \tag{28}$$

where $Q\{\cdot\}$ is the Gaussian-$Q$ function, and $b$ is a constant determined by the modulation format, e.g., $b = 2$ for binary phase-shift keying (BPSK).

In most cases, the sum SER of the two nodes is dominated by the larger one, and thus, a simplified selection criterion can be considered for which the larger instantaneous SER is minimized so as to minimize the average sum SER in a

suboptimal way, referred to as Min-Max selection criterion [19]. The selected pattern is denoted by $\mathcal{P}$, and we have

$$\mathcal{P} = \arg\min_{j} \left( \frac{1}{2} \max_{i} \left( \mathrm{SER}_i^{(j)} \right) \right). \tag{29}$$

Since $Q(\cdot)$ is a monotonic decreasing function, the Min-Max selection criterion in (29) can be further written as

$$\mathcal{P} = \arg\max_{j} \left( \frac{1}{2} \min_{i} \left( \rho_i^{(j)} \right) \right). \tag{30}$$

As the performance of the Min-Max criterion is almost the same as the optimal criterion, which will be verified through the simulations, then the average SER based on the Min-Max criterion is analyzed. Let $\mathrm{SER}_i^{mm}$ denote the instantaneous SER at node $N_i$, based on the Min-Max criterion. The average SER based on the Min-Max criterion can be written as

$$\overline{\mathrm{SER}}_{mm} = \frac{1}{2} \mathbb{E} \left[ \mathrm{SER}_1^{mm} + \mathrm{SER}_2^{mm} \right]. \tag{31}$$

Furthermore, as earlier discussed, the sum SER (31) can be lower bounded by using the larger SER

$$\begin{aligned} \overline{\mathrm{SER}}_{lb} &= \frac{1}{2} \mathbb{E} \left[ \max \left( \mathrm{SER}_1^{mm}, \mathrm{SER}_2^{mm} \right) \right] \\ &= \frac{1}{2} \mathbb{E} \left[ Q \left( \sqrt{b z_{\max}} \right) \right], \end{aligned} \tag{32}$$

where we define $z_{\max} = \max_{j} \left( \min_{i} \left( \rho_i^{(j)} \right) \right)$, and based on (20), the pdf of $z_{\max}$ is given by the order statistics [20]

$$f_{z_{\max}}(z) = \frac{4\eta P_s + 4}{P_s} \left( e^{-\frac{2\eta P_s + 2}{P_s} z} - e^{-\frac{4\eta P_s + 4}{P_s} z} \right). \tag{33}$$

Its corresponding cdf can be written as

$$F_{z_{\max}}(z) = 1 - 2e^{-\frac{2\eta P_s + 2}{P_s} z} + e^{-\frac{4\eta P_s + 4}{P_s} z}. \tag{34}$$

By substituting (34) into (32), the lower bound of the SER can be given by

$$\overline{\mathrm{SER}}_{lb} = \frac{1}{2} + \frac{1}{2} \sqrt{\frac{b P_s}{(b + 8\eta) P_s + 8}} - \sqrt{\frac{b P_s}{(b + 4\eta) P_s + 4}}. \tag{35}$$

At high SNR, (35) can be rewritten as

$$\lim_{P_s\to\infty} \overline{SER}_{lb} = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{b}{b+8\eta}} - \sqrt{\frac{b}{b+4\eta}}. \tag{36}$$

From (36), it indicates that the SER of the Min-Max criterion also converges to an error floor.

The average SER of traditional FD without TRAPS is also provided for comparison

$$\overline{\mathrm{SER}}_t = \frac{1}{2} - \frac{1}{2}\sqrt{\frac{bP_s}{(b+2\eta)P_s+2}}. \tag{37}$$

At high SNR, the SER of the traditional FD without TRAPS can be rewritten as

$$\lim_{P_s\to\infty} \overline{\mathrm{SER}}_t = \frac{1}{2} - \frac{1}{2}\sqrt{\frac{b}{b+2\eta}}. \tag{38}$$

By using Taylor expansion, the SER lower bound of the Min-Max criterion (36) and the average SER of traditional FD (38) can be rewritten as

$$\begin{aligned} \overline{\mathrm{SER}}_t &\approx \tfrac{\eta}{b} - \tfrac{3}{2}\left(\tfrac{\eta}{b}\right)^2 + \cdots, \\ \overline{\mathrm{SER}}_{lb} &\approx 6\left(\tfrac{\eta}{b}\right)^2 + \cdots. \end{aligned} \tag{39}$$

It clearly indicates in (39) that the error floor of TRAPS based on the Min-Max criterion is limited by $\eta^2$, while that of traditional FD is limited by $\eta$. Assuming $\eta \to 0$, we have $\overline{\mathrm{SER}}_{lb} \ll \overline{\mathrm{SER}}_t$, and the improvement of the average SER significantly increases as $\eta$ gets smaller.

Figure 10 compares the Min-SER approach with the traditional FD. The simulations are performed for a BPSK modulation, with $\eta = \{-3, -10, -20\}$ dB. The simulated SER based on the Min-Max criterion is also included, showing that the Min-Max criterion achieves almost the same performance as the optimal criterion. The analytical approximation using only the larger node's SER (35) is included, which is tight at medium to high SINR. Another important observation is that the SER curves converge to different error floors. The Min-SER approach has much lower error floor than the traditional FD, and the average SER gain gets larger with the decrease of the residual self interference.

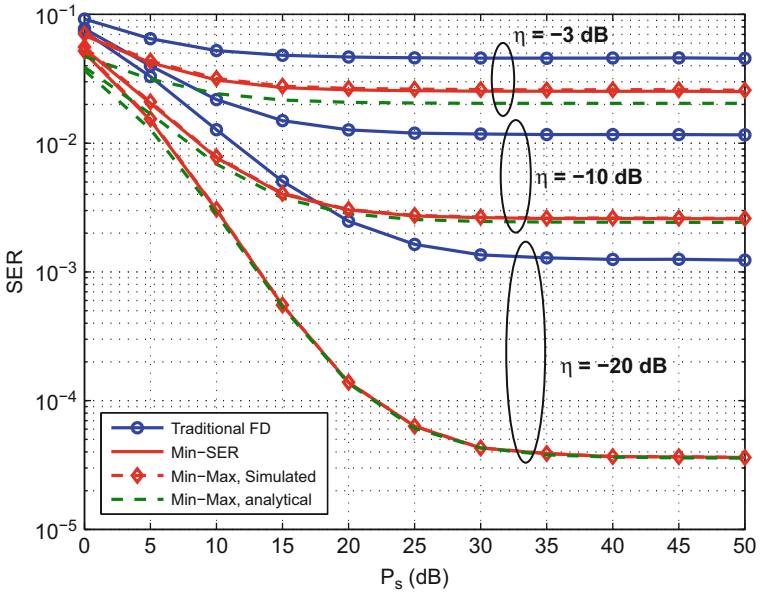


**Fig. 10** Analytical and simulated SER versus transmit power $P_s$ of Min-SER and traditional FD, with $\eta = \{-3, -10, -20\,\text{dB}\}$

## 3 Full-Duplex Cooperative Communications

Cooperative communications, or relaying, is a promising technology that can improve the throughput and coverage of the access points significantly. The traditional relay operates in the HD mode due to its implementation simplicity, leading to inefficient use of the limited spectral efficiency. To integrate the benefits of FD and relaying, it is natural to make relays operate in the FD mode [21–24]. In this section, two schemes in one-way and two-way FD relay communications will be introduced.

### *3.1 One-Way Joint Relay and Tx/Rx Antenna Mode Selection*

In this subsection, the joint relay and Tx/Rx antenna mode selection scheme (JRAMS) [23] is described, where there are one source (S), one destination (D), and $N$ AF relay nodes, as shown in Fig. 11. We assume that there is no direct channel link between the source and the destination due to the strong path loss and attenuation in this link. Both source and relay nodes use the same time-frequency resource, and

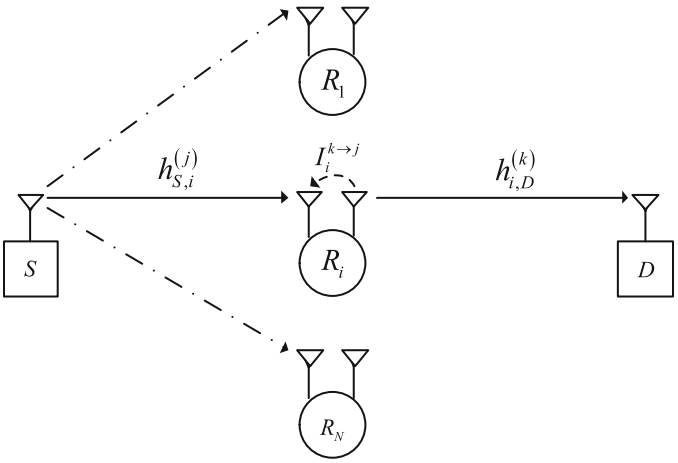


**Fig. 11** System model of the multiple AF relay system, $j, k \in \{1, 2\}$ and $j \neq k$

the relay nodes work in the full-duplex mode with two antennas (one for receiving and one for transmitting), denoted by $T_1$ and $T_2$. In the JRAMS scheme, only the optimal relay with the optimal configuration of FD Tx and Rx antenna mode is selected to forward the signals from the source to the destination. Note that for each relay, the Tx/Rx mode of the two antennas is adaptively determined based on the instantaneous channel conditions between the source/destination and the two antennas of the FD relay nodes.

Assuming that the $i$th relay node, denoted by $R_i$, is selected to forward the signal, and the relay node $R_i$ configures its antenna $T_j/T_k$ to receive/transmit the signal, $i \in \{1, 2, \ldots, N\}, j, k \in \{1, 2\}$ and $j \neq k$.

The signal received by antenna $T_j$ of the relay $R_i$ is given as

$$y_i^{(j)}[t] = h_{S,i}^{(j)}\sqrt{P_S}x[t] + I_i^{k\to j}\sqrt{P_R}x_i[t] + n_i^{(j)}[t], \tag{40}$$

where $h_{S,i}^{(j)}$ is the channel link between the source and the antenna $T_j$ of the relay $R_i$, and $I_i^{k\to j}$ is the residual loop interference from the transmit antenna $T_k$ to the receive antenna $T_j$ of the relay $R_i$. $x[t]$ and $x_i[t]$ are the $t$th signal with the unit power, transmitted from the source and the relay $R_i$, respectively. $P_S$ and $P_R$ are the transmit power of the source and relay node. $n_i^{(j)}$ is the additive white Gaussian noise at the receive antenna $T_j$ with the power $\sigma^2$.

Upon receiving the signal from the source, the relay uses the AF protocol to forward the following signal.

$$x_i[t] = \beta_i^{k\to j} y_i^{(j)}[t], \tag{41}$$

where $\beta_i^{k\to j}$ is the power amplification factor to ensure that the average power of signal $x_{R_i}[t]$ satisfies the following power constraint,

$$\mathbb{E}\left[|x_i[t]|^2\right] = \left(\beta_i^{k\to j}\right)^2 \left(|h_{S,i}^{(j)}|^2 P_S + |I_i^{k\to j}|^2 P_R + \sigma^2\right) \le 1. \tag{42}$$

From Eq. (42), we have

$$\beta_i^{k\to j} = \sqrt{\frac{1}{|h_{S,i}^{(j)}|^2 P_S + |I_i^{k\to j}|^2 P_R + \sigma^2}}. \tag{43}$$

The received signal at the destination is given by

$$y_D[t] = h_{i,D}^{(k)} \sqrt{P_R} x_i[t] + n_D[t], \tag{44}$$

where $h_{i,D}^{(k)}$ is the channel link from the transmit antenna $T_k$ of the relay $R_i$ to destination, and $n_D$ is the additive white Gaussian noise with mean zero and variance $\sigma^2$ at the destination.

The end-to-end SINR can be obtained as

$$\gamma_i^{k\to j} = \frac{P_S P_R \left|h_{S,i}^{(j)} h_{i,D}^{(k)} \beta_i^{k\to j}\right|^2}{P_R^2 \left|h_{i,D}^{(k)} I_i^{k\to j} \beta_i^{k\to j}\right|^2 + P_R \left|h_{i,D}^{(k)} \beta_i^{k\to j}\right|^2 \sigma^2 + \sigma^2}. \tag{45}$$

Substituting (43) into (45), the end-to-end SINR via the relay $R_i$ can be calculated as

$$\gamma_i^{k\to j} = \frac{P_S P_R \gamma_{S,i}^{(j)} \gamma_{i,D}^{(k)}}{P_S \gamma_{S,i}^{(j)} + (P_R \gamma_{i,D}^{(k)} + 1)(P_R \gamma_{LI,i}^{k\to j} + 1)}, \tag{46}$$

where $\gamma_{S,i}^{(j)} = |h_{S,i}^{(j)}|^2/\sigma^2$, $\gamma_{i,D}^{(k)} = |h_{i,D}^{(k)}|^2/\sigma^2$, and $\gamma_{LI,i}^{k\to j} = |I_i^{k\to j}|^2/\sigma^2$, representing the link gains between the source/destination and the relay $R_i$, and the interference gain from the transmit antenna to the receive antenna at the relay $R_i$ respectively.

The block Rayleigh fading are assumed for all links, where the channel remains unchanged during a transmit slot, and changes independently from one slot to another. The mean value of channel gains between the source/destination and the Tx/Rx antenna of the FD relay nodes, $\gamma_{S,i}^{(j)}$ and $\gamma_{i,D}^{(k)}$, are denoted by $\lambda_{S,i}^{(j)}$ and $\lambda_{i,D}^{(k)}$ respectively. We also assume that the average channel gains between the source/destination and the two antennas of the relay $R_i$ is same, $\lambda_{S,i}^{(1)} = \lambda_{S,i}^{(2)} = \lambda_{S,i}$ and $\lambda_{i,D}^{(1)} = \lambda_{i,D}^{(2)} = \lambda_{i,D}$. The residual loop interference between Tx and Rx antenna at relay is modeled as a block Rayleigh fading channel. The mutual interferences between Tx and Rx antenna at each relay node, $\gamma_{LI,i}^{1\to 2}$ and $\gamma_{LI,i}^{2\to 1}$ at the relay node $R_i$, are assumed to be the same due to the same loop interference cancellation method with the same configurations. Therefore, the interference gain of the relay $R_i$, $\gamma_{LI,i}^{1\to 2} = \gamma_{LI,i}^{1\to 2} = \gamma_{LI,i}$, follows the exponential distribution with the expectation of $\lambda_{LI,i}$. The interferences at different relay nodes are independent.

In the system under consideration, each relay node is equipped with two antennas. One antenna is used to receive the signal from the source node, and the other is to forward the signal at the same time. There are two possible channel link pairs between one relay and source/destination nodes, depending on which antenna of $T_1$ and $T_2$ is used as Tx and Rx antenna at relay. Those are $mode^{2\to1}$ where relay $R_i$ configures the antenna $T_1$ as the receive antenna and $T_2$ as the transmit antenna, and $mode^{1\to2}$ where relay $R_i$ configures $T_2$ as the receive antenna and $T_1$ as the transmit antenna. The optimal Tx/Rx antenna mode for relay $R_i$ is determined based on the maximal end-to-end SINR criterion as follows,

$$mode_i = \arg\max\{\gamma_i^{1\to2}, \gamma_i^{2\to1}\}, \tag{47}$$

where $\gamma_i^{2\to1}$ denotes the end-to-end SINR when the relay $R_i$ works under $mode^{2\to1}$, and can be computed via Eq. (46). Similarly, $\gamma_i^{1\to2}$ is the end-to-end SINR via the relay $R_i$ under $mode^{1\to2}$. The CSI of $h_{S,i}^{(j)}$ and $h_{i,D}^{(k)}$ can be measured by the standard pilot-based channel estimation. The CSIs of the loop interference channel $I_i^{j\to k}$ can be obtained through sufficient training. It is also assumed that the very reliable feedback channels are deployed, and CSIs can be estimated and transmitted to the decision node accurately.

In the multiple relay system, the optimal relay with the optimal Tx/Rx antenna mode configuration is selected to optimize the end-to-end SINR performance. Therefore, JRAMS scheme can be formulated as

$$\{R_{opt}, mode_{opt}\} = \arg\max_i\{\gamma_i\}, \tag{48}$$

where $\gamma_i = \max\{\gamma_i^{1\to2}, \gamma_i^{2\to1}\}$. At the beginning of each transmission, all the CSIs are measured and collected at a decision node (say source node). The system computes the end-to-end SINR of the $2N$ candidate relaying modes, and selects the optimal relay $R_{opt}$ with the optimal antenna mode configuration $mode_{opt}$ based on the end-to-end SINR maximization criterion in the JRAMS scheme as shown in equation above. The selected relay node $R_{opt}$ will be used to help to forward the signal from the source to the destination under the optimal mode $mode_{opt}$.

Then, the performance analysis of the JRAMS scheme is provided. The approximate CDF expression of the end-to-end SINR via the $i$th FD relay node, denoted by $\gamma_i$, can be calculated as [23]

$$F_i(x) = 1 - \frac{2}{1+\eta_i x} I_i(x) + \frac{1}{1+2\eta_i x} I_i^2(x), \tag{49}$$

where

$$I_i(x) = 2\sqrt{\frac{x^2+x}{P_S P_R \lambda_{S,i}\lambda_{i,D}}} K_1\left(2\sqrt{\frac{x^2+x}{P_S P_R \lambda_{S,i}\lambda_{i,D}}}\right) e^{-\left(\frac{1}{P_S\lambda_{S,i}} + \frac{1}{P_R\lambda_{i,D}}\right)x}. \tag{50}$$

and $K_1(\cdot)$ is the first order modified Bessel function of the second kind [15], and $\eta_i = P_R\lambda_{LI,i}/P_S\lambda_{S,i}$.

When the average link SNRs, $P_S\lambda_{S,i}$ and $P_R\lambda_{i,D}$, go to infinity, $2\sqrt{\frac{x^2+x}{P_SP_R\lambda_{S,i}\lambda_{i,D}}}$ $K_1\left(2\sqrt{\frac{x^2+x}{P_SP_R\lambda_{S,i}\lambda_{i,D}}}\right)$ approaches one, and thus $I_i(x) \to 1$. The distribution of the received SINR via the $i$th relay in the high transmit power scenarios will approach

$$F_i^{\infty}(x) = 1 - \frac{2}{1+\eta_i x} + \frac{1}{1+2\eta_i x}. \tag{51}$$

In the high link SNR conditions, the distribution of the end-to-end SINR approaches a special distribution which is only determined by the loop interference level $\eta_i$ at the relay node. This indicates that the performance of the FD relay system is similar in the high link SNR scenarios, and increasing the transmit power at the transmitter will be helpless to improving the system performance.

In this scheme, the optimal relay with the largest end-to-end SINR is selected to forward the signal. Therefore, the outage probability of JRAMS scheme of the $N$ relay networks can be calculated as

$$P_{out}(x) = \prod_{i=1}^{N}\left[1 - \frac{2}{1+\eta_i x}I_i(x) + \frac{1}{1+2\eta_i x}I_i^2(x)\right]. \tag{52}$$

The transmit power of the source and the relay is assumed same, $P_S = P_R = P_t$. In the high transmit power scenario, the function $I_i(x)$ approaches $e^{-(\frac{1}{P_t\lambda_{S,i}}+\frac{1}{P_t\lambda_{i,D}})x}$, the outage probability can be written as

$$P_{out}(x) \approx \prod_{i=1}^{N}\left(1 - \frac{2}{1+\eta_i x}e^{-\frac{1}{P_t}(\frac{1}{\lambda_{S,i}}+\frac{1}{\lambda_{i,D}})x} + \frac{1}{1+2\eta_i x}e^{-\frac{2}{P_t}(\frac{1}{\lambda_{S,i}}+\frac{1}{\lambda_{i,D}})x}\right). \tag{53}$$

Based on the finite-SNR diversity order defined in [25], the diversity order $d_{JRAMS}$ of the JRAMS scheme is obtained as

$$d_{JRAMS}(\lambda) = \sum_{i=1}^{N}\frac{2x}{P_t}\left(\frac{1}{\lambda_{S,i}}+\frac{1}{\lambda_{i,D}}\right)\frac{\frac{1}{1+\eta_i x}e^{-\frac{1}{P_t}(\frac{1}{\lambda_{S,i}}+\frac{1}{\lambda_{i,D}})x} - \frac{1}{1+2\eta_i x}e^{-\frac{2}{P_t}(\frac{1}{\lambda_{S,i}}+\frac{1}{\lambda_{i,D}})x}}{1 - \frac{2}{1+\eta_i x}e^{-\frac{1}{P_t}(\frac{1}{\lambda_{S,i}}+\frac{1}{\lambda_{i,D}})x} + \frac{1}{1+2\eta_i x}e^{-\frac{2}{P_t}(\frac{1}{\lambda_{S,i}}+\frac{1}{\lambda_{i,D}})x}}. \tag{54}$$

The diversity order can be well estimated as by using the Taylor expansion

$$d_{JRAMS}(\lambda) \approx \sum_{i=1}^{N} 2\frac{1+A_i}{1+2A_i+2A_i^2}, \tag{55}$$

where $A_i = \frac{P_t \eta_i \lambda_{S,i} \lambda_{i,D}}{\lambda_{S,i} + \lambda_{i,D}}$. It can be observed that when the transmit power $P_t$ increases to infinity, the diversity order approaches zero, representing the performance floor of the FD relaying system. It also shows that in the small loop interference condition, $A_i \to 0$, the diversity order of JRAMS scheme can reach almost $2N$ before the reaching the performance floor.

The performance floor of the outage probability are expressed as

$$P_{out}^{\infty}(x) = \prod_{i=1}^{N} \left(1 - \frac{2}{1 + \eta_i x} + \frac{1}{1 + 2\eta_i x}\right). \tag{56}$$

The performance floor of the proposed scheme is determined only by the loop interference level $\eta_i$ of the $N$ relay nodes, and is lower than that without the Tx/Rx antenna mode selection.

Figure 12 plots the outage performance versus the transmit power for the FD relay system with $N = 1, 2, 3$ and the loop interference level $\eta = 0.05$. The performance of the conventional optimal relay selection (ORS) scheme, proposed in [22], is also illustrated in Fig. 12 for comparison. It can be observed from the figure that there is a performance floor which coincides with the result computed by (56). In the same loop interference environment, the proposed JRAMS scheme outperforms that with the fixed antenna configuration obviously.

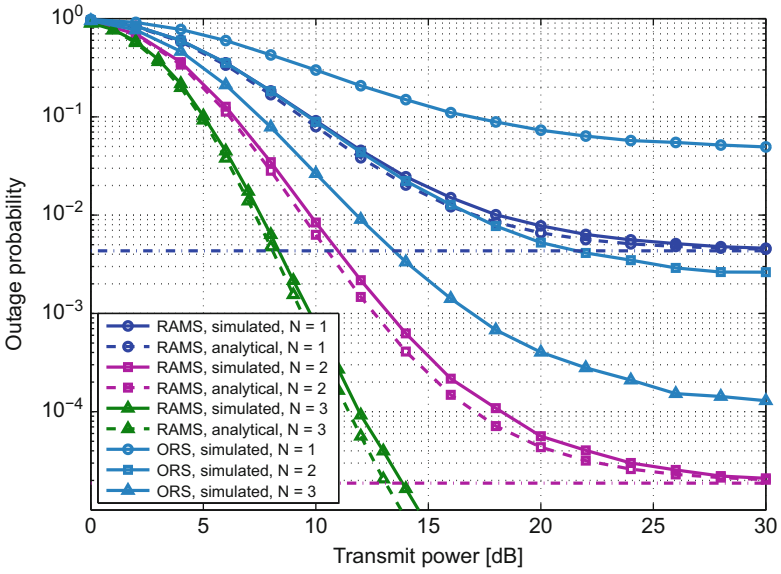


**Fig. 12** Outage probability of JRAMS scheme versus the transmit power, $\eta = 0.05$

### 3.2 Two-Way FD Relay Selection Scheme

In this subsection, a two-way FD relay selection (FRS) scheme with AF protocol [24] is introduced, as illustrated in Fig. 13. The system consists of two sources, denoted by $S_j$, $j = 1, 2$, and $N$ relays, denoted by $R_i$, $i = 1, \ldots, N$. The transmit powers of each source and each relay are denoted by $p_s$ and $p_r$, respectively. The direct link between sources does not exist due to the shadowing effect. The block-fading channel coefficient between $S_j$ and $R_i$ is reciprocal and denoted by $h_{ji}$, which follows independent complex-Gaussian distribution with zero mean and normalized variance.

In contrast to the system model of traditional two-way HD relay [19], each communication node in Fig. 13 operates in the FD mode. We assume that each node is equipped with one FD anttena [26, 27]. The residual self-interference channels at $S_j$ and $R_i$ are denoted by $\tilde{h}_{S_j}$ and $\tilde{h}_{R_i}$. Considering the transmission by the relay $R_i$ at the time slot $\kappa$, the sources $S_j$, $j = 1, 2$, transmit their respective symbols $x_j[\kappa]$ to $R_i$. Simultaneously, $R_i$ broadcasts another symbol $t_i[\kappa]$ to the sources. Therefore, the received signal at $R_i$ is

$$r_i[\kappa] = \sqrt{p_s} x_1[\kappa] h_{1i} + \sqrt{p_s} x_2[\kappa] h_{2i} + \tilde{h}_{R_i} t_i[\kappa] + n_{r_i}[\kappa] \tag{57}$$

where $\mathbb{E}\left\{x_j[\kappa]^2\right\} = 1$, $n_{r_i}[\kappa]$ is the additive white Gaussian noise (AWGN) at $R_i$ with zero-mean and variance of $\sigma_n^2$, and the residual self-interference at the relay $e_{r_i} = \tilde{h}_{R_i} t_i[\kappa]$ is modeled by AWGN with zero-mean and variance of $\sigma_{e,r}^2$.

The relay $R_i$ adopts the AF protocol, and thus $t_i[\kappa]$ is the amplification of the prior received signal, i.e.,

$$t_i[\kappa] = \sqrt{p_r} \beta_{r_i} r_i[\kappa - 1] \tag{58}$$

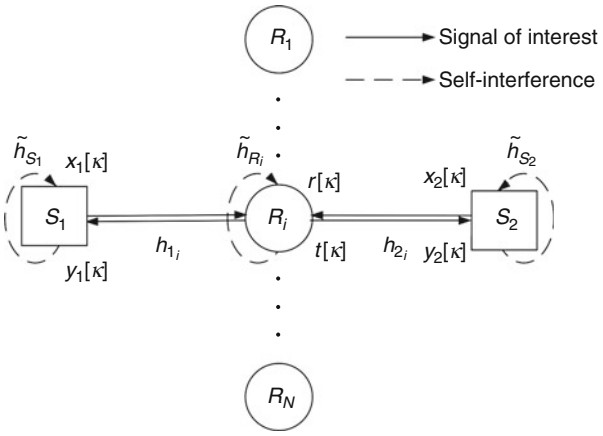


**Fig. 13** System model of two-way FRS scheme

where $\beta_{r_i}$ is the amplification factor of $R_i$. From (57) and (58), $t_i[\kappa]$ can be expanded as

$$t_i[\kappa] = \beta_{r_i}\sqrt{p_r}\sum_{m=1}^{\infty}\left(\tilde{h}_{r_i}\beta_{r_i}\sqrt{p_r}\right)^{m-1}\left\{h_{1i}\sqrt{p_s}x_1[\kappa-m]+h_{2i}\sqrt{p_s}x_2[\kappa-m]+n_{r_i}[\kappa-m]\right\}. \tag{59}$$

To prevent the oscillation at $R_i$, the amplification factor $\beta_{r_i}$ should be

$$\beta_{r_i} = \frac{1}{\sqrt{p_s|h_{1i}|^2 + p_s|h_{2i}|^2 + p_r|\tilde{h}_{R_i}|^2 + \sigma_n^2}} \tag{60}$$

Therefore, according to (59) and (60), the variance of $t_i[\kappa]$ is finite, i.e.,

$$\mathbb{E}\{t_i[\kappa]^2\} = \left(|h_{1i}|^2 p_s + |h_{2i}|^2 p_s + \sigma_n^2\right)\frac{\beta_{r_i}^2 p_r}{1-|\tilde{h}_{R_i}|^2\beta_{r_i}^2 p_r} = p_r. \tag{61}$$

The received signal at $S_j, j = 1, 2$, is expressed as

$$y_j[\kappa] = h_{ji}t_i[\kappa] + n_j[\kappa] + \tilde{h}_{S_j}\sqrt{p_s}x_j[\kappa] \tag{62}$$

where $n_j[t]$ is the AWGN at $S_j$, and the residual self-interference at the source $e_{s_j} = \tilde{h}_{S_j}\sqrt{p_s}x_j[\kappa]$ is modeled by AWGN with zero-mean and variance of $\sigma_{e,s}^2$. Note that the self-interference at sources and that at relays have different impact on the performance. Specifically, the self-interference at the relay is iterative because of (57) and (58), and might make the relay oscillation, whereas the self-interference at the source is non-iterative.

According to (57), (58), and (62), the received signal at $S_j$ can be expanded as

$$y_j[\kappa] = h_{ji}h_{ji}\beta_{r_i}\sqrt{p_r p_s}x_j[\kappa-1] + h_{ji}h_{\bar{j}i}\beta_{r_i}\sqrt{p_r p_s}x_{\bar{j}}[\kappa-1] \tag{63}$$

$$+ h_{ji}\beta_{r_i}\sqrt{p_r}\tilde{h}_{R_i}t_i[\kappa-1] + \tilde{h}_{S_j}\sqrt{p_s}x_j[\kappa] \tag{64}$$

$$+ h_{ji}\beta_{r_i}\sqrt{p_r}n_{r_i}[\kappa-1] + n_j[\kappa] \tag{65}$$

where $\{j, \bar{j}\} = \{1, 2\}$ or $\{2, 1\}$. The first term in (63) can be subtracted totally because of network coding technique [28], and the second term in (63) is the useful signal from $S_{\bar{j}}$; the first term and the second term in (64) are the residual self-interference from $R_i$ and $S_j$, respectively; the first term and the second term in (65) are the noise from $R_i$ and $S_j$, respectively.

After some manipulation, the instantaneous received SINR at $S_j$ via $R_i$ is written as

$$\gamma_{ji} = \frac{\tilde{\psi}_s|h_{ji}|^2\tilde{\psi}_r|h_{\bar{j}i}|^2}{\left(\tilde{\psi}_s+\tilde{\psi}_r\right)|h_{ji}|^2 + \tilde{\psi}_s|h_{\bar{j}i}|^2 + 1}, \tag{66}$$

where $\tilde{\psi}_s = \frac{p_s}{(\sigma_{e,r}^2+1)\sigma_n^2}$ and $\tilde{\psi}_r = \frac{p_r}{(\sigma_{e,s}^2+1)\sigma_n^2}$, and $\sigma_{e,s}^2$ and $\sigma_{e,r}^2$ are the variances of residual self-interference at the sources and the relays, respectively.

According to [24], in the two-way relay, the optimal FRS scheme is

$$k = \max_{i=1,\ldots,N} \min\left(|h_{1i}|^2, |h_{2i}|^2\right) \tag{67}$$

where $k$ is the index of the selected relay. The RS scheme (67) is optimal in minimizing the outage probability and symbol error rate, and also maximizing the ergodic capacity.

In the following, only the bit error rate (BER) analysis is provided for the two-way FRS scheme, other performance analysis can be found in [24]. Prior to that, the distribution function of $\gamma_{jk}$ is calculated, which is determined by the distribution of $|h_{jk}|^2$ according to (66). According to the RS scheme (67), the CDF of $|h_{jk}|^2$ is expressed as

$$F_{|h_{jk}|^2}(z) = 1 - N\sum_{m=0}^{N-1}\binom{N-1}{m}\frac{(-1)^m}{2m+1}\left\{\exp(-z) + \frac{m}{m+1}\exp\left[-2(m+1)z\right]\right\} \tag{68}$$

where $j = 1, 2$.

With the help of the CDF of $|h_{jk}|^2$, a tight lower bound of BER for $S_j$ can be obtained

$$\overline{BER_j} \geq \frac{\alpha}{2} - \frac{3\sqrt{2}\pi\alpha\sqrt{\beta}}{2\tilde{\psi}_r\tilde{\psi}_h}N^2\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}\binom{N-1}{m}\binom{N-1}{n}\frac{(-1)^m}{2m+1}\frac{(-1)^n}{2n+1}\left[f_{11}+f_{12}+f_{21}+f_{22}\right], \tag{69}$$

where

$$f_{11} = \left[\left(\sqrt{\frac{1}{\tilde{\psi}_r}}+\sqrt{\frac{1}{\tilde{\psi}_h}}\right)^2+\frac{\beta}{2}\right]^{-\frac{5}{2}} F\left(\frac{5}{2},\frac{3}{2};2;\frac{\left(\sqrt{1/\tilde{\psi}_r}-\sqrt{1/\tilde{\psi}_h}\right)^2+\beta/2}{\left(\sqrt{1/\tilde{\psi}_r}+\sqrt{1/\tilde{\psi}_h}\right)^2+\beta/2}\right), \tag{70}$$

$$f_{12} = 2n\left[\left(\sqrt{\frac{1}{\tilde{\psi}_r}}+\sqrt{\frac{\xi_n}{\tilde{\psi}_h}}\right)^2+\frac{\beta}{2}\right]^{-\frac{5}{2}} F\left(\frac{5}{2},\frac{3}{2};2;\frac{\left(\sqrt{1/\tilde{\psi}_r}-\sqrt{\xi_n/\tilde{\psi}_h}\right)^2+\beta/2}{\left(\sqrt{1/\psi_r}+\sqrt{\xi_n/\tilde{\psi}_h}\right)^2+\beta/2}\right), \tag{71}$$

$$f_{21} = 2m\left[\left(\sqrt{\frac{\xi_m}{\tilde{\psi}_r}} + \sqrt{\frac{1}{\tilde{\psi}_h}}\right)^2 + \frac{\beta}{2}\right]^{-\frac{5}{2}} F\left(\frac{5}{2}, \frac{3}{2}; 2; \frac{\left(\sqrt{\xi_m/\tilde{\psi}_r} - \sqrt{1/\tilde{\psi}_h}\right)^2 + \beta/2}{\left(\sqrt{\xi_m/\tilde{\psi}_r} + \sqrt{1/\tilde{\psi}_h}\right)^2 + \beta/2}\right), \quad (72)$$

and

$$f_{22} = 4mn\left[\left(\sqrt{\frac{\xi_m}{\tilde{\psi}_r}} + \sqrt{\frac{\xi_n}{\tilde{\psi}_h}}\right)^2 + \frac{\beta}{2}\right]^{-\frac{5}{2}} F\left(\frac{5}{2}, \frac{3}{2}; 2; \frac{\left(\sqrt{\xi_m/\tilde{\psi}_r} - \sqrt{\xi_n/\tilde{\psi}_h}\right)^2 + \beta/2}{\left(\sqrt{\xi_m/\tilde{\psi}_r} + \sqrt{\xi_n/\tilde{\psi}_h}\right)^2 + \beta/2}\right). \quad (73)$$

In addition, $\xi_m = 2m + 2$, $\xi_n = 2n + 2$, $\tilde{\psi}_h = \frac{\tilde{\psi}_s\tilde{\psi}_r}{\tilde{\psi}_s + \tilde{\psi}_r}$, and $F(a, b; c; z)$ is the Confluent Hypergeometric function [15].

Figure 14 depicts the BER performance versus SNR $= p/\sigma_n^2$, when the variance of residual self-interference $\sigma_e^2 = 0.1,\ 0.5,$ and 2, and the number of relay $N = 4$. The figure reveals that the two-way FD relay has significant gain than the two-way HD relay all over the SNR regime, when the residual self-interference is small. For instance, the performance gain is about 1 dB in high SNR when $\sigma_e^2 = 0.5$, and the

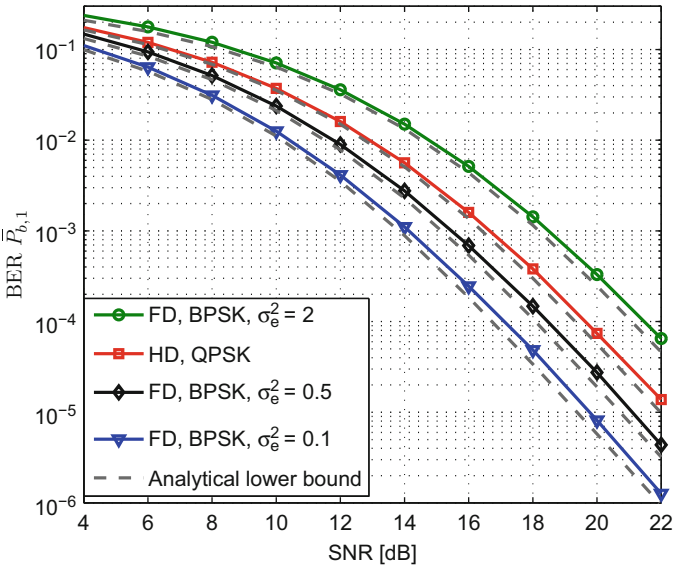


**Fig. 14** BER of the two-way FD relay selection and two-way HD relay selection, when the number of relay $N = 4$

gain increases to 2 dB when the variance of the residual self-interference decreases to $\sigma_e^2 = 0.1$. However, when the residual self-interference is large, the two-way FD relay has worse performance than the two-way HD relay.

## 4 Full-Duplex Cognitive Radio Networks

Cognitive radio, which enables spectrum sharing between licensed (PUs) and unlicensed (SUs) users, has been regarded as a promising solution to spectrum shortage, and caused wide attention for more than a decade [29, 30]. Conventionally, the so-called "listen-before-talk" strategy in which SUs sense the target channel before transmission has been extensively studied [31]. In the listen-before-talk scheme, the sensing and transmission are separated in time domain, which leads to two main problems: (1) transmission time reduction due to sensing, and (2) sensing accuracy impairment due to data transmission. With the help of FD techniques, SUs no longer need to stop their transmissions periodically for sensing. Instead, they can perform simultaneous spectrum sensing and data transmission, which can naturally bypass the above problems. In this section, a "listen-and-talk" (LAT) protocol for FD CR users based on the idea of simultaneous sensing and transmission [32, 33] is introduced. The protocol design as well as detailed performance analysis is provided.

### *4.1 Listen-and-Talk Protocol*

A CRN is considered, consisting of one PU and one SU pair, where $\mathrm{SU}_1$ needs to transmit data to $\mathrm{SU}_2$. Each SU is equipped with two antennas $\mathrm{Ant}_1$ and $\mathrm{Ant}_2$. The spectrum band occupancy by the PU can be modeled as an alternating ON/OFF random process where the PU can access the spectrum at any time. We assume that the probabilities of the PU's arrival and departure remain the same across the time, and the holding time of either state is distributed as the exponential distribution [34]. We denote the variables of the idle period and busy period of the PU as $t_0$ and $t_1$, respectively. And let $\tau_0 = \mathbb{E}[t_0]$ and $\tau_1 = \mathbb{E}[t_1]$ represent the average idle and busy duration.

Figure 15 shows the sensing and spectrum access procedure of the LAT protocol. $\mathrm{SU}_1$ performs sensing and transmission simultaneously by using the FD technique: $\mathrm{Ant}_1$ senses the spectrum continuously while $\mathrm{Ant}_2$ transmits data when a spectrum hole is detected. Specifically, $\mathrm{SU}_1$ keeps sensing the spectrum with $\mathrm{Ant}_1$ with sampling frequency $f_s$, which is shown in the line with down arrows. At the end of each slot with duration $T$, $\mathrm{SU}_1$ combines all samples in the slot and makes the decision of the PU's presence. The decisions are represented by the small circles, in which the higher ones denote that the PU is judged active, while the lower ones denote otherwise. The activity of $\mathrm{SU}_1$ is instructed by the sensing decisions,

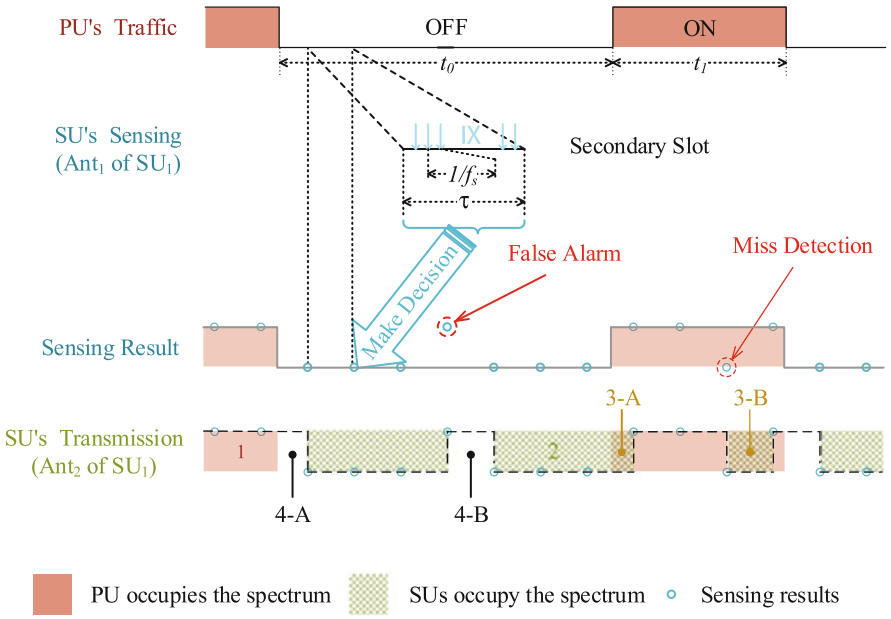


**Fig. 15** The LAT protocol

i.e., $SU_1$ can access the spectrum in the following slot when the PU is judged absent, and it needs to backoff otherwise.

On the implementation of the LAT protocol, the residual self-interference becomes a major impediment to accurate sensing, i.e., when $SU_1$ is transmitting, part of the transmitted signal will be received by the sensing antenna $Ant_1$, which obscures the potential signal from the PU. Specifically, the received signal at $Ant_1$ can be expressed as

$$y = \begin{cases} h_s s_p + \delta_1 w + u, & \mathcal{H}_1, \\ \delta_1 w + u, & \mathcal{H}_0, \end{cases} \tag{74}$$

where $\mathcal{H}_1$ and $\mathcal{H}_0$ are the hypotheses under which the PU is busy or idle, respectively; $s_p$ denotes the signal of the PU, $h_s$ is the channel from the PU to $Ant_1$ of $SU_1$, and $u \sim \mathcal{CN}\left(0, \sigma_u^2\right)$ denotes the complex-valued Gaussian noise. Without loss of generality, we assume that $s_p$ is PSK modulated with variance $\sigma_p^2$, and $h_s$ is a Rayleigh channel with zero mean and variance $\sigma_h^2$. Variable $w$ denotes the residual self-interference at $Ant_1$, which is modeled as Rayleigh distribution with zero mean and variance $\chi^2\sigma_s^2$, with $\sigma_s^2$ denoting the secondary transmit power and $\chi^2$ representing the degree of self-interference suppression, which is defined as

$$\chi^2 := \frac{\text{Power of the RSI}}{\text{Transmit power}}.$$

The parameter $\chi^2$ is commonly expressed in dBs, and indicates how well can the self-interference be suppressed. The Boolean variable $\delta_1$ represents $\text{SU}_1$'s activity. As shown above, the distribution of the received signal for sensing is related to the activity of $\text{SU}_1$, indicating that the sensing strategy needs to be adjusted according to $\text{SU}_1$'s own activity.

For simplicity and without loss of generality, energy detection is adopted as the sensing scheme, and the test statistics $O$ can be given as:

$$O = \frac{1}{N_s} \sum_{n=1}^{N_s} |y(n)|^2, \tag{75}$$

where $y(n)$ denotes the $n$th sample received by $\text{Ant}_1$ in a slot. With a chosen threshold $\epsilon$, the spectrum is judged occupied when $O \geq \epsilon$, otherwise the spectrum is idle, and the false alarm and miss detection probabilities can be defined as, respectively,

$$\begin{aligned} P_f(\epsilon) &= \Pr(O > \epsilon | \mathcal{H}_0), \\ P_m(\epsilon) &= \Pr(O < \epsilon | \mathcal{H}_1). \end{aligned} \tag{76}$$

As mentioned above, the sensing threshold needs to be changed according to $\text{SU}_1$'s activity, i.e., when $\text{SU}_1$ is silent, it chooses $\epsilon_0$ as the sensing threshold, and when it is active, it chooses a different $\epsilon_1$. In the design of the sensing threshold, one of the most important constraints is that the interference to the PU must be under a certain level. This constraint is considered as the collision ratio between SUs and the PU, defined as

$$P_c = \lim_{t \to \infty} \frac{\text{Collision duration}}{\text{PU}'\text{s transmission time during } [0, t]}.$$

Combining the PU's state change and $\text{SU}_1$'s activity, the collision ratio can be derived following the solid arrows in Fig. 16, in which the "state transition" block refers to the Markov chain among the four states of the system: (1) only the PU uses the spectrum; (2) only the SUs use the spectrum; (3) collision; and (4) spectrum waste. Further, for simplicity in threshold design, we artificially add a constraint $P_m^0 = P_m^1 = P_m$, i.e., the miss detection probability remains the same when $\text{SU}_1$ is either active or silent. $P_c - \left(1 - e^{-\frac{T}{\tau_1}}\right)/2$ is a good approximation of $P_m$ in common cases where the PU changes its activity sufficiently slowly.

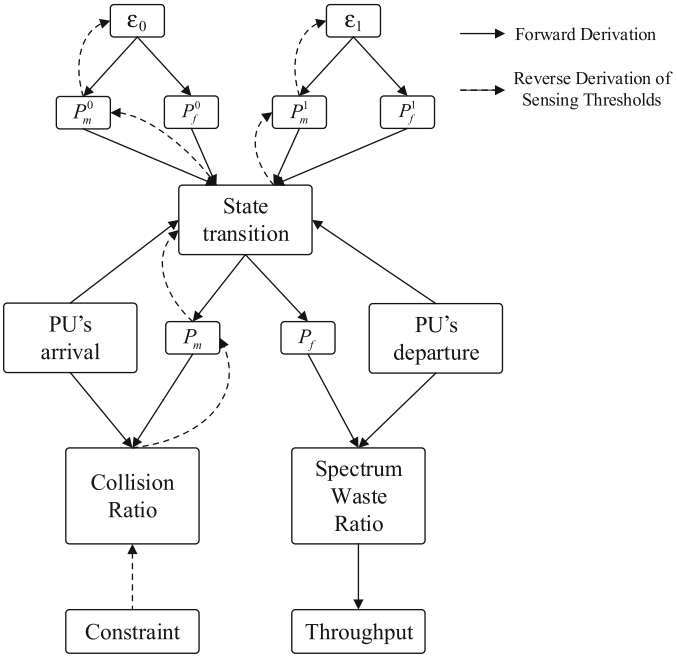


**Fig. 16** The schematic design procedure of the sensing thresholds of the LAT protocol

Then, the sensing thresholds are given by

$$\begin{aligned}\epsilon_{\delta_1} = \epsilon_{\delta_1}(P_c) &= \left(\frac{\mathcal{Q}^{-1}(1-P_m)}{\sqrt{N_s}} + 1\right)(1+\gamma_s+\gamma_i\delta_1)\,\sigma_u^2 \\ &= \left(\frac{\mathcal{Q}^{-1}\left(\frac{3}{2}-P_c-\frac{1}{2}e^{-T/\tau_1}\right)}{\sqrt{N_s}} + 1\right)(1+\gamma_s+\gamma_i\delta_1)\,\sigma_u^2,\ \delta_1 \in \{0,1\}\end{aligned} \tag{77}$$

in which $\gamma_s = \frac{\sigma_p \sigma_h^2}{\sigma_u^2}$ denotes the signal-to-noise ratio (SNR) in sensing, and $\gamma_i = \frac{\chi^2 \sigma_s^2}{\sigma_u^2}$ is the interference-to-noise ratio (INR).

## 4.2 Spectrum Utilization Efficiency

With the thresholds in (77), the spectrum utilization efficiency can be derived by following the procedure in Fig. 16. Specifically, the spectrum waste ratio, defined as

$$P_w = \lim_{t\to\infty} \frac{\text{Duration of unused spectrum hole}}{\text{Total length of spectrum holes during } [0,t]},$$

can be calculated as

$$\begin{aligned} P_w &= \frac{\mu}{2} + \left.\frac{\left(\frac{1}{\mu}-1\right)P_f^1 + 1 - P_m}{1+\left(\frac{1}{\mu}-1\right)\left(1-P_f^0+P_f^1\right)}\right|_{\mu=1-e^{-\frac{T}{\tau_0}}} \\ &= \frac{1-e^{-T/\tau_0}}{2} + \frac{\frac{e^{-T/\tau_0}}{1-e^{-T/\tau_0}}P_f^1 + 1 - P_m}{1+\frac{e^{-T/\tau_0}}{1-e^{-T/\tau_0}}\left(1-P_f^0+P_f^1\right)}, \end{aligned} \tag{78}$$

where $P_f^0$ and $P_f^1$ denote the false alarm probability when $\mathrm{SU}_1$ is silent and active, respectively, which can be expressed as

$$P_f^{\delta_1}(P_m) = \mathcal{Q}\left(\mathcal{Q}^{-1}(1-P_m)\left(1+\frac{\gamma_s}{1+\gamma_i\delta_1}\right)+\frac{\gamma_s}{1+\gamma_i\delta_1}\sqrt{N_s}\right), \tag{79}$$

where $\mathcal{Q}(\cdot)$ is the complementary distribution function of the standard Gaussian distribution.

In Fig. 17, the receiver operating characteristic curves (ROCs) is used to present the sensing performance. Both the ROCs of LAT and conventional "listen-before-talk" (LBT) protocols are presented. For comparison fairness, in the calculation and

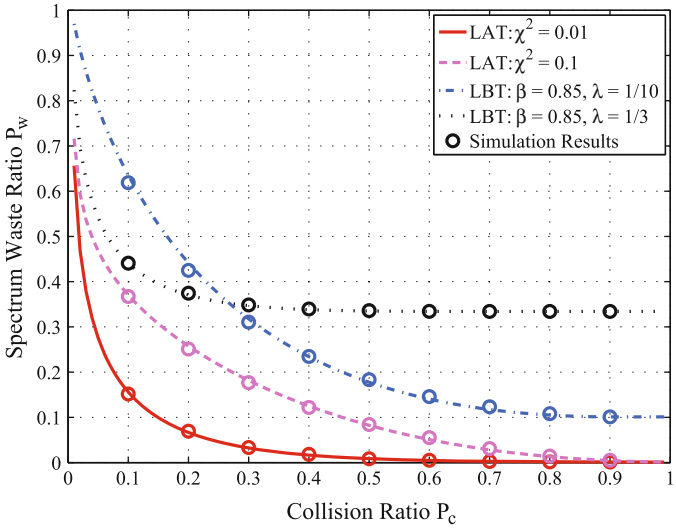


**Fig. 17** ROCs in sensing. In this figure, the probability of the PU's arrival $\mu = 1/500$, departure $\nu = 6/500$, the sample number of a slot $N_s$ is 300, normalized secondary transmit power $\sigma_s^2/\sigma_u^2 = 10\,\mathrm{dB}$, sensing SNR $\gamma_s = -8\,\mathrm{dB}$, the RSI factor $\chi^2$ in the LAT varies between 0.1 and 0.01, the spatial correlation coefficient $\beta$ is 0.85, and the ratio of sensing duration changes between 1/3 and 1/10

simulation of the LBT protocol, we also assume that each SU is equipped with two antennas, which means that $SU_1$ can use two antennas for sensing. The spatial correlation of the two antennas is represented by $\beta$, and the percentage of sensing duration in a slot is denoted by $\lambda$. In Fig. 17, smaller area under a curve denotes better sensing performance, i.e., the sensing performance of the LAT is much better than the LBT. It can be seen that for the curves in the LAT protocol, the solid red line is lower than the dashed pink one, which indicates that smaller residual self-interference leads to better sensing performance. It is noteworthy that the ratio of spectrum waste of the LBT protocol can never be lower than $\lambda$, while that of the LAT can be quite close to zero if the PU's state change is sufficiently slow. Also, in the left side of Fig. 17 where the allowable $P_c$ is small, $P_w$ in the LAT decreases much more sharply with the increase of $P_c$ than $P'_w$ in the LBT, which also implies better sensing performance of the LAT in the real case when $P_c$ is strictly constrained.

### 4.3 Power-Throughput Tradeoff

The secondary throughput can be obtained from (78):

$$C = R \cdot (1 - P_w) = \log_2(1 + \gamma_t) \cdot \left( \frac{1 + e^{-T/\tau_0}}{2} - \frac{\frac{e^{-T/\tau_0}}{1 - e^{-T/\tau_0}} P_f^1 + 1 - P_m}{1 + \frac{e^{-T/\tau_0}}{1 - e^{-T/\tau_0}} \left(1 - P_f^0 + P_f^1\right)} \right), \tag{80}$$

where $\gamma_t = \frac{\sigma_s^2 \sigma_t^2}{\sigma_u^2}$ represents the SNR in transmission, with $\sigma_t^2$ denotes the pass loss of the transmit channel from $SU_1$ to $SU_2$.

In the expression of $SU_1$'s throughput in (80), there are two factors: $R$ and $(1 - P_w)$. On one hand, $R$ is positively proportional to $SU_1$'s transmit power $\sigma_s^2$. On the other hand, however, it can be proved that the spectrum waste ratio $P_w$ increases with the secondary transmit power $\sigma_s^2$. Thus, there may exist a tradeoff between secondary transmit power and secondary throughput in this protocol: when the secondary transmit power is low, the residual self-interference is negligible, the spectrum is used more fully with small $P_w$, yet the ceiling throughput is limited by $R$; when the transmit power increases, the sensing performance get deteriorated, while at the same time, $SU_1$ can transmit more data in a single slot.

Figure 18 shows the throughput performance of the LAT protocol in terms of secondary transmit power. The solid and dotted lines represent the analytical performance of the LAT protocol, and the asterisks (*) denote the analytical local optimal transmit power. The small circles are the simulated results, which match the analytical performance well. The thin solid line depicts the ideal case with perfect self-interference cancelation. Without residual self-interference, the sensing performance is no longer affected by transmit power, and the throughput always goes up with the power. This line is also the upperbound of the LAT performance. The thick dash-dotted, dotted and dash lines in the middle are the typical cases, in

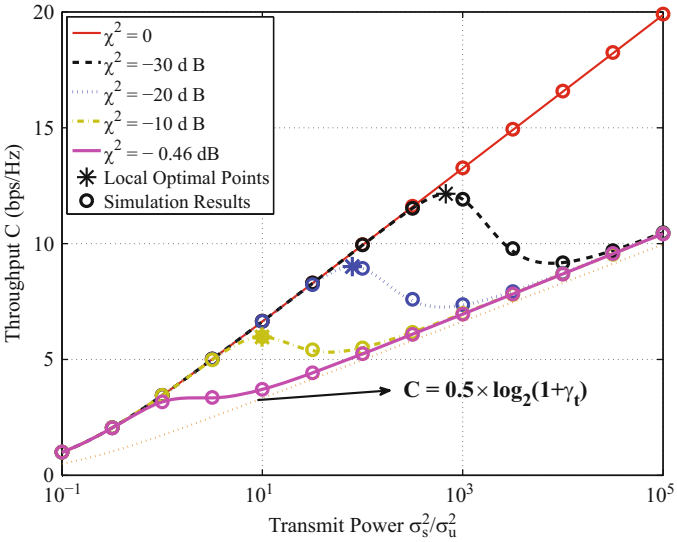


**Fig. 18** Power-Throughput curves in terms of different degree of self-interference cancelation factor $\chi^2$, where the probability of the PU's arrival is 0.002/slot, departure is 0.012/slot, the collision ratio $P_c = 0.1$, the sample number of a slot $N_s$ is 300, sensing SNR $\gamma_s = -5$ dB

which we can clearly observe the power-throughput tradeoff and identify the local optimal power. With the improvement of self-interference cancelation ($\chi^2$ from 0.1 to 0.01 to 0.001), the local optimal transmit power increases, and the corresponding throughput goes to a higher level. This makes sense since the smaller the residual self-interference is, the better it approaches the ideal case, and the deterioration cause by self-interference becomes dominant under a higher power. Also, when $\chi^2$ is sufficiently large, i.e., $\chi^2 = 0.9$ in the figure, there exists no power-throughput tradeoff.

One noticeable feature of Fig. 18 is that when self-interference exists, all curves approach the thin dotted line $C = 0.5 \log_2(1 + \gamma_t)$ when the power goes up. This line indicates the case that the spectrum waste is 0.5. When the transmit power is too large, severe self-interference largely degrades the performance of spectrum sensing, and the false alarm probability becomes unbearably high. It is likely that whenever $SU_1$ begins transmission, the spectrum sensing result falsely indicates that the PU has arrived due to false alarm, and $SU_1$ stops transmission in the next slot. Once $SU_1$ becomes silent, it can clearly detect the PU's absence, and begins transmission in the next slot again. And the state of $SU_1$ changes every slot even when the PU does not arrive at all. In this case, the utility efficiency of the spectrum hole is approximately 0.5, which is clearly shown in Fig. 18. Also, it can be seen that the larger $\chi^2$ is, the earlier the sensing gets unbearable and the throughput approaches the orange line.

## 5 Conclusions and Future Work

In this chapter, FD wireless communications for 5G have been presented, which allows simultaneous transmission and reception over the same frequency band, and potentially increases the spectral efficiency. The most fundamental challenge of FD implementation comes from the performance degradation induced by SI. To this end, we described several SI cancelation techniques in propagation-domain, analog-circuit-domain and digital-domain, which enable FD communications in different application scenarios, such as bidirectional communications, cooperative networks, and cognitive networks. In addition, the performance of several schemes was analyzed, including HD/FD mode switching and TRAPS scheme in FD bidirectional communications, one-way JRAMS and two way RS scheme in FD cooperative networks, and LAT in FD CRN. The existing work mainly focused on the analysis of physical layer, however, much work remains to be done for analysis of higher layers, such as MAC layer and network layer. These are great opportunities for inter-discipline wireless researchers to develop more efficient wireless networks in the future.

## References

1. J.I. Choi, M. Jain, K. Srinivasan, P. Levis, S. Katti, Achieving single channel, full duplex wireless communication, in *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking* (ACM, New York, 2010), pp. 1–12
2. A. Sabharwal, P. Schniter, D. Guo, D.W. Bliss, S. Rangarajan, R. Wichman, In-band full-duplex wireless: challenges and opportunities. IEEE J. Sel. Areas Commun. **32**(9), 1637–1652 (2014)
3. Y. Liao, K. Bian, L. Song, Z. Han, Full-duplex mac protocol design and analysis. IEEE Commun. Lett. **19**(7), 1185–1188 (2015)
4. M. Duarte, A. Sabharwal, Full-duplex wireless communications using off-the-shelf radios: feasibility and first results, in *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)* (IEEE, New York, 2010), pp. 1558–1562
5. E. Everett, A. Sahai, A. Sabharwal, Passive self-interference suppression for full-duplex infrastructure nodes. IEEE Trans. Wirel. Commun. **13**(2), 680–694 (2014)
6. A. Sahai, G. Patel, A. Sabharwal, Pushing the limits of full-duplex: design and real-time implementation, arXiv preprint arXiv:1107.0607, 2011
7. E. Aryafar, M.A. Khojastepour, K. Sundaresan, S. Rangarajan, M. Chiang, MIDU: enabling MIMO full duplex, in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking* (ACM, New York, 2012), pp. 257–268
8. E. Everett, M. Duarte, C. Dick, A. Sabharwal, Empowering full-duplex wireless communication by exploiting directional diversity, in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)* (IEEE, New York, 2011), pp. 2002–2006
9. M. Duarte, C. Dick, A. Sabharwal, Experiment-driven characterization of full-duplex wireless systems. IEEE Trans. Wirel. Commun. **11**(12), 4296–4307 (2012)
10. M. Duarte, A. Sabharwal, V. Aggarwal, R. Jana, K. Ramakrishnan, C.W. Rice, N. Shankaranarayanan, Design and characterization of a full-duplex multiantenna system for WIFI networks. IEEE Trans. Veh. Technol. **63**(3), 1160–1177 (2014)

11. A. Sahai, G. Patel, C. Dick, A. Sabharwal, On the impact of phase noise on active cancelation in wireless full-duplex. IEEE Trans. Veh. Technol. **62**(9), 4494–4510 (2013)
12. B.P. Day, A.R. Margetts, D.W. Bliss, P. Schniter, Full-duplex bidirectional MIMO: achievable rates under limited dynamic range. IEEE Trans. Signal Process. **60**(7), 3702–3713 (2012)
13. M. Kiessling, J. Speidel, Mutual information of MIMO channels in correlated Rayleigh fading environments-a general solution, in *IEEE International Conference on Communications, 2004*, vol. 2 (IEEE, New York, 2004), pp. 814–818
14. M. Zhou, H. Cui, L. Song, Y. Li, Is full duplex configuration better than mimo spatial multiplexing? in *2015 IEEE International Conference on Communications in China* (IEEE, New York, 2015)
15. M. Abramowitz, I.A. Stegun, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables* (Courier Corporation, New York, 2012)
16. A. Paulraj, R. Nabar, D. Gore, *Introduction to Space-Time Wireless Communications* (Cambridge University Press, Cambridge, 2003)
17. M. Zhou, H. Cui, L. Song, B. Jiao, Transmit-receive antenna pair selection in full duplex systems. IEEE Trans. Wirel. Commun. **3**(1), 34–37 (2014)
18. M. Zhou, L. Song, Y. Li, X. Li, Simultaneous bidirectional link selection in full duplex MIMO systems. IEEE Trans. Wirel. Commun. **14**(7), 4052–4062 (2015)
19. L. Song, Relay selection for two-way relaying with amplify-and-forward protocols. IEEE Trans. Veh. Technol. **60**(4), 1954–1959 (2011)
20. H.A. David, H.N. Nagaraja, *Order Statistics*. Wiley Online Library, 1970
21. G.-P. Liu, R. Yu, H. Ji, V. Leung, X. Li, In-band full-duplex relaying: a survey, research issues and challenges. IEEE Commun. Surv. Tutorials **17**(2), 500–524 (2015)
22. I. Krikidis, H. Suraweera, P.J. Smith, C. Yuen et al., Full-duplex relay selection for amplify-and-forward cooperative networks. IEEE Trans. Wirel. Commun. **11**(12), 4381–4393 (2012)
23. K. Yang, H. Cui, L. Song, Y. Li, Efficient full-duplex relaying with joint antenna-relay selection and self-interference suppression. IEEE Trans. Wirel. Commun. **14**, 3991–4005 (2015)
24. H. Cui, M. Ma, L. Song, B. Jiao, Relay selection for two-way full duplex relay networks with amplify-and-forward protocol. IEEE Trans. Wirel. Commun. **13**(7), 3768–3777 (2014)
25. R. Narasimhan, A. Ekbal, J.M. Cioffi, Finite-SNR diversity-multiplexing tradeoff of space-time codes, in *2005 IEEE International Conference on Communications, 2005. ICC 2005*, vol. 1 (IEEE, New York, 2005), pp. 458–462
26. Y. Hua, P. Liang, Y. Ma, A.C. Cirik, Q. Gao, A method for broadband full-duplex MIMO radio. IEEE Signal Process. Lett. **19**(12), 793–796 (2012)
27. D. Bharadia, E. McMilin, S. Katti, Full duplex radios, in *ACM SIGCOMM Computer Communication Review*, vol. 43 (ACM, New York, 2013), pp. 375–386
28. R.H. Louie, Y. Li, B. Vucetic, Practical physical layer network coding for two-way relay channels: performance analysis and comparison. IEEE Trans. Wirel. Commun. **9**(2), 764–777 (2010)
29. J. Mitola III, G.Q. Maguire Jr., Cognitive radio: making software radios more personal. IEEE Pers. Commun. **6**(4), 13–18 (1999)
30. J. Mitola, Cognitive radio - an integrated agent architecture for software defined radio, 2000
31. T. Yücek, H. Arslan, A survey of spectrum sensing algorithms for cognitive radio applications. IEEE Commun. Surv. Tutorials **11**(1), 116–130 (2009)
32. Y. Liao, T. Wang, L. Song, Z. Han, Listen-and-talk: full-duplex cognitive radio networks, in *IEEE Global Communications Conference (GLOBECOM), 2014* (IEEE, New York, 2014), pp. 3068–3073
33. Y. Liao, L. Song, Z. Han, Y. Li, Full duplex cognitive radio: a new design paradigm for enhancing spectrum usage. IEEE Commun. Mag. **53**(5), 138–145 (2015)
34. S. Huang, X. Liu, Z. Ding, Opportunistic spectrum access in cognitive radio networks, in *IEEE INFOCOM 2008. The 27th Conference on Computer Communications* (IEEE, New York, 2008)

# Device-to-Device Communications over 5G Systems: Standardization, Challenges and Open Issues

**G. Araniti, A. Raschellà, A. Orsino, L. Militano, and M. Condoluci**

**Abstract** This chapter deals with a comprehensive analysis of the challenges behind the introduction of Device-to-Device (D2D) communications in the context of 5th Generation (5G) communications. In fact, although D2D communications have emerged as an efficient solution for local traffic between mobile User Equipments (UEs) in proximity in cellular environments, has recently also attracted attention as a key enabling technology for 5G wireless networks. In details, this chapter addresses an overview of the current standardization, its integration in the cellular system architecture, future challenges and open issues. This chapter also provides a performance analysis, which illustrates the important achievements in terms of data rate in a scenario where the proximity communications between devices in an LTE-A system has been introduced for multicast downloading services.

## 1 Introduction

Recently, telco operators have been facing an increasing demand for high data rate as a consequence of a proliferation of innovative applications and services emerging in daily routines of mobile users. In this context, Device-to-Device (D2D) communications have attracted significant attention as a key enabling technology of 5th generation (5G) wireless networks [1] as witnessed by the recent Long Term Evolution (LTE) release 12 [2] from 3rd Generation Partnership Project (3GPP). Direct D2D communications are indeed a promising technological solution for local traffic between mobile User Equipments (UEs) in proximity in cellular

G. Araniti (✉) • A. Orsino • L. Militano
University Mediterranea of Reggio Calabria, Reggio Calabria, Italy
e-mail: araniti@unirc.it; antonino.orsino@unirc.it; leonardo.militano@unirc.it

A. Raschellà
Liverpool John Moores University, Liverpool, UK
e-mail: a.raschella@ljmu.ac.uk

M. Condoluci
King's College London, London, UK
e-mail: massimo.condoluci@kcl.ac.uk



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_12

environments. In particular, devices being close to each other can activate direct links and bypass the base station (BS) or access point (AP) by either using cellular communications resources or using alternative radio technologies such as Wi-Fi.

The conventional cellular communications paradigm involving the BS remains still appropriate when users are not usually close enough to make possible direct communications and for traditional low data rate mobile services such as voice calls and text messages. However, this communication scheme can become inadequate with the modern high data rate services (e.g., video conferencing, interactive gaming, and mobile social networking). When the UEs are in range for D2D communications, this innovative communication paradigm is undoubtedly an advantageous solution to be considered to optimize the spectral efficiency of the network. Noteworthy, the advantages of D2D communications are not only limited to an improved spectral efficiency. In fact, also link coverage, throughput, energy consumption can benefit from this communication technology. Moreover, robust public safety communications can be supported where devices provide at least local connectivity in case of damage to the radio infrastructure.

D2D communications can be enabled through either cellular network transmitting on cellular spectrum, known as *inband communication*, or existing technologies that consider unlicensed bands such as Wi-Fi access points, known as *outband communication*. Concerning the inband approach, the cellular spectrum may be shared between D2D and cellular communications, which is knows as the *underlay inband D2D mode* [3]. In this case a main field of investigation is the mitigation of the interference between D2D and cellular communication [4]. As alternative, other works such as [5] propose to assign a portion of the cellular resources to D2D communications to avoid interference problems, which is known as *overlay inband D2D mode*. With this approach, resource allocation becomes the most important aspect to be investigated in order to avoid wasting precious spectrum resources [6]. The coordination between radio interfaces is either controlled by the BS (i.e, *controlled outband mode*) or by the users (i.e., *autonomous outband mode*). However, in both outband modalities baseline rules are needed to coordinate the communication over two different bands where also a second radio interface is involved (e.g., Wi-Fi Direct). Hence, the studies on outband D2D involve aspects such as power consumption and inter-technology architectural design [7–9].

Despite of the promising features of the D2D communications, there are some important challenges and issues still waiting for a solution before a widespread use of D2D communications in next 5G systems is granted. These include: (1) device discovery procedures with the aim to detect the presence of other UEs in the neighbourhood; (2) link setup strategies in order to properly select the spectrum to be used in the D2D radio links between interested UEs; (3) interference avoidance mechanisms that make possible the coexistence among D2D UEs with cellular network; (4) new devices design. Moreover, some research works has been conducted on as mobility management [10], multicast and broadcast communications and radio access procedures for D2D received little attention so far.

The interest for D2D communications as a key technology in the context of 5G mobile systems is witnessed by the initiatives of the mobile and wireless

communications enablers for the twenty-twenty information society (e.g., METIS European project [11], 5G-PPP association [12], Networld2020 platform [13]). The main objective of these new initiatives is to handle the foundation of 5G, the next generation mobile and the wireless communications systems. In particular, for what concerns the D2D communications, the METIS project is currently evaluating its role in various scenarios such as vehicle-to-vehicle communications, national security and public safety, cellular network offloading, or service advertisement.

In the remaining of the chapter a general overview of D2D communications will be provided firstly introducing its possible role in the context of 5G systems. Then, an overview of the standardization, the system architecture, and some examples of scenarios and use cases will be given. Particular attention will then devoted to D2D communications over cellular networks and possible services exploiting this novel technology. Then a particular scenario will be presented, showing the potentialities offered by this technologies in enhancing the main performance features for the UEs in a downlink multicast service. In the last part of the chapter challenges and open issues, such as mobility and full-duplex communications will be discussed.

## 2 The Role of D2D Communications in Emerging 5G Systems

The D2D communications technology has been addressed in 3GPP LTE release 12 system [14]; notwithstanding, it is expected to have a complete standardization of proximity services in next 3GPP releases 13 and 14 [15]. As mentioned in the introduction to this chapter, the exploitation of D2D communications between UEs in proximity is expected to achieve improvements in terms of spectrum utilization, overall throughput, energy consumption, and to guarantee better public safety networks management. In what is presented next, a general overview of the current D2D standardization process is provided together with the system architecture proposed to integrate this new technology in the current cellular systems, and a number of possible applications in different scenarios and use cases.

### *2.1 Standardization Overview*

The standardization process is an aspect of utmost importance to be considered for the commercial feasibility and future deployment of new technologies. In the particular case of D2D communications, although direct communications are already provided by the use of unlicensed Industrial, Scientific and Medical (ISM) bands (e.g., Wi-Fi Direct), its standardization in the context of the cellular system is currently still ongoing. A first example of the introduction of D2D communications into the LTE-Advanced (LTE-A) network is provided by Qualcomm Company, which developed a mobile communication system called

FlashLinq [16]. In particular, FlashLinq is a PHY/MAC network architecture, which allows cellular devices automatically and continuously discovering thousands of other FlashLinq enabled devices within 1 km and communicating peer-to-peer, at broadband speeds and without the need of intermediary infrastructures. Moreover, peer-to-peer communications enabled through Qualcomm's FlashLinq can share connectivity with a cellular network technology unlike Wi-Fi Direct's-based peer-to-peer. FlashLinq discovery procedure is carried out by broadcasting public/private expressions mapped into tiny 128-bit packages of data, which represent basic information of either devices or users.

From a standardization point of view, 3GPP is focusing its efforts on D2D communications (recently begun in release 12 [14]) for public safety Proximity Services (ProSe) [17]. This strategy has been initially targeted to allow LTE becoming a competitive broadband communication technology for public safety networks used by first responders. However, from a technical perspective point of view, the exploitation of the proximity nature of the communicating devices will provide the further performance benefits: (1) D2D UEs will be able to exploit high data rate with a low delay due to the short range; (2) compared to traditional downlink/uplink cellular communication, D2D will enable energy savings and improve radio resource utilization; (3) cellular data traffic offloading and, consequently, lower overload in the network. In detail, the 3GPP Radio Access Network (RAN) working group has proposed in TR 36.843 Rel. 12 [17] two basic functions for supporting *ProSe discovery* and *ProSe communications* over the LTE radio interface. ProSe discovery allows an UE using the LTE air interface to identify other UEs in proximity. Two kinds of ProSe discovery exist, namely *restricted* and *open*; the difference consists in whether the permission is necessary or not for the discovery for a UE. ProSe communication instead, is the data communication between two UEs in proximity using the LTE air interface. 3GPP Services working group (SA1) has defined in specification TR 22.803 [18] the use cases and scenarios for ProSe. In the document, conditions for service flows and potential requirements for different use cases are analyzed in order to provide a support for D2D systems design. Some examples of use cases and scenarios identified for general commercial/social use and network offloading are summarized below.

The following terms are defined by 3GPP in the description of D2D use cases:

- *ProSe Discovery*: it is a process that identifies a UE in proximity of another, using EvolvedUMTS Terrestrial Radio Access Network (E-UTRAN).
- *ProSe Communication*: it is a communication between two UEs in proximity through an E-UTRAN communication path established between the UEs. The communication path can for example be established directly between the UEs or routed via local evolved-NodeB (eNB).
- *ProSe-enabled UE*: it is a UE that supports ProSe Discovery and/or ProSe Communication.
- *LTE D2D*: it is a series of technologies characterized by ProSe capability.

## 2.2 Uses Cases and Scenarios Presented in 3GPP Rel. 12

Some examples of use cases for ProSe Discovery and ProSe Communication scenarios defined by 3GPP SA1 in specification TR 22.803 [17] are given below.

**Restricted/Open ProSe Discovery:** these are use cases for a basic ProSe Discovery scenario that can be exploited for any kind of application. In case of restricted ProSe Discovery, a ProSe-enabled UE discovers another UE in proximity only if it has previously achieved the permission; while, in case of open ProSe Discovery, a ProSe-enabled UE is able to discover neighbor devices without the necessity of a permission. An example of restricted use case is the friend discovery in a social network where the discovery is constrained by the UE's privacy settings. While a shop/restaurant advertisement is an example of open use case because shops and restaurants are open to be discovered by all the possible ProSe-enabled UEs in proximity, being free of privacy issues.

**Network ProSe Discovery:** it is a use case for ProSe Discovery scenarios where the Mobile Network Operator (MNO) verifies if a UE has the permission to discover another UE and the proximity. Therefore, in this case the network should be able to determine and provide the ProSe-enabled UEs with their proximity.

**Service Continuity Between Infrastructure and E-UTRA ProSe Communication Paths:** this is a use case for a ProSe Communication scenario where the operator is able to switch user traffic from the initial infrastructure communication path to the ProSe communication one. Then, the traffic can be addressed again towards an infrastructure path, without being perceived by the users. Hence, the operator should be able to dynamically control the proximity criteria (e.g., range, channel conditions, achievable QoS) for switching between the two communication paths.

**ProSe-Assisted WLAN Direct Communications:** WLAN direct communication is a use case available between ProSe-enabled UEs with WLAN capability when they are in Wi-Fi Direct communications range. It is based on the ProSe Discovery and the WLAN configuration information from the 3GPP Evolved Packet Core (EPC). In this case the operator is able to switch data session between infrastructure path and WLAN ProSe communication path.

**ProSe Application Provided by the Third-Party Application Developer:** in this case the operator can provide ProSe capability features in a series of APIs to third-party application developers. Through this cooperation between the operator and third-party application developers, the user can download and use a wide variety of new ProSe applications created by third-party developers. In this case the operator's network and the ProSe enabled UE should provide a mechanism that enables to identify, authenticate and authorize the third-party application to use ProSe capability features.

In Table 1 the available specifications together with the corresponding main topics provided by the 3GPP working groups is summarized. It can be noticed the presence of the mentioned SA1 and RAN working group handling, respectively, feasibility study for ProSe and LTE radio interface issues. Other examples of topics

**Table 1** Available documents for D2D

| | | |
|---|---|---|
| SA1 | TR 22.803 | Feasibility study for Proximity Services (ProSe) |
| | TS 22.278 | Service requirements for the Evolved Packet System (EPS) |
| | TS 22.115 | Service aspects; changing and billing |
| | TS 21.905 | Vocabulary for 3GPP specifications |
| SA2 | TR 23.703 | Study on architecture enhancements to support Proximity-based Services (ProSe) |
| | TS 23.303 | Proximity-based Services (ProSe); Stage 2 |
| SA3 | TR 33.833 | Study on security issues to support Proximity Services |
| RAN 1 & RAN 2 | TR 36.843 | Study on LTE device to device proximity services—Radio Aspects |
| CT1 | TS 24.333 | Proximity-services management object (MO) |
| | TS 24.334 | Proximity-services (ProSe) user equipment (UE) to Proximity-services function aspects; Stage 3 |

under investigation supporting ProSe are the study of the architecture, security issues and Management Objects (MOs) representing parameters that handle the configuration of ProSe-enabled UEs.

## *2.3 System Architecture*

In order to support the scenarios illustrated earlier in this chapter, the enhancements in the LTE architecture illustrated in Fig. 1 have been proposed. In details, this architecture aims at meeting the following requirements introduced by the 3GPP specifications to:

- Allow the operator to control the ProSe discovery feature in its network and authorizing the functionalities required for the ProSe discovery of each UE.
- Allow the ProSe communication or ProSe-assisted WLAN Direct communication and seamless service continuity when switching user traffic between an infrastructure path and a ProSe communication of the ProSe-enabled UEs.
- Allow an authorized third party ProSe application interacting with the 3GPP network in order to use the ProSe services provided by the network.
- Be able to control ProSe communication between ProSe-enabled UEs in case the UEs are served by either the same or different eNBs.
- Handle the ProSe-related security functions that correspond to privacy, support for regulatory functions including Lawful Interception, and authentication upon ProSe discovery and ProSe communication.
- Allow the operator's authorization and authentication of the third-party applications before making use of the ProSe features.

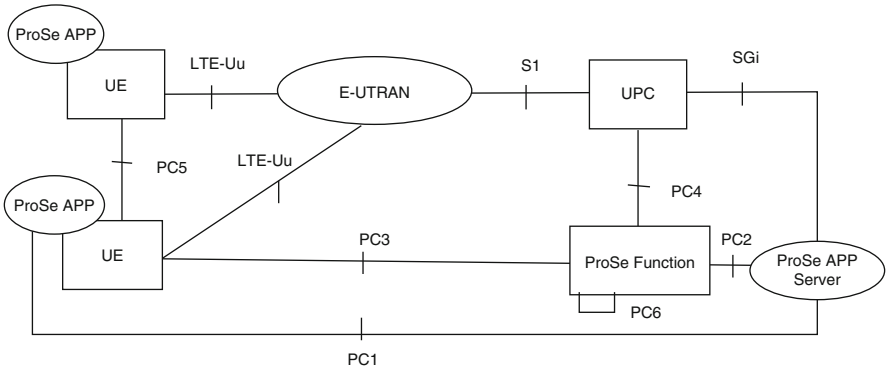


**Fig. 1** 3GPP D2D proposed architecture

As depicted in Fig. 1, in addition to the entities of the conventional LTE architecture (i.e., Evolved UMTS Terrestrial Radio Access Network, E-UTRAN and Evolved Packet Core, EPC), several new entities are required. These new entities are illustrated in the following.

- *Application servers (ProSe App Server)*: include the ProSe capability in order to build the application functionalities, such as specific agencies (PSAP) in case of Public Safety or social media in the commercial cases. These applications are proposed outside the 3GPP architecture; notwithstanding there may be reference points toward 3GPP entities. Moreover, the Application server can communicate with an application in the UE.
- *Applications in the UE (ProSe UEs App)*: exploit the ProSe capability in order to build the application functionality. An example may be the communication between members of Public Safety groups as well as social media application that requires finding other UEs in proximity.
- *ProSe Functions*: are a reference point toward the ProSe App Server, the EPC, and the UE. The functionalities may comprise (but not limited to): (1) internetworking via a reference point toward the third-party applications; (2) authorization and configuration of UEs for discovery and direct communication; (3) allowing the functionality of the EPC-level ProSe discovery, and charging. Notice that for the interconnection of the new entities and the connection with the conventional LTE ones, seven new interfaces/reference points are illustrated in the figure as PC1, PC2, PC3, PC4, PC5, PC6, and SGi (Fig. 1).

## 2.4 Application Scenarios

Applications of 5G D2D communications include local service, emergency communication, and the Internet of Things (IoT) enhancement. A brief description of these applications is provided in the following.

### 2.4.1 Local Service

In this scenario, user data is directly transmitted between terminals without being routed through the network side. Local service is usually utilized for social apps that are a basic D2D application based on the proximity feature. Through the D2D discovery and communication functions, a user can find other close users in order to share data or play games with them.

Another basic application of local service is the local data transmission, which exploits the proximity and direct data transmission characteristics of D2D to extend mobile applications while saving spectrum resources and then, making possible a new source of revenue for operators. In fact, local advertising service based on proximity can accurately target people in order to improve its benefits. Some examples of local transmissions conceived to improve commercial benefits are: a shopping mall where discounts and commercial promotions are sent to people walking into or around the mall; a cinema where information about movies and showtimes can be sent to people close by.

A third application of local service is the cellular traffic offloading that can reduce network overloading problems. In fact, consider that nowadays media services are becoming more and more popular; their massive traffic flows cause an extensive pressure on core networks and spectrum resources. In this context, D2D-based local media services allow operators to save spectrum resources in their core networks. In hotspot areas, operators or content providers can exploit media servers storing popular media services and sending them in D2D modality to the users. Alternatively, users can utilize D2D communications to obtain the media content from close terminals which have obtained media services. This enables to optimize the downlink transmission pressure of operator cellular networks. Furthermore, the cellular communication between short-distance users can be switched to the D2D modality in order to offload cellular traffic.

### 2.4.2 Emergency Communications

Natural disasters such as earthquakes can damage traditional communication network infrastructures making networks not available and causing enormous rescue efforts. This problem could be overcome through the introduction of D2D communications. In fact, although the communication network infrastructures may be irremediably affected, a wireless network can still be created between terminals

based on the D2D connections. This means that an ad hoc network can be set up based on multi-hop D2D to guarantee smooth wireless communication between users. Moreover, a wireless network affected by terrain or buildings can have blind spots. With single-hop or multi-hop D2D communication, users may be connected in the blind spots to other users, which are in coverage areas and then, be connected to the wireless network.

### 2.4.3 IoT Enhancement

One of the main aims of designing new mobile communication technologies is to create an extensive interconnection among different networks involving various types of terminals. This is the motivation, which has pushed forward the development of the Internet of Things (IoT) in the cellular communication framework. The industry forecast says that by 2020 there will be 50 billion cellular access terminals on a global scale and most of them will be devices with the IoT feature. In this context, the connection between D2D with IoT will drive towards a truly interconnected wireless network.

A common application of D2D-based IoT challenge is vehicle-to-vehicle (V2V) communication in the Internet of Vehicles (IoV). For instance, when a vehicle runs at high speeds, it can warn close vehicles in D2D mode before it changes lanes or slows down. According to the received messages, close vehicles warn drivers or even automatically handle the driving in an emergency situation; hence, thanks to this application drivers can react more quickly to diminish the number of traffic accidents. D2D communications provide inherent advantages when they are considered in the context of IoV security issues also thanks to their favorable features in terms of communication delay and neighbor discovery.

As there exist many IoT devices in a 5G network, access load is becoming a serious issue to be taken into account. Nevertheless, D2D-based network access is expected to improve this problem. In a scenario characterized by many terminals, low-cost terminals can access close special terminals in D2D modality instead of direct connections with BSs. Moreover, if multiple special terminals are isolated, the wireless resources for accessing low-cost terminals may be reutilized by these special terminals. Notice that this not only improves access pressure on BSs, but also optimizes the spectrum efficiency. Furthermore, the D2D-based access modality is more flexible and costs less than the small cell structure of the existing 4G networks.

In a smart home application, a smart terminal may be considered as a special terminal. Wireless appliances in the smart home access the smart terminal in D2D modality; while, the smart terminal may access the BS in a traditional cellular mode. The cellular-based D2D communication can represent a real breakthrough for the development of the smart home industry.

### 2.4.4 Other Applications

D2D communications may also be considered in other potential scenarios, such as multiuser MIMO enhancement, cooperative relaying, and virtual MIMO. In the context of the traditional multiuser MIMO, BSs find precoding weights based on the feedback received by the terminals in the respective channel in order to create nulls and delete interference between users. Through the introduction of D2D communications, paired users may directly exchange information about channel status. Hence, terminals can put together channel status information to be sent to the BSs improving the performance of multi-user MIMO.

D2D communications may also contribute to solve problems in new wireless communication scenarios. For instance, in the indoor positioning terminals may not achieve satellite signals if they are indoors. In this case, the traditional satellite-based positioning cannot work efficiently. In case of D2D-based indoor positioning, either pre-deployed terminals with given location information, or usual outdoor terminals with given position can detect the location of terminals to be localized, and support indoor positioning at a low cost in 5G networks.

## 3 State-of-the-art on D2D Communications Over Cellular Networks

D2D communications are expected to play a key role in the ecosystem of future 5G cellular networks. This is motivated by two aspects: (1) the amount of data traffic exchanged over radio mobile systems is exponentially increasing and this dictates novel communications paradigms for radio mobile networks; (2) use cases for D2D communications presented above represent key 5G services. As a consequence, the natively support of D2D communications becomes crucial in 5G systems.

D2D communication was initially proposed in cellular networks as a new paradigm to enhance network performance. Several studies in the literature have already discussed the improvements in terms of spectral efficiency and reduced communication delay that D2D communication can provide in cellular networks [19–24]. On the other hand, this new paradigm presents several aspects to be investigated in terms for instance of interference control overhead and network protocols. Therefore, the feasibility of D2D communications in the context of LTE-A is currently a fascinating topic under investigation by academia, industry, and the standardization bodies. A general overview of state-of-the-art applications based on D2D communications for future 5G wireless systems is given next in both, uplink and downlink scenarios. Then, some examples of services where D2D communications have been efficiently exploited in LTE-A networks will be illustrated and assessed through exhaustive performance evaluation.

Several studies addressing D2D communications for downlink services can be found in the literature, covering several aspects and applications as for instance

mobile data offloading [25], cell coverage extension [26] or content sharing [27, 28]. Recently, D2D communications have been taken into account also for downloading multicast services with focus on direct device communications over short links of a different technology than the cellular one. To cite some of them, in [29] a subset of mobile devices are considered as anchor points in a cell to forward the multicast data received from the BS to other devices in proximity through multihop ad-hoc Wi-Fi links. In [30] cellular users directly communicate to carry out cooperative retransmissions using generic short-range communication capabilities. However, the use of heterogeneous wireless interfaces introduces several issues in terms of content synchronization that become essential in case of multicast video streaming applications. Moreover, as also stated in [31], the use of cellular D2D links provides several benefits compared to *outband* D2D links, like Wi-Fi, in terms of improved user throughput. Although, the focus of the literature has been mainly on technical issues for downlink services, uplink direction scenarios are of undoubted interest as also witnessed by recent publications, such as [32] where relaying on smartphones is proposed to transmit emergency messages from disconnected areas. Multihop D2D communications have been also investigated in a very few recent works. In [33, 34] network-assisted D2D communication is addressed with an analysis on power control and mode selection on the direct links. However, the analysis refers to a more traditional two-hop scenario, with a UE or the eNodeB as the last hop node. Similarly, multihop D2D communication is considered in [35, 36] for end-to-end Machine-to-Machine and human-traffic connectivity.

As an example of D2D communications over cellular LTE-A links a downlink scenario for multicast transmission is considered in order to efficiently overcome the limitations identified in [29, 30]. In details, in the scenario proposed in this study a portion of multicast users, which sense poor channel qualities is split into clusters. The members of these clusters are served through cellular D2D transmissions, while the remaining users (i.e., those with better channel quality) are served over cellular transmission from the BS.

All the solutions illustrated in this section exploit D2D communications relying on LTE-A network infrastructure. In LTE-A, Orthogonal Frequency Division Multiple Access (OFDMA) and Single Carrier Frequency Division Multiple Access (SC-FDMA) are considered, respectively, in case of downlink and uplink. The eNodeB handles the spectrum resources by providing the appropriate number of RBs to each scheduled user and by selecting the Modulation and Coding Scheme (MCS) for each RB. Scheduling solutions are based on the Channel Quality Indicator (CQI) feedback, which is sent by a UE to the eNodeB over dedicated control channels. Each CQI value correspond to a given maximum supported MCS as specified in [37]. The MCS parameters can be adapted at every CQI Feedback Cycle (CFC), which can last one or several Transmission Time Intervals (TTIs) where one TTI is 1 ms.

## *3.1 A Case Study: D2D-Enhanced Conventional Multicast Scheme for Video Delivery*

Nowadays, several research and industrial organizations are considering solutions to best manage the increased traffic and the dissimilar channel quality sensed by users in the same multicast group and efficiently handle the available resources [38]. A basic approach to this issues is the Conventional Multicast Scheme (CMS) [38], which serves all multicast users in a cell at every TTI, by constraining the data rate to the user with the worst channel conditions. However, this approach translates into poor performance in terms of data rate and low satisfaction levels for users with good channel situations. An alternative solution is the Opportunistic Multicast Scheme (OMS) [39], which only serves users with the best channel conditions in each TTI. This enables to enhance the network data rate, but short-term fairness is no more assured to users and this could affect the delivery of time sensitive services such as video to some users. In this study it is demonstrated how multicast delivery strategies in LTE-A systems can benefit from D2D communications. For instance, neighboring UEs that belong to the same multicast group can exploit direct links by utilizing cellular radio resources [21, 40], to cope with adverse cell-edge effects. This study exploits the D2D communications in the multicast communication framework to complement CMS in order to address some of the limitations of the latter. In details, it is proposed a D2D-enhanced Conventional Multicast Scheme (CMS) with Single Frequency (D2D-SF) by combining in an innovative way CMS-based and D2D content delivery in order to improve the aggregate data rate of the cell, while also keeping the short-term fairness between devices. D2D-SF solution makes use of one or more mobile subscriber devices as forwarding devices (FD) to retransmit the data received from the BS over direct local links to other members of the multicast group. It is considered that the D2D links exploit uplink frequencies, as suggested in [22] and all the FDs in the same cell simultaneously use *the same frequency* to transmit multicast data over the D2D links, as illustrated in [41, 42]. The receivers take into account these retransmissions as multipath components of the same signal. D2D-SF is defined in conjunction with a policy for selecting FDs based on the devices clustering. Simulation results demonstrate how significant enhancements of the overall performance when delivering video content can be achieved compared to both state-of-the-art multicast strategies and a D2D novel solution that does not consider the single-frequency paradigm.

### 3.1.1 System Model and Problem Formulation

The LTE/MBMS architecture [42] is composed as follows (i.e., please refers to Fig. 2 to have a general overview). The access network is made of the eNodeB and the MultiCell/Multicast Coordination Entity (MCE), in charge of transmission parameters configuration in single- and multi-cell mode, respectively. The core network involves: Mobility Management Entity (MME) that is in charge for

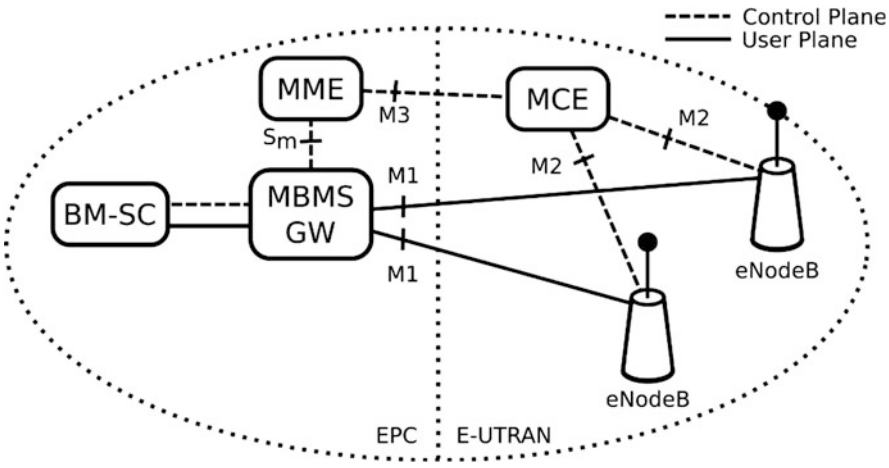


**Fig. 2** LTE/MBMS architecture

authentication, security, and mobility management procedures; MBMS Gateway (MBMS-GW), a logical entity whose principal function is data packets forwarding to eNodeBs; Broadcast Multicast-Service Center (BM-SC), which is the MBMS traffic source accomplishing also service announcement and group membership functions. MBMS entities rely on the LTE-A system illustrated in Sect. 3 considering TDD, by referring to the frame structure *type 2* foreseen by 3GPP [41] and *configuration 1* which guarantees an equal number of downlink and uplink slots over the frame.

Let us consider a group of UEs that is interested in the same multicast content provided by a single LTE-A cell. Under this condition a multicast video delivery can be provided exploiting the enhancements offered by D2D communications among the involved devices. In the considered LTE-A cell, a set of users, denoted by $\mathscr{K}$, is associated to the same multicast group (MG). The eNodeB performs link adaptation procedures on both cellular and D2D links by managing $N$ available RBs according to the CQI feedbacks collected from each user. Let $C$ be the number of available CQI levels and let $c_k \in \{1, 2, \ldots, C\}$ be the CQI reported by multicast member $k$, with $k \in \mathscr{K}$. Furthermore, let $c_{k,j}$ be the CQI value for each D2D link between nodes $k, j \in \mathscr{K}$, $k \neq j$. For a given MCS value $m$, the available data rate depends on the number of assigned RBs and on the spectral efficiency for the given MCS, $b_m$ expressed in bit/s/Hz as explained in [31]. Hence, the spectral efficiency is denoted with $b_m^{dl}$ and $b_m^{ul}$ (where $m = 1, \ldots, C$ ) respectively in downlink and uplink transmissions. Moreover, $f^{dl}(m, n_m)$ and $f^{ul}(m, n_m)$ represent the data rate respectively in downlink and uplink transmissions adopting the MCS associated to the CQI $m$, as a function of $m$ and the assigned RBs. The proposed radio resource management (RRM) scheme is responsible of deciding which multicast configuration to allow, by this meaning: (1) the set of UEs directly served by the eNodeB in downlink, (2) the MCS for the downlink transmission, (3) the cluster

configuration for D2D relaying, and (4) the resource allocation and the MCS assignment for the transmissions of each activated FD. The eNodeB carries out the following steps when the service delivery starts. In details, a single execution of the listed steps is carried out. However, when significant variations in the channel conditions are sensed, these steps should be repeated to update the service configuration.

1. *Service registration*: The eNodeB advertises the multicast service and all interested UEs within the cell join this service to create a single MG.
2. *CQI collection*: The eNodeB collects the CQI feedbacks from all UEs belonging to the MG, i.e., $c_k \ \forall k \in \mathscr{K}$.
3. *D2D CQI collection*: The eNodeB collects the $c_{k,j}$ values from all UEs $k, j \in \mathscr{K}$, $k \neq j$ belonging to the MG. This information will be utilized to discover the UEs that can be reachable through D2D links by selected FDs in the MG.
4. *FD selection and cluster formation*: Being $\tilde{\mathscr{C}} \subseteq \{1, 2, \ldots, C\}$ the set of CQI levels in downlink for the UEs in the MG, for each $m \in \tilde{\mathscr{C}}$ the eNodeB computes: (1) the set of UEs that can correctly decode data if served by the BS, i.e., $\mathscr{K}_m^{dl} = \{k \in \mathscr{K} | c_k \geq m\}$; *(ii)* the subset of served UEs $\mathscr{R}_m \subseteq \mathscr{K}_m^{dl}$, that can act as FDs; (3) the remaining UEs that are not served by the eNodeB, but can be served by a FD through D2D connections. Hence, the eNodeB determines a D2D CQI matrix (DCM) based on the $c_{k,j}$ values (where $k \in \mathscr{K}_m^{dl}$ and $j \in \mathscr{K} \setminus \mathscr{K}_m^{dl}$) for all the links between the potential FDs (the matrix rows) and the remaining nodes (the DCM columns). A $c_{k,j} = 0$ value in the DCM represents that a D2D link cannot be activated between nodes $k$ and $j$. Following the values in the DCM, the eNodeB will select the subset of UEs $\mathscr{D}_{m,r} \subseteq \mathscr{K} \setminus \mathscr{K}_m^{dl}$ to be associated to each allowed FD $r \in \mathscr{R}_m$.
5. *D2D link configuration*: For each CQI level $m \in \tilde{\mathscr{C}}$ evaluated for downlink transmissions, the eNodeB determines the resource $N_{m,r}^{ul}$ and the MCS level $l_{m,r}$, to be utilized on the D2D link for each FD $r \in \mathscr{R}_m$. D2D links can be either unicast or multicast. A conservative approach is considered in the multicast case; hence, the FD serves all UEs in the D2D cluster in a single transmission by using the MCS corresponding to the worst CQI value in the DCM, i.e., $l_{m,r} = \min_{k \in \mathscr{D}_{m,r}} \{c_{r,k}\}$ for FD $r$. This study takes into account two alternative policies according to which the FDs manage the uplink frequencies to send data in their own D2D cluster. The first policy associates different resources to the different FDs; the second one implements the novel single-frequency-based D2D paradigm, i.e., all the FDs considers the same RBs. In the former case, disjoint sets of RBs are assigned to the D2D links. In the latter case the amount of resources assigned to the D2D links are constrained by the cluster with the lowest activated MCS. In general, devices connected on a D2D link are expected to be at a short distance and with good channel conditions, thus they need a lower amount of resources compared to those needed for a direct cellular communication.
6. *Multicast service activation and resource allocation*: Finally, the eNodeB chooses the solution to activate, which is the one that optimizes the system data rate under the constraint that all the UEs in a MG are served, either through

direct cellular links or through D2D links. In details, after the selection of the MCS level $m^*$ to activate in downlink and of the corresponding $\mathscr{K}_{m^*}^{dl}$, $\mathscr{R}_{m^*}$, $\mathscr{D}_{m^*,r}$, $N_{m^*,r}^{ul}$, and $l_{m^*,r}$ values, the eNodeB assigns the available resources.

All values $m \in \tilde{\mathscr{C}}$ are potential CQI levels to activate in downlink. For each of the CQI levels a cluster formation algorithm is implemented to provide a *configuration* of FDs and corresponding D2D clusters. If a given tested level is *eligible*, then the corresponding data rate $\Omega_m$ is determined. A cluster configuration is taken into account *eligible* if the FDs are able to forward the total amount of bits received from the eNodeB over the D2D links to all users not served by the cellular link. This needs the following conditions: (1) the enabled FDs can successfully serve all the nodes belonging to $\mathscr{K} \setminus \mathscr{K}_m^{dl}$ via D2D links, and (2) the $N$ available resources are enough to relay all data to the D2D receivers. If instead, no cluster configuration for the tested CQI level $m$ can be found, thus the iteration on the $m \in \tilde{\mathscr{C}}$ value is stopped and the final selection is carried out. In details, the iteration can be stopped as the tested CQI levels follow an order from the minimum to the maximum CQI value, and with higher values for the CQI level in downlink the probability of getting an *eligible* configuration is reduced. Fundamental steps in the implementation of the proposed RRM previously discussed, are the FD selection and cluster formation and the D2D link configuration with the radio resource allocation. Concerning the first strategy, let us consider the generic iteration where the $m$-th CQI level is tested for downlink transmission. Given $\mathscr{K}_m^{dl}$, the set of UEs that can correctly decode the data according to the considered CQI, and based on the DCM, the eNodeB determines which nodes can potentially act as FDs for the remaining $\mathscr{K} \setminus \mathscr{K}_m^{dl}$ nodes. Based on this information the eNodeB can assign the resources to each D2D link. Let us focus now on the BSC policy, which is based on the idea that the eNodeB chooses "the best" FD for each UE not served in downlink. In particular, the best FD for each node $j$ belonging to $\mathscr{K} \setminus \mathscr{K}_m^{dl}$ is considered as the node $r \in \mathscr{K}_m^{dl}$ which assures the best D2D link conditions. In those cases where more than one FD can assure the same CQIs, the eNodeB selects the FD serving more users in order to limit the number of FDs.

### 3.1.2 Performance Results

Main simulation parameters are listed in Table 2. $R = 100$ RBs are available in the LTE system on a 20 MHz channel bandwidth. Channel conditions for the UEs are evaluated in terms of signal to interference and noise ratio (SINR) experienced over each sub-carrier [43] when path loss and fading phenomena affect the signal reception. The effective SINR is mapped onto the CQI level that guarantees a block error rate (BLER) smaller than 1 % [43, 44].

The following metrics have been considered to assess the performance of the proposed solutions with respect to CMS [38] and OMS [39]:

- *mean data rate* is the mean data rate value experienced by the multicast members;

**Table 2** Main simulation parameters

| Parameter | Value |
|---|---|
| Cell radius | 500 m |
| Frame Structure | Type 2 (TDD) |
| TTI | 1 ms (11 OFDM data symbols plus 3 control symbols) |
| Cyclic prefix/Useful signal frame length | 16.67 μs / 66.67 μs |
| TDD configuration | 1 |
| Carrier Frequency | 2.5 GHz |
| eNodeB Tx power | 46 dBm |
| D2D node Tx power | 23 dBm |
| Noise power | −174 dBm/Hz |
| Path loss (cell link) | 128.1 + 37.6 log(d), d[km] |
| Path loss (D2D link, NLOS) | 40 log(d) + 30 log(f) + 49, d[km], f[Hz] |
| Path loss (D2D link, LOS) | 16.9 log(d) + 20 log (f/5) + 46.8, d[m], f[GHz] |
| Shadowing standard deviation | 10 dB (cell mode); 12 dB (D2D mode) |
| RB size | 12 sub-carriers, 0.5 ms |
| Sub-carrier spacing | 15 kHz |
| BLER target | 1 % [44] |

- *aggregate data rate (ADR)* is the sum of the data rates experienced by the multicast users;

The performance analysis in terms of the metrics indicated focuses on the video streaming towards multicast users lasting 1 s; within this interval the BS adapts the transmission parameters every scheduling frame (i.e., 10 ms) [45]. In particular, for this analysis, two different study cases are considered:

- **Case A:** This case analyses the impact that the channel bandwidth has on the considered strategies. In this case the multicast group size $|\mathscr{K}|$ is set to 200, whereas a variable number of resources $R$ (ranging from 10 to 100 RBs) is dedicated to the service. UEs are distributed within an area of 100 m × 100 m located near the cell-edge;
- **Case B:** This case analyses the impact a varying multicast group size has on the considered strategies. The number of available resources $R$ is set to 100 RBs, whereas the number of UEs $|\mathscr{K}|$ ranges from 20 to 200. The same cell-edge distribution of UEs as in case A is considered.

Video parameters are set in accordance to [46], where adaptive video coding [47] is considered to be performed at the BS. Video parameters are tuned such that the video stream has an average bit rate between 256 kbps and a maximum value, which depends on the channel quality experienced by multicast users.

**Case A** The results achieved are plotted in Fig. 3. As expected, both mean data rate and ADR improve with the number of available RBs for all solutions. The proposed D2D-based strategies outperform CMS, with a better performance for *D2D-SF* . The

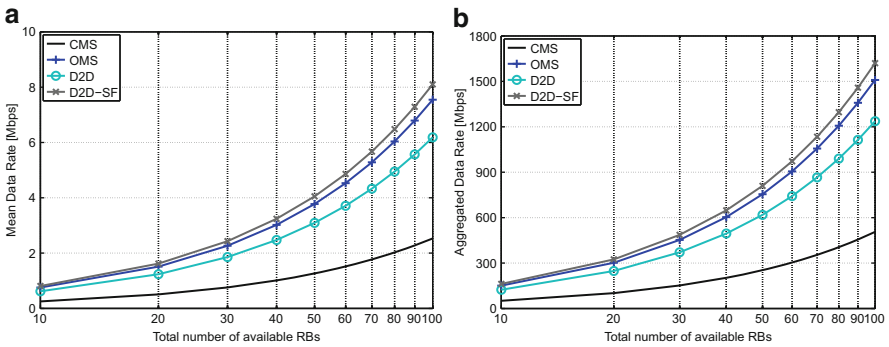


**Fig. 3** Performance analysis for video streaming analysis: study case A. (**a**) Mean UE data rate, (**b**) aggregate data rate

single-frequency solution also outperforms OMS. The benefit compared to CMS is constant with the number of available RBs and equal to 144 % for the *D2D* and *D2D-SF* solutions. The mean data rate for CMS is lower than *D2D* (left-hand in Fig. 3). The data rate of the OMS solution gets higher values than the *D2D* policy, but lower values compared to the solution based on the single-frequency paradigm. The price to pay when adopting OMS is the decrease in the number of served users and in short-term fairness. A similar behavior is achieved for ADR (right-hand in Fig. 3).

**Case B** The performance obtained by varying the multicast group size is shown in Fig. 4. Also in this case, the novel scheme considering the single-frequency paradigm, i.e., *D2D-SF*, outperforms the others. It is worth noting that for all the solutions, the mean data rate slightly diminishes when the number of users in the cell increases, as illustrated in left-hand plot in Fig. 4. In details, the CMS shows a performance reduction of about 29 % when changing from 20 to 200 UEs. When considering *D2D* there is a 25 % of reduction. This is an expected result as greater the number of users in the group, higher the risk of having users with very low channel conditions, which limit the overall performance. Also OMS senses a reduction in its offered mean data rate with the increase in number of UEs. Nevertheless, this phenomenon is less evident, as the mean data rate diminishes from 7.9 Mbps (with 20 UEs) to 7.5 Mbps (with 200 UEs). As expected, the right-hand in Fig. 4 illustrates how the ADR value increases with the number of multicast members in the cell. Furthermore, the gain introduced by *D2D* with respect to the CMS policy is larger when the number of multicast users increases.

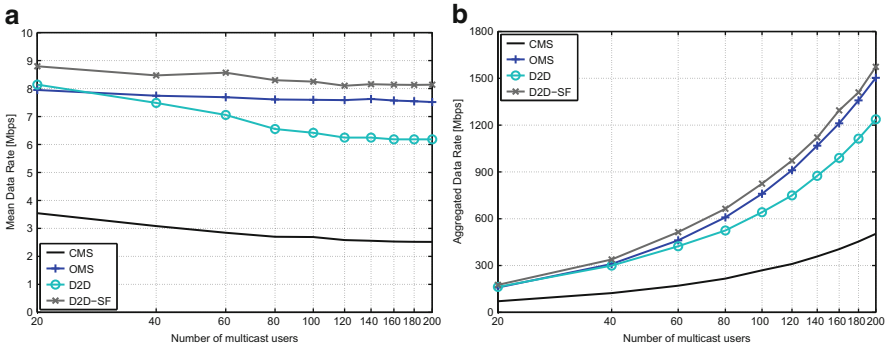


**Fig. 4** Performance analysis for video streaming analysis: study case B. (**a**) Mean UE data rate, (**b**) aggregate data rate

## 4 Challenges and Ongoing Research

Future 5G cellular networks are attracting the interest of the research community and engineers from the entire world. In particular, 5G technologies are expected to attain 1000 times higher mobile data volume per unit area, 10–100 times higher number of connecting devices and user data rate, 10 times longer battery life, and five times reduced latency [48]. However, 5G is currently without an official definition. Notwithstanding, the research community has already highlighted how the above objectives can be potentially reached through a multi-tier heterogeneous network architecture together with several technologies such as spatial modulation, millimeter wave (mmWave), visible light communication (VLC), and massive MIMO. In the following new challenges in the context of D2D communications as innovative technology supporting 5G will be illustrated.

### *4.1 Mobility*

The development of wireless networking technologies has brought enormous changes to people's everyday life. In particular, the vision of 5G networks empowers users to communicate through different wireless access technologies enabling new applications in modern communication paradigms, such as vehicular ad hoc networks (VANETs) and mobile cloud computing (MCC). These scenarios, however, introduce to one of the major challenging features related to D2D communications, namely user mobility. In particular, it affects network connectivity and the design of network protocols. On the other hand, another important aspect to be considered is the complex traffic flow related to a dynamic network topology and the double communication opportunities (i.e., infrastructure or infrastructureless wireless networks). The interest for D2D communications in Vehicle-to-Vehicle

(V2V) communication is related to the strict delay requirements in traffic safety use cases [49]. In particular, for collision avoidance systems it will be essential to have very low latency. Hence, it can obtain more complete and reliable information, which would be useful to avoid accidents. By using D2D communications rather than relying on infrastructure, it is possible to both diminish the latency and to develop a solution, which works without cellular network coverage so that a car can get more complete and reliable information to avoid accidents.

## *4.2 Heterogeneous Networks*

D2D communications may be also considered as an efficient solution for Multiuser cooperative communication (MUCC). For instance, a benefited user (BU) is in an area experiencing a poor cellular signal, but there is another user in the proximity sensing a good cellular signal. This user may be considered for helping the BU to enhance its experienced signal and then, acting as a so-called supporting user (SU). There exist two radio links for the BU: BU towards eNB (i.e., link directly) and SU directly towards picocells. Moreover, both BU and SU communicate with each other through LTE-A D2D communication. This kind of communication enhances the throughput because the system will always be able to schedule the best user experiencing the most favorable channel quality. There exist different channel/pipes and the probability that all channels are degraded at the same time is quite small, and this will improve reliability. Moreover, any single pipe improvement may be also considered at the same time (e.g., MIMO). Furthermore, MUCC improves the performance of this scheme through the context information (e.g., location information) of the users.

## *4.3 mmWave*

Millimeter wave (mmWave) communication operating in the 10–300 GHz band, appear to be an innovative candidate for next-generation cellular systems by which multiple gigabit-per-second data rates can be supported. On the other hand, enabling mmWave cellular systems in practice requires properly dealing with the channel impairments and propagation characteristics of the high frequency bands. In fact, the main propagation-related obstacles in realizing mmWave cellular are that free-space path loss is much larger in mmWave due to the higher carrier frequency, scattering is less significant which reduce the available diversity, and no line-of-sight (LOS) links are weaker making blockage and coverage holes more pronounced. Moreover, the noise power is larger due to the use of larger bandwidth channels. Furthermore, it is difficult for mmWave signals to penetrate through solid materials that confines outdoor mmWave signals to streets and other outdoor structures, although some signal power may get inside the buildings through glass windows and wood doors.

The mentioned propagation features give opportunities and advantages in allowing D2D communications over directional mmWave networks. In fact, mmWave can be considered in D2D-enabled wireless devices for direct short range communications amongst users or machines in near proximity. Moreover, D2D communication can also provide a path between two wireless devices and between wireless devices and mmWave BSs by relay if LOS paths are no available. Furthermore, mmWave communication with highly directional antennas gives wireless connections characterized by a high data rate for BS to BS (B2B) communications, in contrast with 4G cellular networks where fiber links communications between BSs are considered. Finally, interference management is one of the most important issue that may be enhanced through the exploitation of D2D communications. Due to the directional antennas and large propagation loss, mmWave communication has relatively low multi-user interference (MUI) that may support simultaneous communications over the same radio spectrum. Enabling multiple concurrent D2D links, the network capacity and spectrum efficiency may be efficiently enhanced. Hence, new network architectures and resource sharing strategies that will account for the directional interference sources are needed in mmWave 5G cellular networks to fully exploit the enhancements of both technologies.

### *4.4 Massive-MIMO*

Massive MIMO is a technology whose features can be efficiently exploited in the context of D2D communications in uplink scenarios of cellular networks. In details, in a massive MIMO system, each BS utilizes a very large antenna array to serve multiple users in each time-frequency resource block [50]. If the number of antennas at a BS is considerably greater than the number of served users, the channel of each user towards the BS is closely orthogonal to that of any other user. This aspect enables to process techniques for very simple transmit or receive such as matched filtering to be nearly optimal with enough antennas even in the presence of interference. This means that, a large antenna array at a BS may drive D2D communications towards a very low interference at the uplink massive MIMO BS, making D2D very simple and appealing in massive MIMO systems. However, although D2D-to-cellular interference can be effectively managed by the large antenna array at a BS, cellular interference provoked in D2D communications holds and can be worse in a massive MIMO system. In details, massive MIMO is a multi-user transmission strategy developed in order to support multiple users in each time-frequency block; the number of users simultaneously active in an uplink scenario is scalable with the number of antennas at the BS. Considering this increased number of uplink transmitters, the D2D connections reutilizing uplink radio resources will sense increased interference. In order to protect D2D connections against such increased interference, the number of simultaneously active uplink users might have to be limited.

### *4.5 Full-Duplex*

Duplex mode is a recent innovative research direction related to D2D communications development. In particular, when considering FDD, two receivers and two antennas are needed. The Tx antenna and Rx one should have enough isolation distance. It is always challenging implementing into UE with small physical size. While TDD has the advantage that uplink downlink transmissions/receptions are performed in different time slots and in the same frequency band. This is basically why most of D2D researchers are currently focusing on TDD modality [51]. It is predicted that in 5G D2D will be supported by both TDD and FDD. Notwithstanding, D2D TDD mode will be the principal type considered thanks to its lower cost, lower complexity, and the possible benefit of better frequency availability.

## 5 Conclusion

In this chapter an overview of the D2D-communications has been presented as one of the promising technological features for the definition of future 5G wireless networks. The details related to the standardization process, the possible applications and technological issues, have been browsed together with the important benefits such a technology can offer. In the second part of the chapter a practical application of D2D communications has been discussed, with reference to a scenario where the proximity communications between devices in an LTE-A system, introduces benefits to a multicast downloading service. Important improvements in terms of data rate increase for the involved users has been demonstrated through simulative results. Finally, in the last part of this chapter open challenges and research field directions have been introduced for the design of novel technological solutions integrating also mobility issues, mmWave communications, massive MIMO and Full-Duplex.

## References

1. X. Lin, J.G. Andrews, A. Ghosh, R. Ratasuk, An overview on 3GPP device-to-device proximity services. IEEE Commun. Mag. **52**(4), 40–48 (2014)
2. 3GPP, 3rd Generation Partnership Project; Technical Specification Group SA; Feasibility Study for Proximity Services (ProSe) (Release 12). TR 22.803 V1.0.0, August 2012
3. C.-H. Yu, K. Doppler, C. Ribeiro, O. Tirkkonen, Performance impact of fading interference to device-to-device communication underlaying cellular networks, in *Proceedings of IEEE PIMRC* (IEEE, Piscataway, 2009), pp. 858–862
4. L. Lei, Y. Zhang, X. Shen, C. Lin, Z. Zhong, Performance analysis of device-to-device communications with dynamic interference using stochastic petri nets. IEEE Trans. Wirel. Commun. **12**(12), 6121–6141 (2013)

5. Y. Pei, Y.-C. Liang, Resource allocation for device-to-device communication overlaying two-way cellular networks. IEEE Trans. Wirel. Commun. **12**(7), 3611–3621 (2013)
6. L. Lei, Y. Kuang, X. Shen, C. Lin, Z. Zhong, Resource control in network assisted device-to-device communications: solutions and challenges. IEEE Commun. Mag. **52**(6), 108–117 (2014)
7. A. Asadi, V. Mancuso, Energy efficient opportunistic uplink packet forwarding in hybrid wireless networks, in *Proceedings of the Fourth International Conference on Future Energy Systems* (2013), pp. 261–262
8. A. Asadi, V. Mancuso, On the compound impact of opportunistic scheduling and D2D communications in cellular networks, in *Proceedings of the 16th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM '13)* (ACM, New York, 2013), pp. 279–288
9. M. Condoluci, L. Militano, A. Orsino, J. Alonso-Zarate, G. Araniti, LTE-direct vs. WiFi-direct for machine-type communications over LTE-A systems, in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 30 August–2 September 2015, pp. 2298–2302
10. A. Orsino, M. Gapeyenko, L. Militano, D. Moltchanov, S. Andreev, Y. Koucheryavy, G. Araniti, Assisted handover based on device-to-device communications in 3GPP LTE systems, in *IEEE Globecom Workshop on Emerging Technologies for 5G Wireless Cellular Networks*, December 2015
11. EU Project METIS, https://www.metis2020.com
12. The 5G Infrastructure Public Private Partnership (5GPPP), https://5g-ppp.eu
13. NetWorld2020, http://networld2020.eu
14. D. Astely, E. Dahlman, G. Fodor, S. Parkvall, J. Sachs, LTE release 12 and beyond [Accepted From Open Call]. IEEE Commun. Mag. **51**(7), 154–160 (2013)
15. Ericsson white paper: LTE Release 13, http://www.ericsson.com/res/docs/whitepapers/150417-wp-lte-release-13.pdf
16. W. Xinzhou, S. Tavildar, S. Shakkottai, T. Richardson, L. Junyi, R. Laroia, A. Jovicic, FlashLinQ: a synchronous distributed scheduler for peer-to-peer ad hoc networks. IEEE/ACM Trans. Networking **21**(4), 1215–1228 (2013)
17. 3GPP TR 36.843, Study on LTE device to device proximity services; Radio aspects, v12.0.1, March 2014
18. 3GPP TR 22.803, Feasibility study for proximity services (ProSe), v12.2.0, June 2013
19. IEEE Standard 802.16e-2005, Air interface for fixed and mobile broadband wireless access systems amendment for physical and medium access control layers for combined fixed and mobile operation in licensed band (2005)
20. B. Kaufman, B. Aazhang, Cellular networks with an overlaid device to device network, in *Proceedings of Asilomar Conference on Signals, Systems and Computers* (2008), pp. 1537–1541
21. K. Doppler, M. Rinne, C. Wijting, C. Ribeiro, K. Hugl, Device-to-device communication as an underlay to LTE-advanced networks, *IEEE Commun. Mag.* **47**(12), 42–49 (2009)
22. K. Doppler, M.P. Rinne, P. Janis, C. Ribeiro, K. Hugl, Device-to-device communications; functional prospects for LTE-Advanced networks, in *Proceedings of IEEE ICC Workshops* (2009), pp. 1–6
23. A. Osseiran, K. Doppler, C. Ribeiro, M. Xiao, M. Skoglund, J. Manssour, Advances in device-to-device communications and network coding for IMT-Advanced, in *ICT Mobile Summit* (2009)
24. T. Peng, Q. Lu, H. Wang, S. Xu, W. Wang, Interference avoidance mechanisms in the hybrid cellular and device-to-device systems, in *Proceedings of IEEE PIMRC* (2009), pp. 617–621
25. S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, Y. Koucheryavy, Analyzing assisted offloading of cellular user sessions onto d2d links in unlicensed bands. IEEE J. Sel. Areas Commun. **33**, 67–80 (2015)
26. G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, Z. Turanyi, Design aspects of network assisted device-to-device communications. IEEE Commun. Mag. **50**, 170–177 (2012)

27. J. Seo, T. Kwon, V. Leung, Social groupcasting algorithm for wireless cellular multicast services. IEEE Commun. Lett. **17**, 47–50 (2013)
28. L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, G.-M. Muntean, Single frequency-based device-to-device enhanced video delivery for evolved multimedia broadcast and multicast services. IEEE Trans. Broadcast. **61**(2), 263–278 (2015)
29. S.C. Spinella, G. Araniti, A. Iera, A. Molinaro, Integration of ad-hoc networks with infrastructured systems for multicast services provisioning, in *International Conference on Ultra Modern Telecommunications and Workshops, 2009. ICUMT '09*, 12–14 October 2009, pp. 1–6
30. Q. Zhang, F.H.P. Fitzek, V.B. Iversen, Design and performance evaluation of cooperative retransmission scheme for reliable multicast services in cellular controlled P2P networks, in *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, (PIMRC)*, Athens, Greece, September 2007
31. L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, F.H.P. Fitzek, Wi-Fi cooperation or D2D-based multicast content distribution in LTE-A: a comparative analysis, in *IEEE International Conference on Communications Workshops (ICC)*, June 2014, pp. 296–301
32. H. Nishiyama, M. Ito, N. Kato, Relay-by-smartphone: realizing multihop device-to-device communications. IEEE Commun. Mag. **52**(4), 56–65 (2014)
33. J. da Silva, G. Fodor, T. Maciel, Performance analysis of network-assisted two-hop d2d communications, in *Globecom Workshops (GC Wkshps)*, December 2014, pp. 1050–1056
34. L. Lei, X. Shen, M. Dohler, C. Lin, Z. Zhong, Queuing models with applications to mode selection in device-to-device communications underlaying cellular networks. IEEE Trans. Wirel. Commun. **13**(12), 6697–6715 (2014)
35. G. Rigazzi, F. Chiti, R. Fantacci, C. Carlini, Multi-hop d2d networking and resource management scheme for m2m communications over lte-a systems, in *International Wireless Communications and Mobile Computing Conference (IWCMC)* (IEEE, 2014), pp. 973–978
36. L. Militano, A. Orsino, G. Araniti, A. Molinaro, A. Iera, A constrained coalition formation game for multihop D2D content uploading. IEEE Trans. Wirel. Commun., **15**(3), 2012–2024 (2015)
37. X. Lu, P. Wang, D. Niyato, A layered coalitional game framework of wireless relay network. IEEE Trans. Veh. Technol. (63), 472–478 (2014)
38. A. Richard, A. Dadlani, K. Kim, Multicast scheduling and resource allocation algorithms for OFDMA-based systems: a survey. IEEE Commun. Surv. Tutor. **15**, 240–256 (2013)
39. T.P. Low, M.O. Pun, Y.W.P. Hong, C.C.J. Kuo, Optimized opportunistic multicast scheduling (OMS) over wireless cellular networks. IEEE Trans. Wirel. Commun. **9**, 791–801 (2009)
40. L. Lei, Z. Zhong, C. Lin, X. Shen, Operator controlled device-to-device communications in LTE-advanced networks. IEEE Wirel. Commun. **19**, 96–104 (2012)
41. 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Rel. 11, September 2012
42. 3GPP TS 36.440, General aspects and principles for interfaces supporting Multimedia Broadcast Multicast Service (MBMS) within E-UTRAN, Rel. 11, September 2012
43. C. Mehlfuhrer, M. Wrulich, J.C. Ikuno, D. Bosanska, M. Rupp, Simulating the long term evolution physical layer, in *17th European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, August 2009, pp. 1471–1478
44. A. Urie, A. Rudrapatna, C. Raman, J.M. Hanriot, Evolved multimedia broadcast multicast service in LTE: an assessment of system performance under realistic radio network engineering conditions. Bell Labs Tech. J. **18**, 57–76 (2013)
45. S. Deb, S. Jaiswal, K. Nagaraj, Real-time video multicast in WiMAX networks, in *IEEE INFOCOM 2008. The 27th Conference on Computer Communications*, April 2008, pp. 1579–1587
46. S. Sharangi, R. Krishnamurti, M. Hefeeda, Energy-efficient multicasting of scalable video streams over WiMAX networks. IEEE Trans. Multimedia **13**, 102–115 (2011)
47. Y. Wang, L.P. Chau, K.H. Yap, Bit-rate allocation for broadcasting of scalable video over wireless networks. IEEE Trans. Broadcast. **56**, 288–295 (2010)
48. J.G. Andrews, S. Buzzi, C. Wan, S.V. Hanly, A. Lozano, A.C.K. Soong, J.C. Zhang, What will 5G be. IEEE J. Select. Areas Commun. **32**(6), 1065–1082 (2014)

49. G. Piro, A. Orsino, C. Campolo, G. Araniti, G. Boggia, A. Molinaro, D2D in LTE vehicular networking: system model and upper bound performance, in *2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 6–8 October 2015, pp. 281–286
50. X. Lin, R.W. Heath, J.G. Andrews, The interplay between massive MIMO and underlaid D2D networking. IEEE Trans. Wirel. Commun. **14**(6), 3337–3351 (2015)
51. L. Wang, T. Fei, T. Svensson, D. Feng, M. Song, S. Li, Exploiting full duplex for device-to-device communications in heterogeneous networks. IEEE Commun. Mag. **53**(5), 146–152 (2015)

# M2M Communications in 5G

**Jesus Alonso-Zarate and Mischa Dohler**

**Abstract** This chapter provides a summary of the State-of-the-Art and future trends related to wireless connectivity solutions for Machine-to-Machine (M2M) Communications and Machine-Type Communications (MTC) that will be part of 5G networks, and technologies beyond 5G.

## 1 Introduction

In 2009, the Wireless World Research Forum (WWRF) envisioned a total of 7 Trillion (7,000,000,000,000) of connected devices by 2017 all around the world [1]. Such expectations led the magazine Forbes to claim in August 2014 that the Internet of Things (IoT) was the most hyped technology, overtaking Big Data (Fig. 1).

Today, it is possible to assure that the figure envisioned by the WWRF back in 2009 was maybe a bit too optimistic, according to the lower figures brought down to billions or even millions as stated in various market research reports. However, even though the figures may differ by some orders of magnitude, there is a common agreement on the fact the Internet of Things (IoT), or Internet of Everything (IoE), is here to stay, grow, and change the world as we know it. The reality is that, according to Machina Research [2], the number of sensor and actuator shipments in 2014 reached 23.64 billion (which mainly includes the dozen sensors found in a typical smart phone), the number of total connected devices was of more than 14.4 billion, and the number of M2M connections were above 4 billion.

Such promising market figures are motivated by the fact that the concept of having all objects interconnected and Internet-connected is very appealing and can foster the creation of innovative applications and business models, and boost societal

J. Alonso-Zarate (✉)
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC),
Castelldefels, Barcelona, Spain
e-mail: jesus.alonso@cttc.es

M. Dohler
King's College London, London, UK
e-mail: mischa.dohler@kcl.ac.uk



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_13

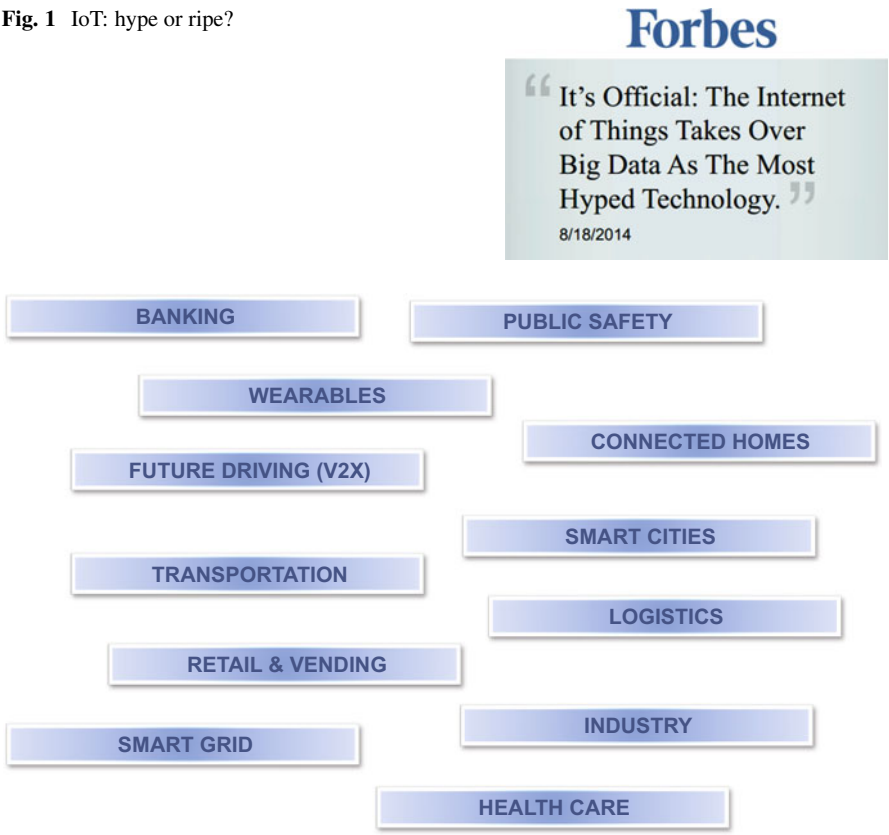


**Fig. 1** IoT: hype or ripe?

**Fig. 2** Key IoT verticals

well-being altogether. The wide range of IoT applications is overwhelmingly wide, including smart grids, smart cities, industries 4.0, smart automotive driving, eHealth, and wearable technology, among many others (Fig. 2).

**Smart grids** constitute one of the key markets, and they have received a lot of attention from both academia and industry. The vision is to improve the efficiency of the electric grid so that the energy demands and waste can be reduced by making a smarter use of energy, the dependency on natural resources can be softened, and the distribution of electricity can become more efficient. The use of automated communicating devices monitoring the state of the grid and the instantaneous use of the resources is essential to attain the objective of achieving a truly smart grid. Very much related to smart grids is the concept of automatic smart meter infrastructure (AMI), which may pose a wider view of the application scenario. Besides electricity, the measurement of water or gas consumption can improve the efficiency in which these resources are dispatched and consumed. A smart and automated reading of meters is beneficial for all involved players: consumers can accurately track their

consumption and thus learn to be more efficient; on the other side, facilities, besides avoiding the manual reading and reporting of meters, can better dimension the production of resources, thus leading to cost savings.

**Smart cities** constitute one of the most appealing applications and markets. Automated parking search, public lights, container levels, notification in case of incident, warning in case of vandalism of security threat, are some of the example of applications that could make cities more efficient by enabling communications between devices. Indeed, the Smart City concept has triggered the attention of many cities, which are struggling to become the worldwide references in applying technology to their city operations. Barcelona, London, New York, or Sydney are just few examples of emerging technology hubs devoted to the Smart City concept.

The concept of the **Industrial Internet**, as introduced by General Electric (GE) a few years ago, promises improvement of efficiencies in industries, driving down the costs of manufacturing, and creating a revolution in a sector which has an impact in many verticals. According to the numbers presented by GE in Fig. 3, improving the efficiency of industrial processes by a factor 1 % can lead to huge economic gains and savings, thus becoming a primary application domain for the IoT.

The **automotive** sector has become a key vertical market for the IoT. Enabling remote monitoring, in-vehicle diagnosis, or car-to-car communication to improve

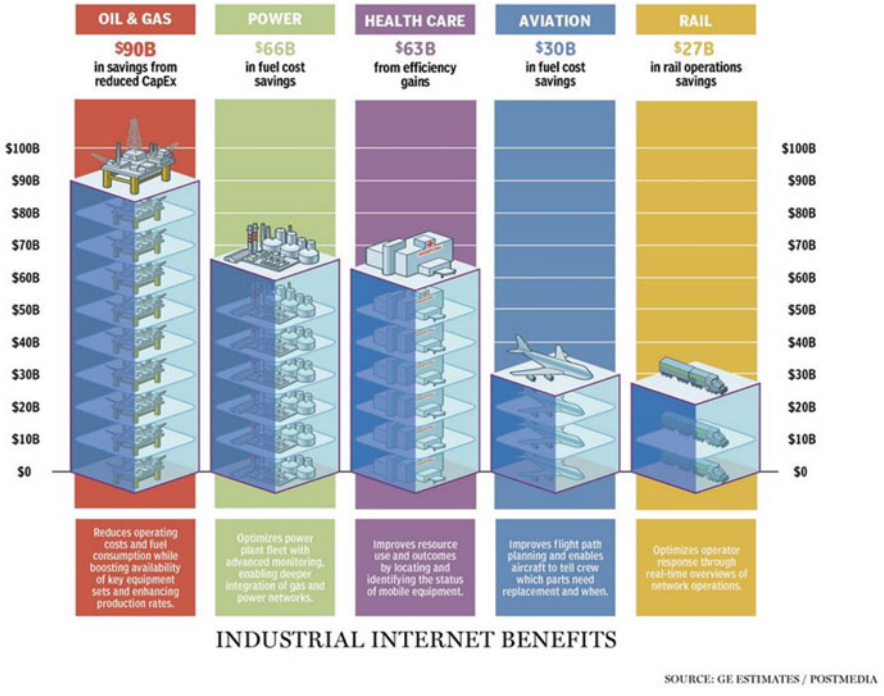


**Fig. 3** Industrial Internet: the power of 1 % (benefits over 15 years when scaled up across the economic system)

safety on-road emerge as key domains to drive the innovation of the IoT. The European eCall initiative, that targets the installation of SIM cards in each vehicle for automatic notification in case of accident, will enable more sophisticated vehicle-based applications in the near future. This is just the beginning.

Today, we also see the advent of **wearable technology**. Bringing technology down to the individual domain, having sensors and actuators interacting with individuals becomes the latest trend in the development of the IoT. Smart clothes, watches, earrings, or glasses, will completely change the world as we know it. **eHealth** raises as a big subdomain of wearable technology. Remote diagnosis and even drug dosing are just two simple examples of applications that can benefit from having automated communications between in- or on-body sensors, actuators located at the patient home, and the centralized servers installed at hospitals.

All in all, the use of automated communications among machines, devices, and people can lead to a smarter and more efficient world. A world which is different from what we now today. However, before this new world becomes a reality, there are still many challenges that need to be faced. Some are related to the new economics triggered by the IoT, or even the societal implications that such hyper-connectivity will impose. However, in this chapter we focus on the technological challenges, many of which are still pertinent because prior technology attempts failed. In particular, the focus will be on the challenges that the definition of the next generation of wireless communications will need to cope with. This is the **5G**.

5G will be the next generation of wireless communications, and its deployment is expected to start by 2020. The Wireless Radio communications Conference (WRC) was held in November 2015, officially starting-off discussions about what 5G will be and what spectrum is needed for it. Over the last 30 years, every new generation of wireless communications has had to deal with the growing need for more data bits and capacity. The evolution from 2G to 4G, passing through 3G, has seen an increase in complexity to satisfy the needs of Mobile Broad Band (MBB) services, requiring more and more data. However, for the first time in history, the arrival of the Internet of Things, dealing with Machine-to-Machine (M2M) communications, also referred to as with the more special 3GPP term Machine-Type Communications (MTC), calls for something different. This new generation of communications has to deal with the coexistence of Human-Type Traffic (HTC), possibly calling for more capacity following the trend of previous evolutions of the various generations, with the emergence of MTC traffic, which is fundamentally different from HTC.

This chapter is devoted to discuss the emerging MTC wireless communications and their role in the definition of 5G. Towards this end, the chapter is organized as follows. In Sect. 2, we describe the elements of an IoT application and define the concept of MTC and M2M, highlighting the differences with HTC traffic. In Sect. 3, we discuss existing technologies to cope with MTC. In Sect. 4, we set the objectives to be met by 5G in order to cope with the various requirements posed by MTC applications. Towards the consecution of these objectives, key enabling technologies are identified and discussed in Sect. 5. Finally, in Sect. 6, an outlook and future perspective of MTC in 5G is provided to conclude the chapter.

## 2 M2M Communications: What Is New?

M2M or MTC constitute a fundamental part of any IoT application. Making things simple for the sake of understanding, the elements of an IoT application are (Fig. 4):

- **End M2M devices**: electronic platforms consisting of at least some kind of energy source (battery or energy harvester), a microprocessor to execute "smart" instructions, a radio transceiver and antenna to communicate with other devices, and sensors and actuators to interact with the physical world, hence the frequently coined term cyber-physical system (see Fig. 5).
- **M2M gateways**: devices in charge of coordinating and concentrating data from a number of end devices in an M2M area network and providing them with connectivity to the Internet.
- **Communications Networks**: backhaul and core networks to connect either the end devices or M2M gateways to the Internet or other external networks. These networks can be either static of virtualized dynamically.
- **M2M Platforms**: hardware/software entities (e.g. data centers, computers, data bases in the cloud, etc.) in charge of providing data ***storage*** and (big) data ***processing*** capabilities to process the data obtained from the sensors and take some actions (*business intelligence*).
- **End User devices**: any kind of device which enables the interaction between the cyber-physical system and humans. This could be, for example, a mobile app running on a smartphone or tablet, or a web service browsed from a laptop.

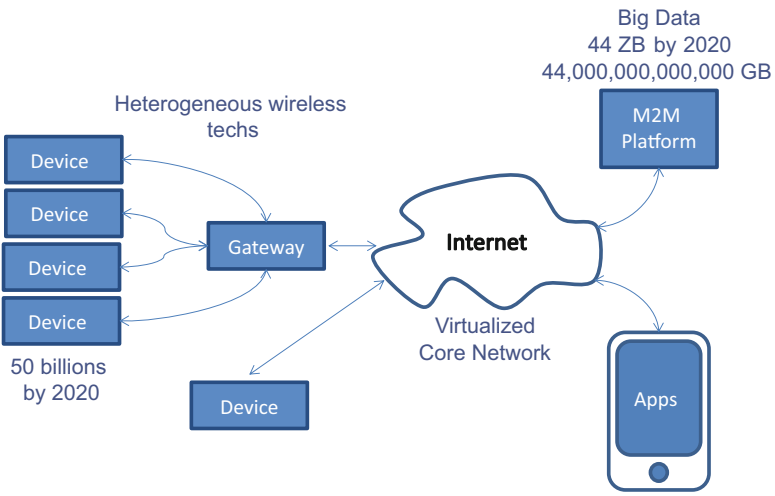


**Fig. 4** Elements of an IoT application

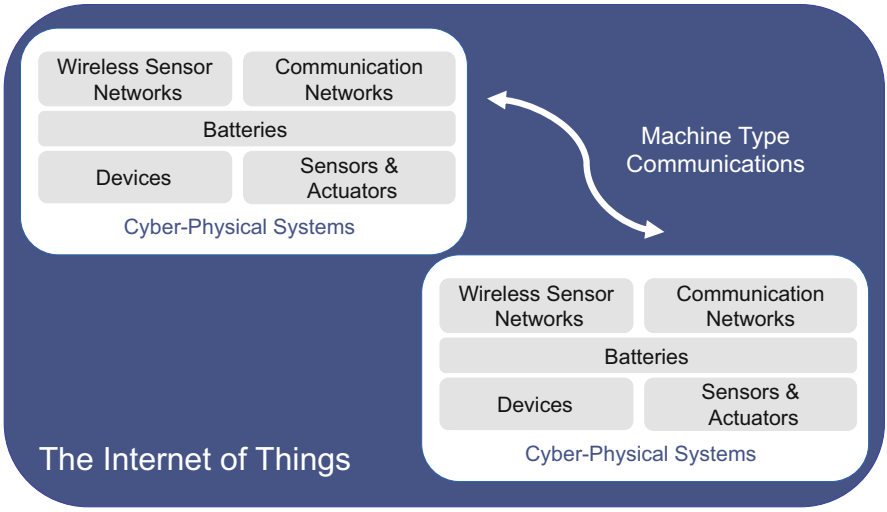


**Fig. 5** The Internet of Things and Machine-Type Communications

The term M2M or MTC is typically used to refer to any exchange of information between machines into this ecosystem. Note that MTC traffic can travel in the access network from end M2M devices to other end M2M devices or to M2M gateways, from the M2M gateways to the core communication networks, within the communication networks themselves, from these networks to/from M2M platforms, and in data exchanges to/from the end user devices. Essentially, M2M or MTC refer to any communication where at least one of the ends is a machine.

These communications between machines, or between machines and people, MTC, are conceptually different from HTC [3]. This implies that the design of future communication networks has to take these differences into account. Some of the main differences that the specification of 5G will have to deal with are [4]:

- **Small Data Transmission**: the data exchanges of MTC are typically small. Think for example of the transmission of a "flag" to indicate whether a light is on or off, or the case of a reading from a smart meter. Few bytes will suffice for that purpose. Compared to HTC traffic, where the payloads are typically very large, MTC deals with small data.
- **Uplink transmissions**: the weight of the uplink of MTC is comparable, or even higher, than downlink. This implies a design shift from traditional HTC communications systems, primarily designed for the downlink. Despite the emergence of interactive applications, still most applications rely on the downlink. The growth of video transmissions and broadcast of multimedia material points at an even higher asynchrony in the data flows for HTC.
- **Energy efficiency**: yes, all communications must be as efficient as possible from the energy point of view to avoid the need for constant battery recharges. However. In MTC, this becomes particularly critical, since in many cases, it will

be not feasible to recharge the batteries of all deployed devices due to scalability (size of the network) or reachability. Note that in some cases, the devices will be installed in wide spread areas, and thus it will turn unfeasible to recharge or replace the batteries individually. The energy efficiency needs to be boosted by a factor 10 at least, so that the lifetime of devices can be measured in tens of years.
- **Amount of devices**: it is expected that the amount of devices connected per aggregation point will increase by a factor 1000 with the IoT. Therefore, MTC devices will be deployed in massive networks.
- **Infrequent transmissions**: exploiting "deep" duty cycles will be key to ensure scalability and energy efficiency of MTC applications. Therefore, the infrequent transmission of data is expected from MTC devices. This has a great impact into the design of communication networks, since the systems need to be designed for a huge number of devices, knowing that just a small proportion of them will be active at any given time.
- **Low Mobility**: many MTC applications are characterized by having very low mobility patterns; in many cases, the devices will be completely immobile. This has strong implications in those systems that integrate control signaling to allow mobility and roaming.
- **Heterogeneous requirements**: the wide variety of applications related to the IoT pose diverse, and sometimes counterpoised requirements. For example, some MTC applications will be delay-constrained, while other will be highly delay tolerant. Therefore, the design of communications networks for MTC will have a to deal with a great variety of different requirements involving capacity, reliability, scalability, delay, Quality of Service (QoS), etc.

The new generation of wireless communications will need to cope with all these requirements of MTC, while ensuring that all HTC services are not jeopardized.

## 3 Existing Communication Technologies for MTC

As it has been discussed in the previous section, MTC are fundamentally different from HTC. Therefore, new emerging communications systems will need to handle with the coexistence of both types of traffic. However, before starting to define a new technology in 5G to cope with MTC, it is necessary to understand which technologies are available today and how suitable (or not) they are to handle such complex scenario that is being formed today.

Figure 6 shows some of the key available wireless communications technologies used to today to power the IoT.

Before continuing with the discussion, they reader may have noted that there are no wired solutions represented in the figure. A wired solution, understood as a solution where sensors or actuators are physically connected to a gateway that provides connectivity to the backbone network, provides an extremely reliable communication channel, with high rates, low delays, and extremely high security.

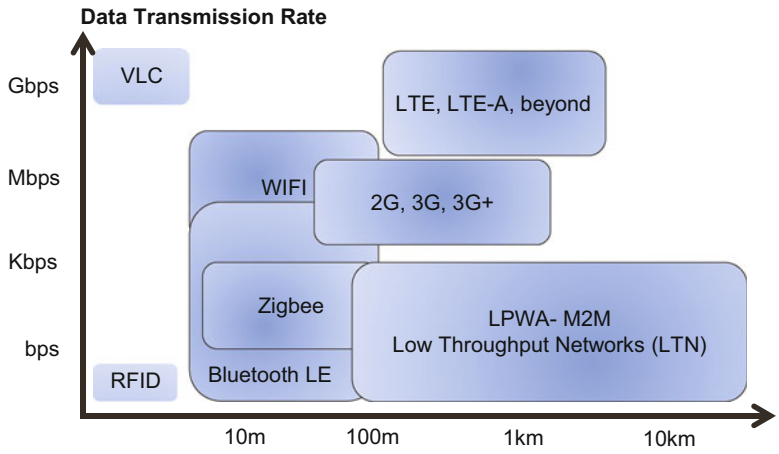


**Fig. 6** Existing wireless technologies

Information cannot be overheard by a third party unless cables are physically intercepted. However, although wired solutions are cheap to maintain, their deployment might be an important entrance barrier for some IoT applications. The main drawback of wired solutions is the lack of mobility and scalability, as the addition of new devices to an M2M network requires the cabling of the new devices.

Therefore, the IoT will be mainly driven by wireless solutions. Connecting billions of devices through wires does not seem a promising approach. Therefore, even though some applications will still rely on cables, it is expected that the majority of M2M applications running the show of the IoT will be based on wireless technologies.

Looking at Fig. 6, it is possible to see how available wireless communication technologies offer different solutions in terms of range and transmission rate. The ones shown in the figure and which will be discussed later on in this chapter are:

- **Visible Light Communications** (high rate, short range).
- **Radio Frequency Identification** (low rate, short range).
- **WiFi** (high rate, medium range).
- **Zigbee-like** (low rate, shot single-hop range and long range via multi-hop).
- **Bluetooth Low Energy** (low rate, short range).
- **2G, 3G, 3G+** (medium rate, long range).
- **LTE, LTE-A**, and beyond (high rate, long range).
- **Low Power Wide Area (LPWA)**—(low rate, very long range).

It is important to note that when it comes to selecting the most suitable technology, other performance indicators may become even more relevant that range or rate; note for example the case of eHealth solutions where security and privacy may be the key performance indicators, or automotive applications, where latency and reliability may become the critical parameters to evaluate. An example

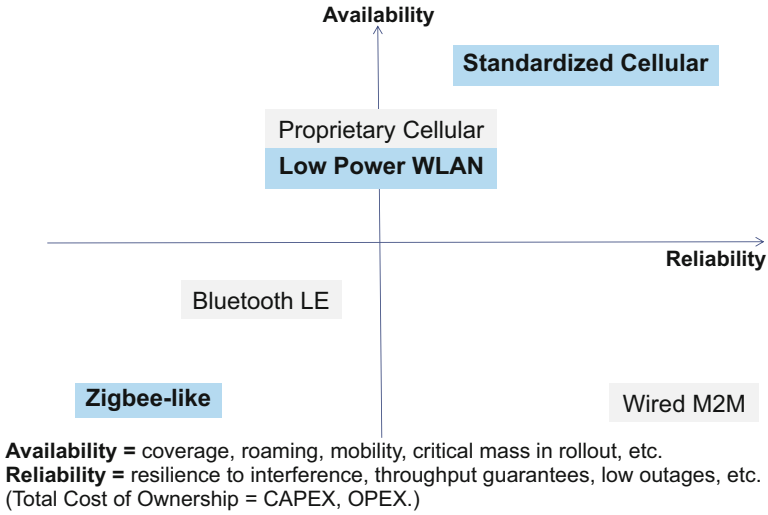


**Fig. 7** Business criteria for technology selection

of business criteria for selecting various technologies for M2M is shown in Fig. 7 where availability and reliability are traded for the purpose of selecting the most suitable technology for each IoT application.

In the following sections, each of the technologies shown in Fig. 6 are briefly described and their suitability for MTC is quantitatively assessed.

## *3.1 Visible Light Communication (VLC)*

The use of modulated LED (Light Emitting Diode) to transmit data is not a mainstream technology yet, but it has a huge potential to become ubiquitous in the coming years [5]. The reasons are pretty simple:

- It is an extremely secure technology; since the light cannot go through walls, only receivers in line of sight of the transmitter can receive data.
- It has no associated radio emissions, thus avoiding limitations of use in "sensitive" areas such as hospitals, primary schools, or chemical plants, among others.
- It has high bandwidth compared to radio technologies, thus leading to high data transmission rates.

On the counter-side, VLC suffers from not being a mainstream technology, thus not mature enough as compared to other radio solutions. Therefore, its cost is still too high to be widely accepted to deploy billions of devices. Also, the limited range and need for line of sight, while providing strong security, it constitutes a limitation of its usage for certain applications.

### 3.2 Radio Frequency Identification (RFID)

Many runners all around the world use the technology based on the ISO 18000 Standard to measure the time they need to complete 10k, 21k or marathons. This is probably the most popular usage of RFID in the world. However, this technology is used in a plethora of other applications, such as proximity access cards, tracking pets or goods for logistics, among many others. By the year 2014, the RFID market was quantified in being close to a 9 billion US dollar market, showing its strength as a tagging technology. Some market research reports claim that the potential of RFID will make it grow up to a 27 billion US dollar by the year 2024.

The relationship between RFID and the IoT is undisputed, and its usage to tag things will play a key role in the future, especially due to its very low cost of deployment. However, the limited data transmission rate and very low coverage limits its applicability to very specific applications, where, by the way, it may see no competition at all in the coming years.

### 3.3 Wireless Local Area Networks (IEEE 802.11)

The IEEE 802.11 Standard is the most successful commercialized standard for Wireless Local Area Networks WLANs. The WiFi alliance was very successful in maximizing the market penetration of WiFi devices that today can be found almost everywhere. This technology operates in license-free bands, the so-called Industrial, Scientific and Medical (ISM) bands.

The ubiquity of WiFi is becoming undisputed. The number of Access Points (Aps) available in urban cities is growing, and the footprint of WiFi is becoming denser. Today's data plans to not allow sharing these WiFi APs, but, being the infrastructure already there, it is just a matter of finding the business case to make all these connecting points available to the public.

In addition, the IEEE 802.11 standards family is extremely dynamic, and new amendments are constantly being developed and evolved to increase capacity (e.g., 802.11ac and ad) or to enable new uses cases such as the IoE, home automation, smart grid or wearable consumers (e.g., 802.11ah).

The latter, i.e., the novel IEEE 802.11ah project, is particularly interesting since it can potentially become an accelerator for IoT uptake. The standardization work of the corresponding ah task group had commenced in November 2010 where the prospective technology is generally based on a variation of IEEE 802.11ac standard but down-clocked by a factor of 10. It is currently being developed to enable low-cost long-range (up to 1 km) connectivity across massive M2M deployments with high spectral and energy efficiencies. Today, thousands of M2M devices may already be found in dense urban areas, which required providing support for up to 6 K machines connecting to a single access point.

Fortunately, IEEE 802.11ah technology does not need to maintain backward compatibility with the other representatives of the IEEE 802.11 family. Operating over different frequencies, 802.11ah could thus afford defining novel compact frame formats, as well as offering more efficient mechanisms to support a large population of devices, advanced channel access schemes, as well as important power saving and throughput enhancements. As the result, 802.11ah is believed to significantly enrich the family of 802.11 protocols, which already receive increasing attention from mobile network operators willing to introduce low-cost connectivity in unlicensed bands.

All these amendments are being already developed with security measures in place, thus making WiFi a very trustful and secure technology. For all these reasons, the cost of WiFi chips is dramatically going down, thus becoming very competitive to become a major driver of the IoT [6].

On the downside, all IEEE 802.11-based Standards operate in an increasingly crowded license-free ISM band, thus suffering from inter-system interference, suffering from limitations in the transmit power, and suffering from a lack of network planning. The overlapping of access point's coverage can lead to very low performance in highly dense environments.

### *3.4 IEEE 802.15.4 (Zigbee-Like Solutions)*

The IEEE 802.15.4 Standard, promoted by the Zigbee Alliance, was originally defined as the most energy efficient solution for short-range communications also operating in ISM bands. Unfortunately, some recent research works have shown that WiFi used in Low Power mode can be more efficient than an IEEE 802.15.4 Network [7]. WiFi in Low Power mode is nothing else than using smart duty cycling mechanisms to let devices go to sleep whenever they have no data to be transmitted or received. Since energy consumption is the result of power multiplied by time, it has been shown that is it more energy efficient to transmit for short periods of time at high power (also reaching long distances and reducing the complexity of multi-hopping), rather than transmitting for long periods of time at low power.

Further, Zigbee-like solutions suffer from many drawbacks that limit their applications to very specific and delimited applications, such as industrial environments. Zigbee has not followed the market success story of WiFi, and the infrastructure is barely found. In addition, the need for multi-hop transmissions to cover medium and large distances makes its design, implementation, and operation very costly and complex.

## 3.5 Bluetooth Low Energy (BLE)

Bluetooth also operates in ISM bands. Therefore, it also suffers from the effects of operating in an increasingly crowded band. The last generation of Bluetooth products, based on the specification Bluetooth 4.0 and featuring very low energy consumption have become very popular for the IoT. An example can be seen with the iBeacon technology made popular by Apple, which lets BLE receivers to detect "beacons" transmitted from a base station at close proximity (<10 m) and retrieve data of interest from the Internet based on the geo-location information. This has a huge potential for delivering personalized services, and this is the reason why a number of companies exploiting the iBeacon technology is emerging. BLE has something on favor; the low cost and extremely high ubiquity it has achieved. Almost all smartphones today are shipped with a BLE transceiver, and thus can communicate via BLE with other devices and access points. In this case, the low transmission rate and short range, play on favor of a technology that is increasingly reaching the IoT market for very specific applications.

## 3.6 3GPP Solutions

### 3.6.1 The Power of Standardized Cellular Technologies

All the technologies discussed so far in previous sections operate in ISM bands; this means that they are frequency bands free of license which can be used by anyone under certain regulations (transmission power and transmission time). The inter- and intra-system interference thus limits the reliability of communications conducted in these bands. If a Service Level Agreement (SLA) has to be committed and some guarantees of QoS are to be provided, communication in license-bands becomes the only alternative.

3GPP Standards for cellular communications operate in license bands. Mobile Network Operator (MNOs) pay big amounts of money to bid for chunks of frequency bands where they can operate without (legal) interference from other systems. For this reason, they invest in deploying a vast infrastructure which can provide customers with reliable communications. Today, the coverage of standardized cellular systems is enormous, and almost very populated inch in the planet has some kind of cellular coverage.

In addition, these networks can provide mobility and roaming, something that cannot be offered by any other radio system today.

Unfortunately, as it is discussed below, the evolution of standardized cellular systems has followed the increasing demand for higher data rates, and not the specific needs of MTC described in Sect. 2 of this chapter. However, from release 12 on of 3GPP, new emerging releases for cellular systems are starting to deal with both HTC and MTC.

### 3.6.2 From 2G to 4G

Global System for Mobile Communications (GSM) broke into the market in 1990, providing a cellular network infrastructure designed for the transmission of voice calls. This technology showed an extraordinary market penetration and today provides connectivity worldwide. With the irruption of the Internet by the mid-1990s, the transmission of data and the sense of ubiquitous connectivity has become almost a commodity for humans, who demand more and more data transmission capability and some degree of ensured QoS. Unfortunately, GSM was designed for the transmission of voice, and its capability to provide efficient data transmission was very limited. The response of the standard bodies towards this need was the continuous evolution of the cellular technology.

General Packet Radio Service (GPRS) was the first evolution of GSM to provide increased data transmission capabilities, raising the concept of the 2.5 generation of mobile communications (2.5G). Then came the third generation networks (3G), with technologies such as EDGE (Enhanced Data Rates for GSM Evolution), UMTS (Universal Mobile Telecommunications Systems) and CDMA200. High Data Packet Access (HDPA) is called the 3.5G, and still offers increased data transmission capabilities. Even though the data transmission capabilities with 3G and 3.5G are considerably higher than the capability of GSM, the non-top increasing demand for multimedia contents and interactive applications such as social networking, still feeds the need for increased data rates and higher performance. This is the motivation for the fourth generation (4G) of cellular mobile communications, with transmission rates up to 1 Gbps are envisioned in the downlink. This is the Long Term Evolution (LTE)—Release 8 and LTE Advanced—Release 10.

From 2G to 4G, the evolution of cellular networks has been driven by the need for higher transmission rate and increased QoS to meet the increasing demands of humans to use multimedia and interactive applications. However, now is time to evolve and design communication networks so that they can efficiently cope with HTC and MTC. Indeed, the Release 12 of 3GPP already included, for the first time, some improvements towards MTC traffic [8]. One was, for example, the inclusion of a power saving mode for a newly defined Category-0 of user terminals, with reduced complexity for the sake of lower cost and complexity of MTC devices. Further progress will be achieved in Release 13 bringing more complexity reduction (75 % complexity reduction compared to Cat-1 modems), reduced bandwidth, down to 1,4MHx, 10+ years of battery, and 15–20 dB coverage extension.

### 3.6.3 Release 13 and Road to 5G

3GPP Standards continue their evolution to meet the requirements of next generation networks and enable new use cases, new services, and new applications. For this reason, the next releases of 3GPP will continue aiming at improving the capacity of networks for Mobile Broad Band, but also for MTC services.

For the MBB, the next efforts aim at the so-called LTE-Unlicensed. The idea is that LTE can operate also in unlicensed bands via coexistence with WiFi networks, but also via usage of LTE radio in unlicensed bands. While the former option has been progressively integrated into the releases of 3GPP, the concept of LTE-U has been a brand fresh approach for 3GPPP standards to operate in the 5 GHz band. The design of Licensed Assisted Access (LAA) will be the key to ensure the right operation of LTE-U. Essentially, this consists in using unlicensed under the control and cooperation of the main cellular network. Critical data would travel through the main site using licensed spectrum, while best effort traffic can be offloaded via the unlicensed links. Other techniques being included in next releases of 3GPP aim at including enhanced LTE carrier aggregation, exploitation of full-dimension MIMO, and downlink multi-user transmission using superposition coding.

As for the IoT, the activities of the 3GPP towards 5G can be divided into two big groups:

(1) **Further enhancements of LTE for Machine-Type Communications**, building on the work initiated in the release 12 (which introduced the Category-0 of terminals and new power saving states), reducing the bandwidth channel operation down to 1.4 MHz (in contrast with the 20 MHz of release 12), enhancing coverage with an increased 15 dB budget for extended coverage, simplifying PHY/MAC functions, and further improving power usage.
(2) Design of a brand new narrow-band radio optimized for the IoT; on June 2016, the 3GPP completed the standardization of NB-IOT. This is part of Release 13 and includes optimizations for MTC. NB-IOT can operate in-band or out-of-band of licensed frequencies, thus offering very high flexibility. NB-IOT features very narrowband operation with a bandwidth below 200 kHz (thus compatible with GSM bands). NB-IOT focuses on very low data rate transmissions (below 100 bps), with very limited mobility support (to reduce signaling), and ensures very large coverage and both extremely low cost and power consumption.

Transversal to both HTC and MTC, 3GPP is also working on innovative Device-to-Device communications, enabling the by-pass of the communication through the eNodeB (base stations) to dramatically reduce end-to-end latency.

## *3.7 Low Power Wide Area Networks (LPWA)*

While 3GPP and IEEE do their homework standardizing technologies for MTC, some proprietary solutions have started to gain the market. These are the so-called Low Power Wide Area (LPWA) networks, which offer **very large coverage**, being able to give service to an entire city with a very little number of connection points, **offer extremely low cost**, since they have been optimized for MTC and cannot provide services for HTC, **offer a SLA to the customer**, thus becoming very interesting option for the end users, and what is more important: **they are already in the market** [9].

Examples of such solutions are those proposed by Neul, Sigfox, and the LoRa Alliance (out of Cycleo/Semtech), each of them offering different business models and technical solutions.

## 4 Key Objectives for MTC in 5G

Future communication networks must be designed to ensure that the following objectives can be meet to satisfy the new needs posed by emerging MTC services [10]:

1. *Reduce* **latency** of wireless communications, enabling end-to-end ***latencies below 5 ms*** for mission-critical applications where real-time transmissions are required [11, 12]. Today's wireless technology is unable to provide such delay figures, thus blocking some potential applications, which are expected to have a huge impact on economy, industry, and society in various vertical sectors such as automotive, industrial automation, or smart grid, among others.
2. *Improve* **reliability** of wireless communications, defined as the probability that a certain amount of data to/from an end user is successfully transmitted to another peer within a predefined time frame, i.e., before a certain deadline expires. The amount of data to be transmitted and the deadline depend on the service characteristics of the underlying use case. In this context, a reliability of at least 99.9999 % is targeted within the project, i.e., the probability that a packet is not delivered within the specified deadline must be below $10^{-6}$ (or equivalently, at most one in a million packets does not arrive within the specified deadline, on average). For certain types of applications (e.g., industrial automation), an even higher reliability of up to $10^{-9}$ packet loss rate (outside latency) is sought.
3. *Ensure* **availability** of communication networks by extending network coverage, ensuring network access, and providing roaming support to make sure that critical applications are not in outage when needed. An availability of 99.999 % is sought (i.e., less than ∼300 s of accumulated outage per annum), aiming at equating availability of wireless and wired solutions.
4. *Guarantee* **interoperability** of heterogeneous communication technologies; in a hyper-connected world with humans, things, and machines connected, it is necessary to make sure that 5G can cope with a great variety of heterogeneous applications, devices, and radio technologies.
5. *Increase* the **number of connected devices** by a factor of up to 100 with respect to currently deployed LTE networks (typically up to 250 simultaneous users per cell); regardless of the particular numbers of the many publicly available market forecasts, there is no doubt that 5G networks will need to cope with an unprecedented number of connected devices [13].

6. *Boost* **energy efficiency** by a factor of 10 with respect to currently deployed MTC technology, to significantly reduce energy consumption and operational costs of the network infrastructure, and allow battery-driven devices to operate without the need to recharge or replace batteries frequently, which would constrain the number of potential applications [14].
7. ***Secure*** communications [15]; many MTC applications will involve personal or confidential data that must not reach the public domain, nor be modified or replayed by unauthorized parties. This becomes a severe challenge when the amount of data to be transmitted is very small and the acceptable securing overhead must be kept to the minimum.

Note that not all specific objectives above are equally relevant to all applications. For example, ultra-low latency is essential for certain future driving applications, whereas energy efficiency is essential for ultra-low power MTC. Furthermore, achieving certain objectives may entail sacrificing others (e.g., very high energy efficiency can usually only be achieved with long sleep modes, which increases latency and has an impact on reliability and availability). All these trade-offs must be well understood in the long run.

## 5 Key Enabling Technologies for MTC in 5G

In order to achieve the objectives defined in the previous sections, many solutions have to be explored at various domains. Considering the particular focus of this chapter towards available wireless communication networks, the following non-comprehensive list of key enabling technologies are identified:

(1) In order to ensure scalability of networks, it is necessary to seek for innovative ways of managing the access to shared resources. Looking at the radio access, still today most of wireless technologies are based on ALOHA-based protocols, which suffer from congestion as the number of users increase. Solutions such as those proposed in [16, 17], based on **DQ technologies** constitute innovative ways of sharing resources that must be further understood. Beyond the radio access, such sharing techniques could be applied to any domain where common resources need to be shared, such as available resources in a cloud.
(2) Collisions in the wireless medium must be further exploited to attain useful information. **Advanced coding techniques** from the signal processing domain can be brought to the protocol design domain to turn collisions into useful realizations of data transmissions [18]. Success Interference Cancellation (**SIC**) techniques constitute a promising technology to be further studied and whose implementation must be better matured [19].
(3) Enabling **Device-to-Device** communications, avoiding the two-hop path from devices to infrastructure and from infrastructure to devices, can dramatically reduce the energy of devices, offload main cellular networks, and achieve end-to-end latencies which cannot be attained today [20]. The use of D2D has

started to be integrated into 3GPPP releases, but still many functionalities need to be further studied and understood to make them viable in future generation networks. Notably, inter-operator D2D functionalities are needed to make D2D a truly scalable solution; i.e. a Siemens sensor on a T-Mobile data plan should be able to communicate directly with an ABB actuator on a Telefonica data plan. The usage of LTE in unlicensed bands (**LTE-U**) is also emerging as an alternative to the D2D facilitated by other radios such as WiFi.

(4) **Heterogeneous networks**, combining different cell-sizes and technologies to create a unique and ubiquitous communication infrastructure. As it was pointed in [21], the greatest increases in capacity over the last 35 years have been achieved thanks to making cells smaller. Future 5G networks will be comprised of a gamut of technologies and sites that need to coexist together in harmony.

(5) Network Function Virtualization (**NFV**) and Software Defined Networks (**SDN**) are promising technologies to make the backhaul and core networks flexible and capable to managing a variable and possibly unexpectedly dynamic network conditions [22]. Moving processing power and functions along the network, closer to the edge, can improve the overall QoS perceived by the users, enabling applications which today cannot be realized due to the limitations of an extremely static core network configuration.

## 6 Summary and Outlook

The definition of 5G must take the emergence of the IoT and its need to coexist with Mobile Broad Band services into account all along the way of the specification process. Involving communication between machines, either M2M or MTC, communication networks need to deal with traffic patterns which are fundamentally different from HTC. These have been discussed in Sect. 2 of this chapter. Existing wireless communication technologies presented and discussed in Sect. 3 were originally designed for HTC, and thus still further improvements for the future are to be sought. Standardization bodies all around the world are conducting great efforts to standardized new technical solutions that can cope with the particular requirements of MTC. In the meanwhile, some proprietary solutions are already available in the market to cover the needs to today's application. In the long run, it is difficult to ensure that the IoT will become mainstream unless real standards are defined and widely accepted. This was the example of WiFi, and this must be the example for the IoT.

The definition of requirements for the IoT, due to the great variety of possible applications and markets, needs to be agreed with vertical players which are the final end users of the technology. Once these requirements are set, new emerging technology enablers need to be explored to make sure that the necessities are covered, as it has been discussed in Sect. 5. At the air interface, massive MIMO, disruptive access methods based on something different to ALOHA, such as DQ technologies, exploitation of Successive Interference Cancellation, or Device-to-Device communications have been identified as key enablers, identifying the use

of LTE in unlicensed bands (LTE-U) as an emerging area for research. At the architectural level, both Network Function Virtualization (NFV) and Software Defined Networks (SDN) will be core enablers for the core networks of the future.

Global initiatives are working together to make the realization of 5G a reality so that it can efficiently deal with both traffic generated by humans and machines in good harmony.

## References


1. User Scenarios 2020 – A Worldwide Wireless Future. WWRF Outlook, July 2009. Available online
2. The need for low cost, high reach, wide area connectivity for the Internet of Things: A Mobile Network Operator's Perspective. White Paper, February 2014
3. A. Laya, K. Wang, L. Alonso, J. Alonso-Zarate, J. Markendahl, Future challenges in efficiently supporting M2M in the long term evolution (LTE) standards, in *Machine-to-Machine (M2M) Communications, Architecture, Performance and Applications* (Woodhead Publishing, 12 January 2015)
4. K. Zheng, S. Ou, J. Alonso-Zarate, M. Dohler, F. Liu, H. Zhu, Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications. IEEE Wirel. Commun. **21**(3), 12–18 (2014)
5. A. Jovicic, Li. Junyi, T. Richardson, Visible light communication: opportunities, challenges and the path to market. IEEE Commun. Mag. **51**(12) (2013)
6. S. Tozlu, M. Senel, Wi Mao, A. Keshavarzian, Wi-Fi enabled sensors for the Internet of Things: a practical approach. IEEE Commun. Mag. **50**(6) (2012)
7. S. Tozlu, Feasibility of Wi-Fi enabled sensors for the Internet of Things, in *Proceedings of the Wireless Communications and Mobile Computing Conference (IWCMC)*, 2011
8. 3GPP Technical Report 37.869, "Study on Enhancements to Machine-Type Communications (MTC) and other Mobile Data Applications", V12.0.0, Sep. 2013
9. X. Xiong, K. Zheng, X. Rongtao, X. Wei, P. Chatzimisios, Low power wide area machine-to-machine networks: key techniques and prototype. IEEE Commun. Mag. **53**(9), 64–71 (2015)
10. E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, Y. Selén, J. Sköld, 5G wireless access: requirements and realization. IEEE Commun. Mag. (2014)
11. N.A. Johansson, Y.-P.E. Wang, E. Eriksson, M. Hessler, Radio access for ultra-reliable and low-latency 5G communications, in *IEEE International Conference on Communication (ICC)*, London, June 2015
12. O.N.C. Yilmaz, Y.-P.E. Wang, N.A. Johansson, N. Brahmi, S.A. Ashraf, J. Sachs, Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case, in *IEEE International Conference on Communication (ICC)*, London, June 2015
13. A. Laya, L. Alonso, J. Alonso-Zarate, Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives. IEEE Commun. Surv. Tutorials, Special Issue on Machine-to-Machine Communications, **16**(1) (2014)
14. T. Tirronen, A. Larmo, J. Sachs, B. Lindoff, N. Wiberg, Machine-to-machine communication with long-term evolution with reduced device energy consumption. Trans. Emerg. Telecommun. Technol. **24**(4), 413–426 (2013)
15. A. Grau, Can you trust your fridge? IEEE Spectrum **52**(3) (2015)
16. J. Alonso-Zárate, E. Kartsakli, A. Cateura, C. Verikoukis, L. Alonso, A near-optimum cross-layered distributed queuing protocol for wireless LAN. IEEE Wirel. Commun. Mag., Special Issue on MAC protocols for WLAN, **15**(1), 48–55 (February 2008), IF. 3.180, ISSN 1536-1284

17. A. Laya, C. Kalalas, F. Vazquez-Gallego, L. Alonso, J. Alonso-Zarate, Goodbye, ALOHA!, IEEE Access, April 2016, Special Section: The Plethora of Research in Internet of Things (IoT)
18. N.K. Pratas, H. Thomsen, C. Stefanovic, P. Popovski, Code-expanded random access for machine-type communications, in *2012 IEEE Global Communications Conference (Globecom) Workshops*, Dec. 2012
19. E. Paolini, C. Stefanovic, G. Liva, P. Popovski, Coded random access: applying codes on graphs to design random access protocols. IEEE Commun. Mag. **53**(6), 144–150 (2015)
20. A. Laya, K. Wang, A. Widaa, J. Alonso-Zarate, J. Markendahl, L. Alonso, Device-to-device communications and small cells: enabling spectrum reuse for dense networks. IEEE Wirel. Commun. Mag. **21**(4), 98–105 (2014)
21. M. Dohler, R.W. Heath, A. Lozano, C.B. Papadis, R.A. Valenzuela, Is the PHY layer dead? IEEE Commun. Mag. **49**(4) (2011)
22. M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, K. Zheng, Toward 5G Densenets: architectural advances for effective machine-type communications over femtocells. IEEE Commun. Mag. **53**(1), 134–141 (2015)

# Design Techniques of 5G Mobile Devices in the Dark Silicon Era

**Imed Ben Dhaou and Hannu Tenhunen**

**Abstract** In the internet of things age, future communication technologies should provide the necessary bandwidth and latency for the connection of billion devices and the development of ubiquitous applications to improve the quality of life. The design of the prospected mobile communication system needs wide skills in wireless communication, analog circuit design, embedded system, microwave technology, and so forth. System level analyses, design space exploration, performance tradeoffs are some key steps that enable the design of low-cost, energy efficient, ubiquitous and flexible transceiver. This chapter provides comprehensive design techniques for 5G mobile communication in the dark silicon era and using More than Moore technology (MtM).

## 1 Introduction

Wireless sensor networks have enabled the development of a plethora of smart and ubiquitous applications (ambient intelligence, smart-home, intelligent transportation system, smart-city, etc.) Fig. 1 shows an electronic system that aims to combat traffic noise in urban areas [1]. The system uses an acoustic sensor (microphone) combined with a DSP system, a communication block, and a CCTV camera. The communication block is used to access the Internet wirelessly. The Long-term evaluation system (4G LTE) offers the required data-rate and latency to the anti-traffic noise system.

The growing importance of the Internet of Things (IoT) and the ever increasing demands for smart and ubiquitous systems has led to major technical challenges beyond the capabilities of the 4G system.

To address the growing need for higher-bandwidth, low-latency, and the connection of billion devices, the international communities have started to look for a replacement to the current 4G system.

I. Ben Dhaou (✉)
College of Engineering, Qassim University, Saudi Arabia
e-mail: phd.imed.bendhaou@ieee.org

H. Tenhunen
School of Information and Communication Technology Royal Institute of Technology, Sweden



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_14

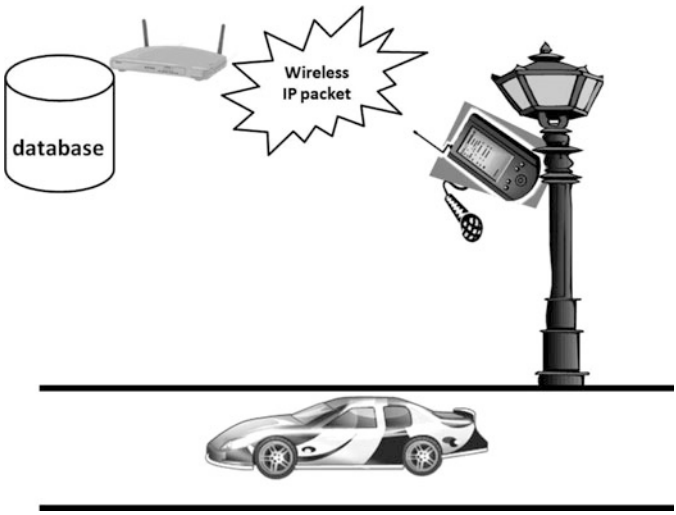


**Fig. 1** An anti-traffic noise system proposed in [1]

As discussed in [2], there are five disruptive technologies for the 5G. These technologies are: (1) a device-centric architecture, (2) a millimeter wave communication, (3) a massive MIMO system, (4) a smarter device, and (5) a native support for machine-to-machine communication.

It is expected that the 5G system will enable the integration of the various wireless communication technologies (GSM, WCDMA, 3G, 4G, DVB-T, WLAN, RFID, Zigbee, to name a few) [3].

To quench the future wireless communication requirements, the authors of [4] suggest that the 5G communication network should offer for example a round-trip latency less than 1 ms, 1000× aggregate data-rate compared to 4G and guarantees 100 Mbps for 95 % of users.

The success of the 5G terminals will be contingent to the fabrication technology. CMOS is the technology of choice for the manufacturing of the base-band signal processing blocks. For the RF front-end, CMOS has also been used to fabricate the various RF front-end blocks (mixers, amplifiers, oscillators, down-converters, filters, and so forth) [5].

The growing number of wireless standards, the need for a single radio transceiver and the advances in CMOS technology are the main factors that have contributed to the birth of the software defined radio (SDR). The SDR aims to design a flexible and reconfigurable radio architecture [6]. The reconfiguration is done using a specialized software. The SDR has brought into light various design techniques for the RF front-end and the base-band signal processing. Among the emerging paradigm is the *dirty RF* [7]. The design of the SDR architecture using the dirty-RF is accomplished using design trade-offs at the system level. The analog front-end is designed using a direct-conversion receiver architecture. The advantages of the direct-conversion

receiver (DCR) compared to the superheterodyne are low-cost, reduced complexity and energy efficiency [8]. However, the DCR architecture suffers from various impairments such as even-order distortion, dc-offset, I/Q mismatches, and 1/f noise [9].

The dirty-RF paradigm allows to design imperfect analog front-end. To mitigate the noisy RF front-end, the receiver should use effective numeric algorithms. These algorithms seek to minimize the bit-error rate or symbol error rate for a wide-range of signal to noise ratios.

Among the popular techniques in the dirty-RF paradigm are: (1) scrambling algorithms to reduce the peak-to-average power ratio for OFDM signal [10], and (2) a frequency offset compensation scheme combined with channel estimation for OFDM DCR architecture [11, 12].

In recent years, the mobile telephony systems (devices and network) have migrated from the voice-oriented applications to ubiquitous computing. Current smart phones offer the user a plethora of pervasive applications some of which are: location based services, navigation [13], augmented reality, and vehicle to pedestrian communication [14].

The development of the smart phone is accredited to the progress in sensor technology, the development of the system-in-package, the progress in the embedded system design, and so forth. Figure 2 shows a generic smart-phone architecture.

A smart-phone contains the following blocks: (1) a heterogeneous multi-core processor (base-band, protocol, application and GPU processors), (2) a RF front-end (amplifiers, mixers, filers, oscillator, ADC, DAC, and so forth), (3) sensors (GPS, gyroscope, micro-phone, cameras, accelerometer, etc.), (4) a machine-to-machine communication platform (Bluetooth, NFC, V2X communication, and so on), and (5) a user interface (keyboard, loudspeaker, tactile screen, and so forth).

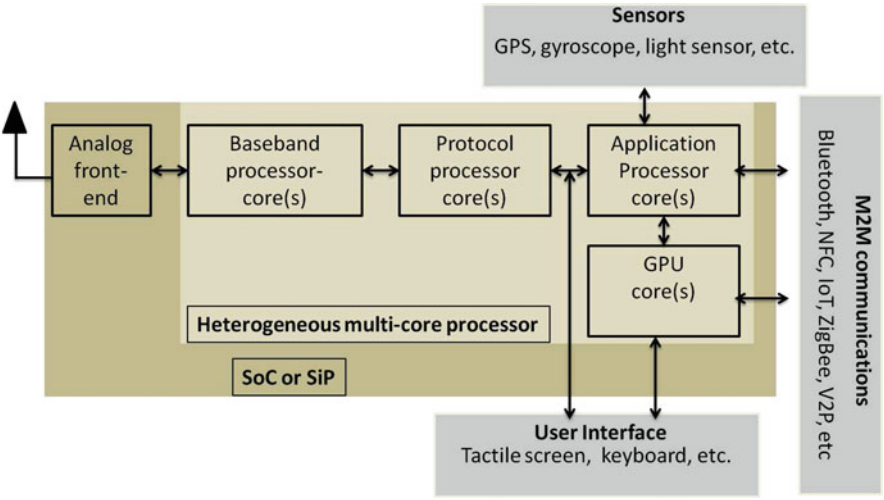


**Fig. 2** Smart-phone architecture

For years, Moore's law has been the vehicle for aggressive transistor scaling. The technology scaling paradigm allows to design faster, cheaper and energy-efficient systems. However the non-digital blocks do not benefit equally from the scaling. Moreover, in the nano-meter regime, the transistor feature sizes and interconnects are hitting the borderline. Hence, many research efforts have started to look for a replacement to the CMOS technology beyond Moore's law [15–18]. The international roadmap for the semiconductor has provided a dual trends for the technology roadmap. The first trend is the continuous scaling of the transistor feature size (More Moore). The second trend is the continuous integration of more non-digital blocks (sensors, RF, actuators, and so forth). Figure 3 depicts these dual trends.

The rest of the chapter is organized as follows. Section 2 describes the problems associated with the interconnect centric design, reports the NoC as a new methodologies to address the complexity associated with on-chip communication and elaborates several techniques to design baseband circuitry in the dark silicon age. Section 3 presents the multi-gate MOSFET technology for the design of heterogeneous multi-core process and the RF front-end for the feature size smaller than 22 nm. Section 5 overviews the various technology and methodologies to build sensors for 5G ubiquitous devices. Finally, Sect. 7 summarizes the chapter.

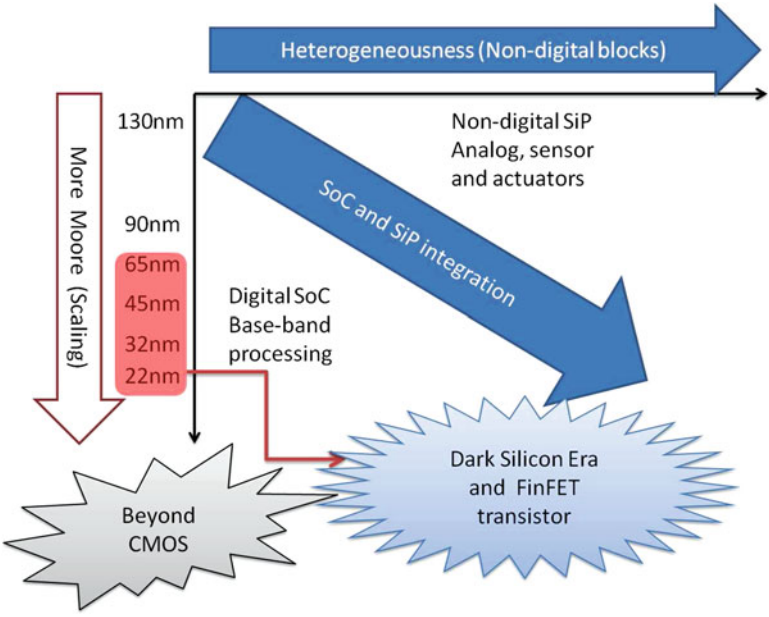


**Fig. 3** ITRS Roadmap for MtM Technology

## 2 Signal Processor Design in the Dark Silicon Era

CMOS is the mainframe technology for the design of digital integrated circuitry. Its popularity is attributed to a number of factors such as: (1) low cost, (2) high yield, (3) low power consumption, (4) high-noise immunity, and (5) the scaling capabilities.

In 1965, Gordan Moore postulated his famous prediction, stating that the number of transistor will double every 18 months. Ever since its postulation, Moore's law has been the guideline for aggressive transistor scaling and large scale integration. Figure 4 plots the transistor density (number of transistor per square millimeter) for the generations of Intel microprocessor. The figure clearly shows that technology scaling spurs an exponential increase in the transistor density.

Technology scaling enables the implementation of faster, smaller and energy efficient transistor. To understand the underlying mechanisms behind technology scaling, we will revisit the device characteristics in the nano-meter regime (short-channel effect) [19]. The NMOS transistor (shown in Fig. 5) is built using specially selected layers of metal and insulator. The transistor is a four terminal device (body, gate, drain and source). The body of the NMOS transistor is the p-type semiconductor.

The NMOS device operates in linear, saturated or cutoff mode. Using the long-channel model, the expression for the drain-source current of an NMOS transistor of length $L$ and width $W$ is shown in (1).

$$I_{ds} = \begin{cases} 0 & \text{if } V_{gs} < V_t \\ \beta(V_{gs} - V_t - \frac{V_{ds}}{2})V_{ds} & \text{if } V_{ds} < V_{gs} - V_t \\ \frac{\beta}{2}(V_{gs} - V_t)^2 & \text{if } V_{ds} > V_{gs} - V_t \end{cases} \quad (1)$$

where $\beta = \mu C_{ox}\frac{W}{L}$, $C_{ox}$ is the gate oxide capacitance, $\mu$ is the electron mobility, $V_t$ is the threshold voltage, $V_{ds}$ is the drain-source voltage, and $V_{gs}$ is the gate-source voltage.

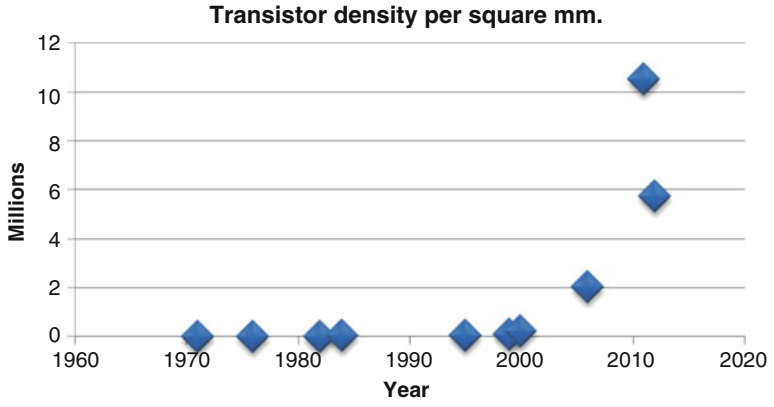


**Fig. 4** Evolution of transistor density for Intel microprocessor

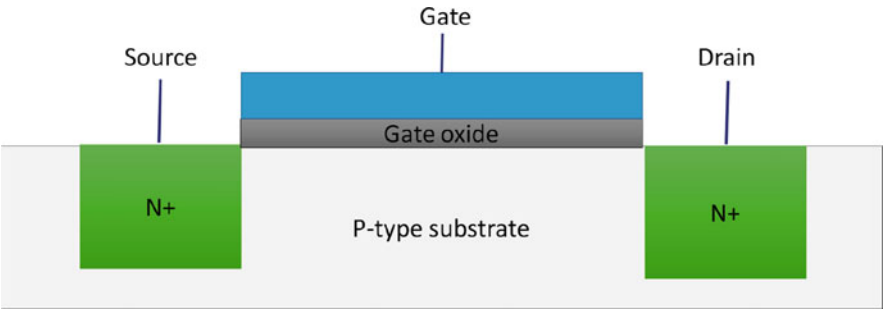


**Fig. 5** Layout of NMOS transistor

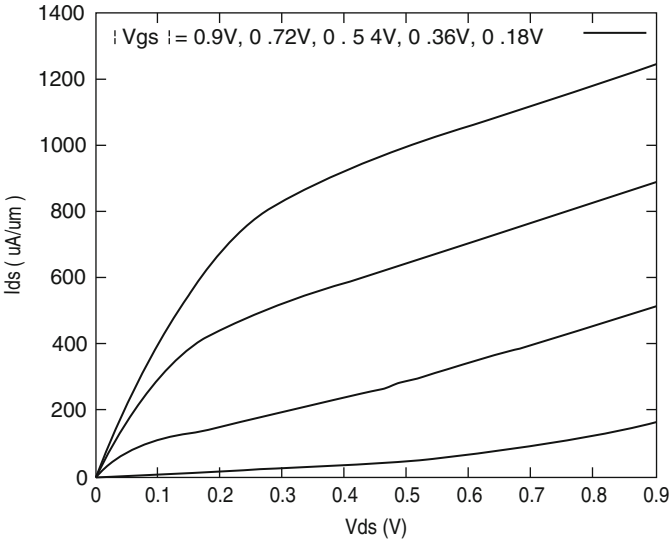


**Fig. 6** IV characteristic for a 32 nm NMOS transistor

The expression for the IV characteristics of the NMOS transistor has been calculated under the assumption that the lateral electric field between the source and the drain is constant.

In sub-nanometer technology (below 90 nm), the secondary effects neglected in long-channel model have become significant. The prominent ramifications are for instance mobility degradation, velocity saturation, channel length modulation, drain-induced barrier lowering, and sub-threshold leakage [19].

To illustrate the IV characteristic of NMOS transistor in sub-nanometer technology, we used the predictive technology model described in [20] to obtain the IV characteristics of the transistor. Figure 6 depicts the relationship between $I_{ds}$ and $V_{ds}$ in the three operating regions for the 32 nm NMOS transistor.

## 2.1 Low-Power Design

The power consumption of the CMOS circuits is composed of the dynamic and static components. The dynamic power consumption is the sum of the short-circuit power consumption and the power consumed due to the switched capacitance. The latter is the dominant. As a result, the widely accepted expression for the dynamic power consumption is shown in (2).

$$P_{dynamic} \approx \alpha C V_{dd}^2 f, \tag{2}$$

where $f$ is the clock frequency, $\alpha$ is the switching activity, $C$ is the circuit capacitance, and $V_{dd}$ is the supply voltage.

In the idle status, the CMOS circuitry consumes static power which is due to leakage currents (sub threshold, gate, junction) and contention current.

In nano-meter design, the sub-threshold leakage current has started to represent over 40 % of the total power consumption [21]. Several techniques have been proposed to bring down the leakage power consumption. Those techniques have been developed for the various abstraction levels (system, algorithm, architecture, and circuit). The power gating technique has been reported as an effective way to cancel the leakage current [19]. There are many schemes and circuit styles for power gating (fine-grain, coarse-grain, and so forth). Figure 7 exemplifies circuits for power gating [19].

Over the last decades, several approaches have been proposed to reduce the dynamic power consumption [22]. It has been widely accepted that higher abstraction levels have larger savings than lower level. The system level power optimization can lead to a sturdy power savings. The dynamic voltage combined with the

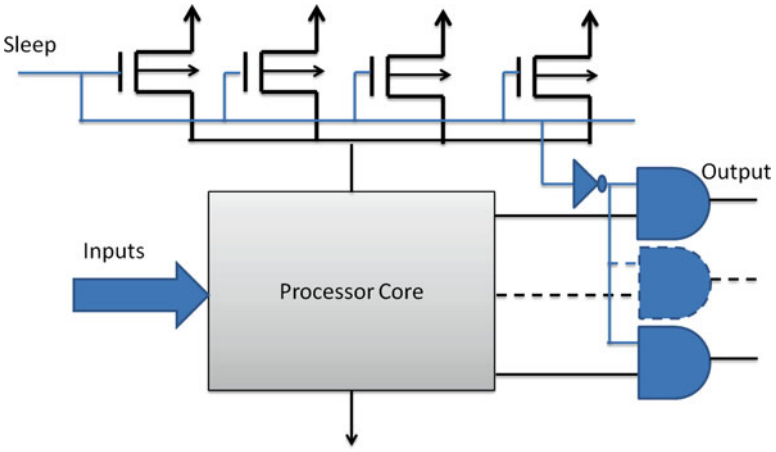


**Fig. 7** Circuit for power gating

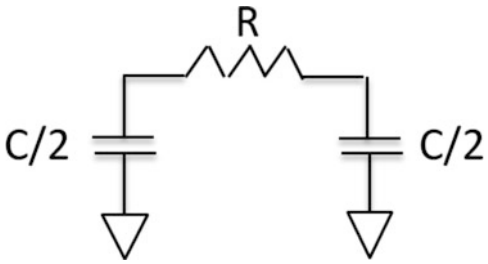


**Fig. 8** $\pi$ RC model for on-chip interconnect

frequency scaling (DVFS) is a well established technique and has been shown very effective to reduce the power consumed by a modern processor [23].

## *2.2 Interconnect Centric Design: The NoC Era*

The increasing number of devices has created routability and electromigration problems. To cope with the problem, the semiconductor foundries have increased the number of metalization levels, proposed new interconnect architectures and developed interconnect prediction models.

In nanometer regime, interconnect capacitances have started to be several orders of magnitude larger than the intrinsic capacitances of the CMOS transistors. There exist several ways to model the on-chip wires [24]. The most widely used model is the RC wire depicted in Fig. 8. The model can be used to estimate the propagation delay, power consumption, and so forth. It can also be used to optimize the wire under a given delay budget (buffer insertion, wire resizing, and so forth).

Using the RC wire model shown in Fig. 8, ITRS predicted that delay for intermediate wire will increase as shown in Fig. 9.

Both the interconnection delay and the coupling capacitance for the intermediate wire is experiencing an increase. The interconnect model used by ITRS to predict the coupling coefficient is depicted in Fig. 10. A 2-D static field solver has been used to determine the values for the self and coupling capacitances. The model takes into account the fringing capacitance.

The ITRS predicts that the coupling ratio between the intermediate wires will remain constant as depicted in Fig. 11.

In [25], the authors showed using a first-order model that the dynamic power dissipation for a loaded inverter is the sum of the intrinsic power consumption (power consumed by the inverter) and the power required to send the voltage across a global wire. Several techniques to reduce the signaling power consumption have been proposed (low-voltage swing, current mode, multi-level signaling, etc.) [24, 26–28].

The coupling capacitance has created a new design problems. The widely reported issues associated with the capacitive couplings are timing closure, crosstalk noise, glitches, and power consumption. The impact of the crosstalk on the physical parameters of the circuit depends on the alignment of the aggressors and the

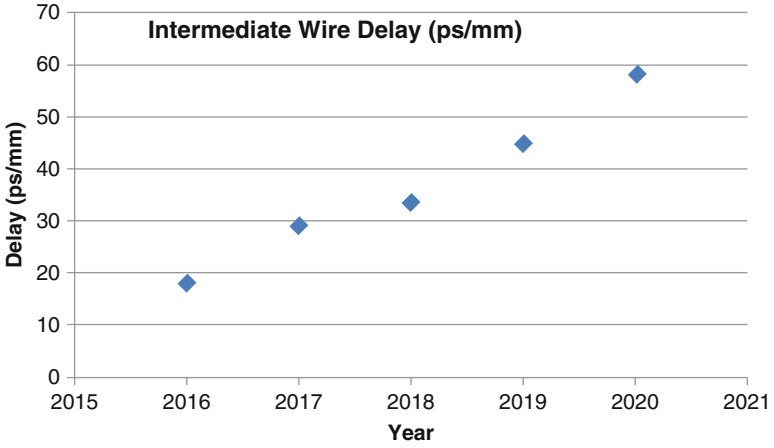


**Fig. 9** Projected intermediate wire delay (ps/mm)

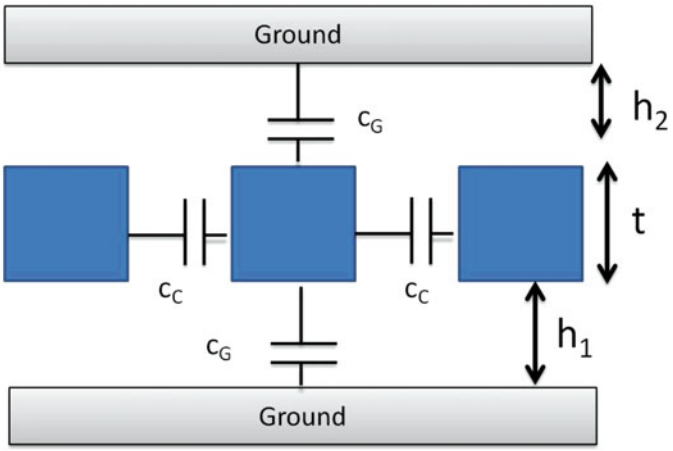


**Fig. 10** Interconnect model used by ITRS to compute self and coupling capacitance

victims [24]. The worst case occurs when the aggressor and the victim switch in the opposite direction as exemplified in Fig. 12.

In [29], the author presented a time-dependent power estimation model for capacitively coupled circuits. Figure 13 quantifies the impact of the crosstalk capacitance on the dynamic power consumption. The dynamic power has been obtained through spice simulation. The curves clearly show that the maximum power is consumed when the aggressor and the victim switch in the opposite direction.

To overcome the bottleneck caused by wires and to sustain the performance improvements, Networks-On-Chip (NoC) has emerged as a new design paradigm for SoC. NoC motivations come from many design corners such as the migration towards multi-core implementation (many core architecture), the growing silicon

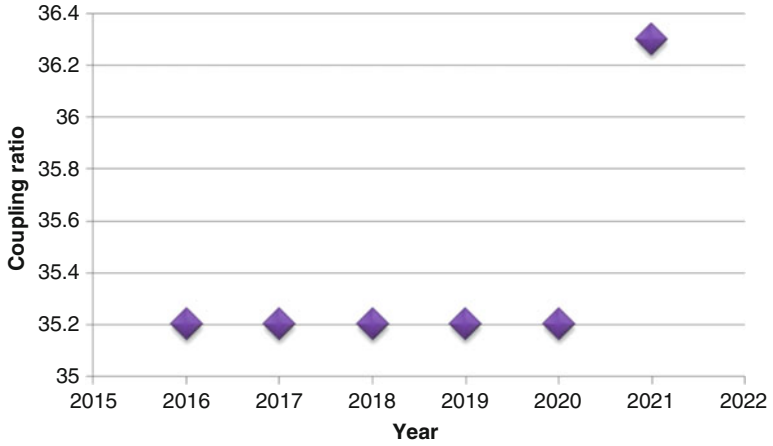


**Fig. 11** Coupling ratio for intermediate and global wires

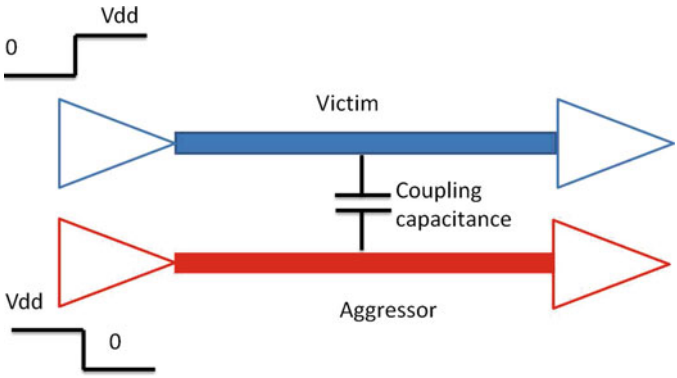


**Fig. 12** Worst case scenario for the impact of crosstalk noise

area devoted to routings, and the difficulties to implement point-to-point communication among cores or Processing Elements (PEs) [30]. Figure 14 depicts a 3x3 2D NoC architecture. Its worth pointing out that there are plenty of NoC architectures beyond the 2D mesh network.

The growing number of cores has created design bottlenecks for the NoC architecture. The widely reported design challenges are the high communication latency for distant cores and the extra power consumed by the switches [31]. Several network and communication technologies have been reported in the literature. Millimeter-wave NoC architecture using a hierarchical small-world networks proposed in [32] is a promising solution that solves some the issues related to the NoC design using multi-hope network.

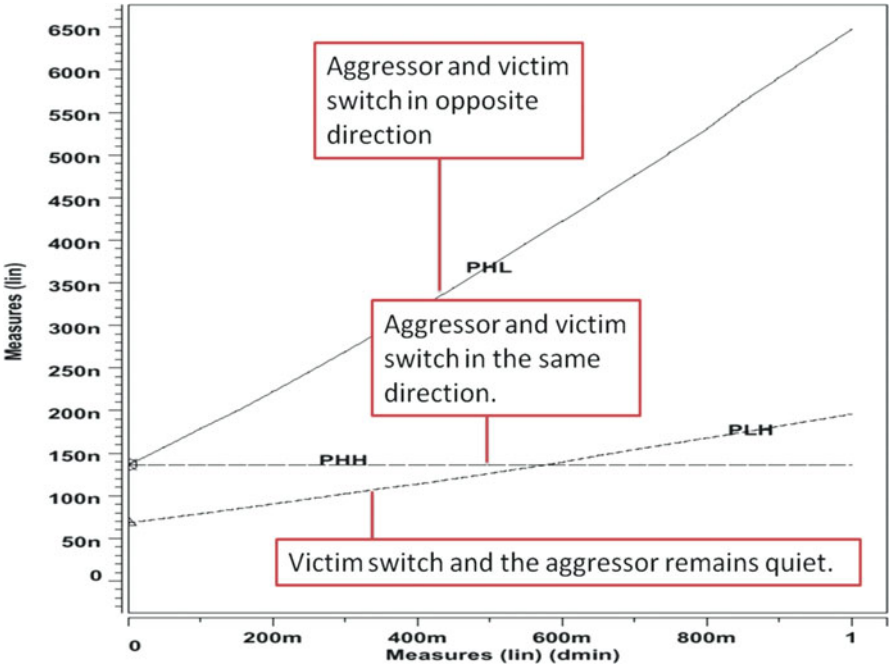


**Fig. 13** Impact of the coupling capacitance on the dynamic power consumption

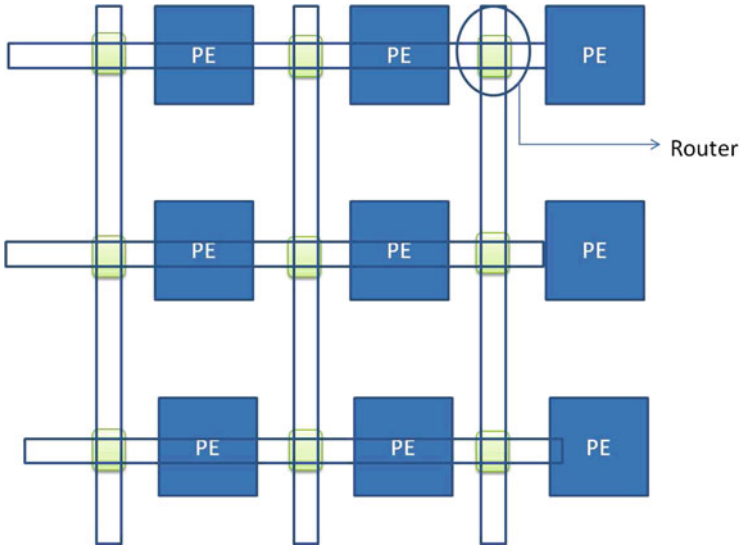


**Fig. 14** 3x3 2D mesh network

## 2.3 *Dark Silicon Era*

As a result of the aggressive transistor scaling and the ever increasing number of transistor density, voltage scaling can no more follow the Dennard's scaling law. This phenomena has led to a steady increase in power density. To prevent the processor from overheating, it has been observed that the number of powered transistors need to be shrunk using for example power gating. As a result, for a given period of time a percentage of the transistors is dark. This effect is referred to as *the dark silicon* [33]. The amount of the dark silicon is expected to escalate as the transistor feature sizes decrease (cf. Fig. 15.)

The dark silicon has emerged as a new design challenges that prevent circuit designers from increasing processor performance. As a results, new design methodologies and solutions should be proposed to reduce the dark silicon area.

In [34], the authors developed scaling models for device, core, and multicore. The models have been used to (1) study the percentage of dark silicon for future nodes, (2) develop the Pareto frontiers, and (3) investigate the impact of various architecture parameters on the processor's performance in the dark silicon era. The authors have considered the following parameters: (1) memory bandwidth, cache size, simultaneous multi-threading (SMT), and parallelism. Table 1 summarizes the reported results.

The table clearly shows that the parallelism is one important design consideration that can alleviate the impact of the dark silicon. For the 5G mobile communication system, most of the DSP algorithms can easily be parallelized using the MIMO system (Multiple Input Multiple Output) instead of the SISO system (Single Input Single Output) [35] (cf. Fig. 16). Unfolding techniques can also be applied to increase the level of parallelism for the DSP algorithms.

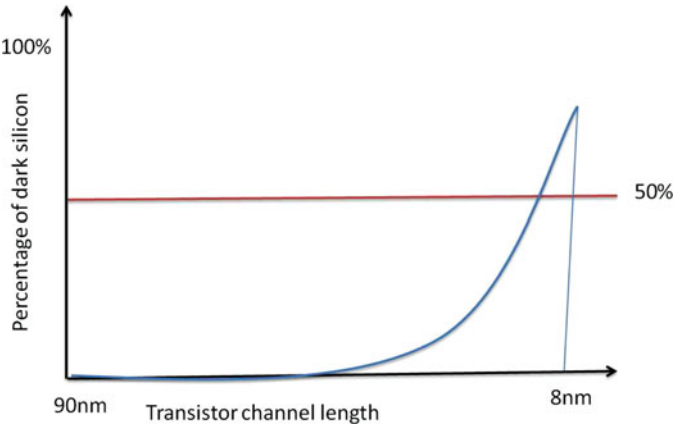


**Fig. 15** Growing problem of dark silicon area [34]

**Table 1** Impact of architecture parameters on processor performance

| Parameter | Maximum improvements |
|---|---|
| Cache area | 20 % |
| Memory bandwidth | 2× |
| 2-way SMT | 1.5× |
| 8-way SMT | 0.3–2.5× |
| Parallelism | 6.3× |

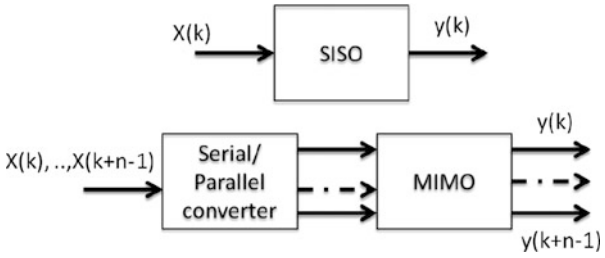


**Fig. 16** SISO to MIMO transformations for DSP algorithms

The systolic architecture is yet another design technique that can be used to implement the DSP algorithm using concurrent and identical processing elements (PEs). The dependence graph is used to determine the number of PEs [35, 36]. Contrary to the MPSOC, the Amdahl's law is inapplicable to the systolic processor. This suggest that the prediction reported in [34] can be too pessimistic for the DSP systems implemented using systolic array.

In [37], the authors firstly analyzed the source of energy dissipation in CPU. Secondly, they defined the utilization wall. Finally they proposed the deployment of energy saving cores (c-cores) to reduce the ratio energy/instruction.

In [38], the authors showed the advantage of the heterogeneous cores (processor) over homogeneous ones. They devised an algorithm to determine the optimal number of the heterogeneous cores for a given power budget and dark silicon constraints. The authors showed that the synthesized heterogeneous multi-cores can lead to up-to 60 % improvement in performance compared to the homogeneous design.

As the NoC has become a design solution to address the growing problem of on-chip communication, it has been observed that the NoC circuitry can consume a significant portion of the energy. The highest reported value is 33 % which is consumed by Intel's cloud computing processor [23]. To bring this energy down, many published reports considered schemes based on dynamic voltage and frequency scaling (DVFS) [39, 40]. However, in the dark silicon era, DVFS scheme has become very ineffective because of the various factors such as the growing leakage power dissipation. In [41], the authors devised multi-layer NoC routers. Each layer has a unique voltage and frequency pair, referred to as a VF level. The levels have been defined for a target CMOS technology node. This scheme sacrifices silicon area to reduce the percentage of the dark silicon.

## 3 FinFeT for Future SoC Designs

It is well documented that the scaling process for the bulk CMOS is reaching a dead end beyond 22 nm [42, 43]. This fact is attributed primarily to the short-channel effect and leakage current. In 70 nm node technology, it has been reported that the active mode leakage current is responsible for over 40 % of the total power consumption [21].

To address the shortcomings of the bulk CMOS technology and to enable the progress of technology scaling, the FinFET has been proposed as a substitute for the bulk CMOS transistor. It is widely reported that the FinFET overcomes the problem associated with short-channel effect, offers higher current drive, and dissipates less standby current compared to the bulk MOSFET technology [44].

The FinFET transistor is a MOSFET with double gates. There exist several competing realizations for the double-gated transistor. In [45] double-gated transistor on SOI, independent double-gate FinFET and planar double-gate SOI were reported. Figure 17 shows these three structures.

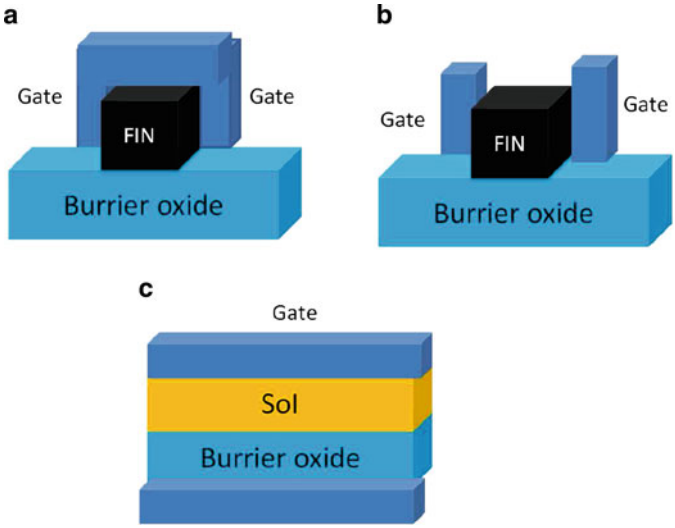


**Fig. 17** FinFET structure:(**a**) shorted-gate FinFET, (**b**) independent double-gate FinFET, (**c**) planar double-gate FinFET on SOI

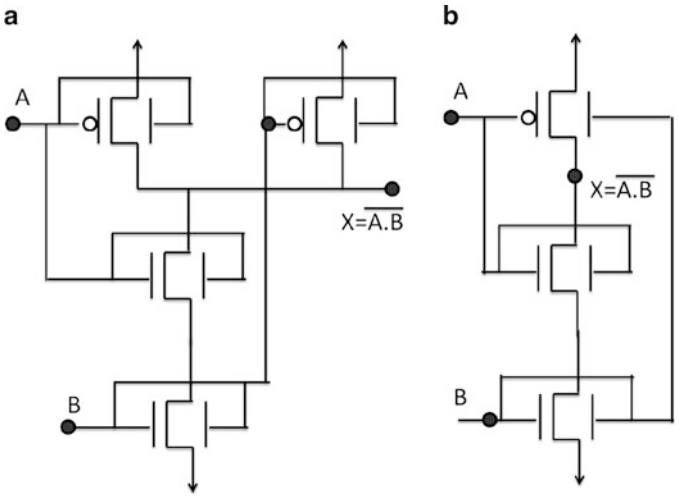


**Fig. 18** A 2-input NAND logic gate built using: (**a**) a shorted-gate FinFET, and (**b**) independent gate FinFET transistor

## *3.1 Design Techniques Using FinFET Transistor*

The primitive gates built using multi-gate MOSFET is fundamentally different from the bulk CMOS. Figure 18 shows a 2-input NAND gate built using, respectively, shorted-gate and independent-gate FinFET transistor [46].

From Fig. 18, we can see that (1) the IG-FinFET consumes less transistor than the SG-FinFET, and (2) the SG-FinFET transistor can be fabricated easily using the bulk CMOS process.

The results on the ISCAS benchmark reported in [47] show that the IG-FinFET consumes less power but slower than the SG-FinFET. Figure 19 shows a low-power NAND gate built using FinFET and IG-FinFET [47]. The supply voltages $V_L$ and $V_H$ are used to reverse-bias the IG-FinFET transistor.

The FinFET transistor did not completely solved the leakage current problem. In [47], it is reported that the leakage power consumption for the ISCAS'85 benchmark represents on average 33.1 % of the total power consumption. The authors presented several schemes to reduce the leakage power dissipation under delay or relaxed delay constraints.

The emerging FinFET technology for the manufacturing of future integrated circuits raised a deep concern for the projection made for the dark silicon area reported in [33]. The authors of [48] studied the dark-silicon problem using the FinFET transistor. The study was performed for the 7 nm FinFET technology. Two types of processor cores were used in the study. These processor cores are the Nehalem out-of-order and LEON3. The reported results show that the 50 % dark silicon prediction reported in [33] is very pessimistic. They further demonstrated that the near-threshold voltage regime is a powerful way to reduce the percentage of dark-silicon.

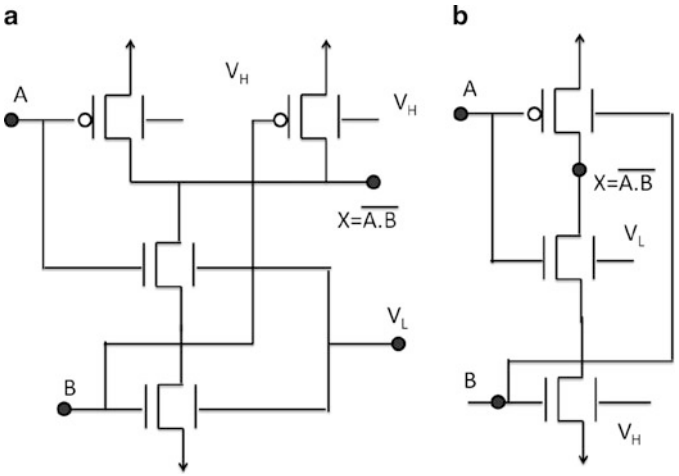


**Fig. 19** A low-power NAND gate built using: (**a**) Low-power FinFET, (**b**) Low-power IG-FinFET [47]

## 4 RF Design Using FinFET Transistor

For the RF design, it has been reported in [49] that the FinFET outperforms the bulk CMOS for frequency below 5 GHz. However, for high frequency (millimeter wave band) the authors found that the FinFET suffers from low drive current. The authors of [50], concluded that for the RF frequency higher than 10 GHz, the bulk CMOS is better fitted to design RF blocks.

In recent years, several reports showed the potential of FinFET for the construction of RF circuits in millimeter wave band [51, 52].

## 5 Sensors for 5G Device

In modern days, thanks to the fast development of the integration technology, sensors have become widespread, ubiquitous and low-cost. In addition, the sensor technology has been subjected to conspicuous development over time. There exist several technologies to fabricate sensors. To determine the preponderant technology, we conducted extensive web search for the up-to-date commercial sensors. The results are presented in Table 2. The table shows that the MEMS is the prevalent technology for sensor manufacturing.

**Table 2** Sensor technology and applications

| Sensor type | Application | Technology |
|---|---|---|
| Mechanical | Pressure sensing | MEMS |
| Thermal | Air conditioning | MEMS |
| Biological | Blood glucose | MEMS |
| Chemical | Environmental monitoring | MEMS |
| Optic | Ambient light detection | Silicon |
| magnetic | Gyroscope | Hall IC |

## 6 Microelectromechanical System

Microsystem or Microelectromechanical System or MEMS for short is a system that has both electronic and non-electronic devices [53]. MEMS contains the following blocks:

1. A Transducer to convert physical quantities into electrical signal and vice-versa.
2. An actuator and sensor for action.
3. A controller which is an integrated circuit devised to take decision.

As opposed to the integrated circuits, MEMS is diverse in fabrication, functionalities, and applications. In other words, MEMS fabrication is not confined to a certain material and steps. What makes MEMS attractive for sensor is it is size, fidelity and yield. Furthermore, the current success of MEMS is credited to abundant of factors, such as its similarity with the integrated circuit, low-cost, and the existing of CAD tools.

Contemporary smart phones and tablets contain a bunch of sensors fabricated using MEMS (gyroscope, accelerometer, and microphone). There is also a growing interest to build RF front-end using MEMS (RF-MEMS). To develop smart applications, future smart phones will integrate electronic nose. The nose will be responsible for detecting volatile compound.

Using the mobile Internet device as a driving force for the development of MEMS, a road map has been determined for the following devices: accelerometers, gyroscopes, integration path for inertial measurement units, microphone, RF MEMS, and electronic nose.

## 7 Summary

Future wireless transceiver will be implemented using MtM technology. In this technology, the number of sensors is expected to increase at a rate proportional to the Moore scaling paradigm. For the base-band processor, the new design challenge is the increasing portion of the dark silicon area. In this chapter, we discussed potential solutions to reduce the impact of dark silicon and we proposed a multitude of architecture solutions. These solutions are spread across all the design abstraction

levels (system, algorithm, circuit and device). Multi-gate FET transistor has plenty of advantages compared to the bulk CMOS for the design of both RF front-end and base-band processor in the technology below 20 nm.

**Acknowledgements** The authors deeply acknowledge Dr. Mehrdad Dianati for his support and interest that has shown to the chapter.

## References

1. I.B. Dhaou, An electronic system to combat drifting and traffic noises on Saudi roads, in *Proceeding of IV* (2012), pp. 217–222
2. F. Boccardi, R.W. Heath Jr., A. Lozano, T.L. Marzetta, P. Popovski, Five disruptive technology directions for 5G. IEEE Commun. Mag. **2**(52), 74–80 (2014)
3. K.E. Skouby, P. Lynggaard, Smart home and smart city solutions enabled by 5G, IoT, AAI and CoT services, in *Proceedings of IC3I* (2014), pp. 874–878
4. J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, J. Zhang, What will 5G be? IEEE J. Select. Areas Commun. **32**(6), 1065–1082 (2014)
5. T.H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, 2nd edn. (Cambridge University Press, Cambridge, 2003)
6. P.B. Kenington, *RF and Baseband Techniques for Software Defined Radio* (Artech House, Boston, 2005)
7. G. Fettweis, M. Lohning, D. Petrovic, M. Windisch, P. Zillmann, W. Rave, Dirty RF: a new paradigm, in *Proceeding of PIMRC*, vol. 4 (2005), pp. 2347–2355
8. W. Namgoong, T.H. Meng, Direct-conversion RF receiver design. IEEE Trans. Commun. **49**(3), 518–529 (2002)
9. L. Noor, A. Anpalagan, Direct conversion receiver for radio communication systems. IEEE Potentials **24**(5), 32–35 (2005)
10. T. Jiang, Y. Wu, An overview: peak-to-average power ratio reduction techniques for OFDM signals. IEEE Trans. Broadcast. **54**(2), 257–268 (2008)
11. I.B. Dhaou, C.C. Logothetis, H. Tehnunen, On the robustness and performance tradeoffs for OFDM channel estimation, in *Proceedings of the 13th International Conference On Wireless Communication*, vol. 1 (2001), pp. 41–48
12. D. Tandur, M. Moonen, Compensation of RF impairments in MIMO OFDM systems, in *Proceeding of ICASSP* (2008), pp. 3097–3100
13. I.B. Dhaou, Client-server network architecture for safe pilgrim journey in the Kingdom of Saudi Arabia, in *Proceeding of IV* (2010), pp. 1043–1048
14. W.Z. Khan, X. Yang , M.Y. Aalsalem, Q. Arshad, Mobile phone sensing systems: a survey. Commun. Surv. Tutor. **15**(1), 402–427 (2013)
15. R.K. Cavin, P. Lugli, V.V. Zhirnov, Science and engineering beyond moore's law. Proc. IEEE **100**, 1720–1749 (2012)
16. K. Roy, J. Byunghoo, A. Raghunathan, Integrated systems in the more-than-moore era: designing low-cost energy-efficient systems using heterogeneous components, in *Proceeding of VLSID* (2010), pp. 464–469
17. W. Arden, M. Brillout, P. Cogez, M. Graef, B. Huizing, R. Mahnkopf More-than-moore white paper, International Roadmap for Semiconductor, ITRS, 2010
18. G.Q. Zhang, A.J. van Roosmalen (eds.), *More than Moore Creating High Value Micro/Nanoelectronics Systems* (Springer, New York, 2009)
19. N. Weste, D. Harris, *Principles of CMOS VLSI Design*, 4th edn. (Pearson, Boston, 2011)
20. [Online]. Available: http://ptm.asu.edu

21. J. Kao, S. Narendra, A. Chandrakasan, Subthreshold, leakage modeling and reduction techniques, in *Proceeding of ICCAD* (2002), pp. 141–148
22. J. Rabaey, *Low Power Design Essentials* (Springer, New York, 2009)
23. J. Howard et al., A 48-Core IA-32 processor in 45 nm CMOS using on-die message-passing and DVFS for performance and power scaling. IEEE J. Solid-State Circuits **46**(1), 173–183 (2011)
24. I.B. Dhaou, Efficient interconnect modeling for SoC systems with application to interconnection delay estimation, Technical Report, TUCS 2003
25. I.B. Dhaou, H. Tenhunen, V. Sundararajan, K.K. Parhi, Energy efficient signaling in deep-submicron technology, in *Proceeding of ISCAS*, vol. 5 (2001), pp. 411–414
26. I.B. Dhaou, Low-power design techniques in deep-submicron with application to wireless transceiver design, Ph.D. Dissertation, Royal Institute of Technology, Sweden, 2002
27. E. Mensink, D. Schinkel, E.A.M. Klumperink, E. van Tuijl, B. Nauta, Power efficient gigabit communication over capacitively driven RC-limited on-chip interconnects. IEEE J. Solid-State Circuits **45**(2), 447–457 (2010)
28. J. Nurmi, H. Tenhunen, J. Isoaho, A. Jantsch (eds.), *Interconnect-Centric Design for Advanced SoC and NoC* (Springer, New York, 2005)
29. D. Khalil, D. Sinha, H. Zhou, Y. Ismail, A timing-dependent power estimation framework considering coupling. IEEE Trans. Very Large Scale Integr. VLSI Syst. **17**(6), 843–847 (2009)
30. H. Tenhunen, A. Jantsch (eds.), *Networks on Chip* (Springer, New York, 2003)
31. R. Marculescu, U. Ogras, L. Peh, N. Jerger, Y. Hoskote, Outstanding research problems in NoC design: system, microarchitecture, and circuit Perspectives. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **28**(1), 3–21 (2009)
32. S. Deb, K. Chang, X. Yu, S.P. Sah, M. Cosic, A. Ganguly, P.P Pande, B. Belzer, D. Heo, Design of an energy-efficient CMOS-compatible NoC architecture with millimeter-wave wireless interconnects. IEEE Trans. Comput. **62**(12), 2382–2396 (2013)
33. H. Esmaeilzadeh et al., Dark silicon and the end of multicore scaling, in *Proceeding of 38th International Symposium on Computer Architecture (ISCA)* (2011), pp. 365–376
34. H. Esameilzadeh, E. Blem, R.St. Amant, K. Sankaralingam, D. Burger, Power limitation and dark silicon challenge the future of multicore. ACM Trans. Comput. Syst. **30**(3), 11:1–11:27, August 2012
35. K.K. Parhi, *VLSI Digital Signal Processing Systems: Design and Implementation* (Wiley Interscience, New York, 1999)
36. N. Petkov, *Systolic Parallel Processing* (North Holland Publishing, New York, 1992)
37. N.G. Hotta, J. Sampson, Q. Zheng, V. Bhatt, J. Auricchop, GreenDroid: an architecture for the dark silicon age, in *Proceeding of ASP-DAC* (2012), pp. 100–105
38. Y. Turkhia, B. Raghunathan, S. Garg, D. Marculescu, HaDeS: architectural synthesis for heterogeneous dark silicon chip multi-processors, in *Proceeding of DAC* (2013), pp. 1–7
39. R. David, P. Bogdan, R. Marculescu, U. Ogras, Dynamic power management of voltage-frequency island partitioned networks-on-chip using intel's single-chip cloud computer, in *Proceeding of NoCS* (2011), pp. 257–258
40. M.K. Yadav, M.R. Casu, M. Zamboni, A simple DVFS controller for a NoC switch, in *Proceeding of PRIME* (2012), pp. 131–134
41. H. Bokhari, H. Javaid, M. Shafique, J. Henkel, darkNoC: designing energy-efficient network-on-chip with multi-vt cells for dark silicon, in *Proceeding of DAC* (2014), pp. 1–6
42. D. Frank, R. Dennard, E. Nowak, P. Solomon, Y. Taur, H.-S. P. Wong, Device scaling limits of Si MOSFETs and their application dependencies. Proc. IEEE **89**(3), 259–288 (2001)
43. Y. Bin et al., FinFET scaling to 10 nm gate length, in *Proceeding of IEDM* (2002), pp. 251–254
44. T.-J. King, FinFETs for nanoscale CMOS digital integrated circuits, in *Proceeding of ICCAD* (2005), pp. 207–210
45. Y.S. Chauhan, D.D. Lu, V. Sriramkumar, S. Khandelwal, J.P. Duarte, N. Payvadosi, A. Niknejad, Ch. Hu, *FinFET Modeling for IC Simulation and Design: Using the BSIM-CMG Standard* (Academic, London, 2015)
46. D. Chen, N.K. Jha (eds.), *Nanoelectronic Circuit Design* (Springer, New York, 2011)

47. A. Mutterja, N. Agarwal, N.K. Jha, CMOS logic design with independent gate FinFETs, in *Proceeding of ICCAD* (2007), pp. 560–567
48. A. Shafaei, Y. Wang, S. Ramadurgam, Y. Xue, P. Bogdan, M. Pedram, Analyzing the dark silicon phenomenon in a many-core chip multi-processor under deeply-scaled process technologies, in *Proceeding of GLSVLSI* (2015), pp. 127–132
49. V. Subramanian, B. Parvais, J. Borremans, A. Mercha, D. Linten, P. Wambacq, J. Loo, M. Dehan, C. Gustin, N. Collaert, S. Kubicek, R. Lander, J. Hooker, F. Cubaynes, S. Donnay, M. Jurczak, G. Groeseneken, W. Sansen, S. Decoutere, Planar bulk MOSFET versus FinFETs: an analog/RF perspective, IEEE Trans. Electron Devices **53**(12), 3071–3079 (2006)
50. M. Dehan et al., Perspectives of (sub-) 32 nm CMOS for analog/RF and mm-wave applications, in *Proceeding of EuMIC*, 27–28 October 2008, pp. 103–106
51. S.K. Lin, J.L. Kuo, W. Huei, A 60 GHz sub-harmonic resistive FET mixer using 0.13 μm CMOS technology. IEEE Microw. Compon. Lett. **21**(10), 562–564 (2011)
52. S.C. Woo et al., Device design guidelines for nanoscale FinFETs in RF/analog applications. IEEE Electron Device Lett. **33**(9), 1234–1236 (2012)
53. S.E. Lyshevski, *MEMS and NEMS: Systems, Devices, and Structures* (CRC Press, Boca Raton, 2001)

# Part III
# Networking Techniques and Applications for 5G Networks

# Ultra-Dense Network Architecture and Technologies for 5G

**Shanzhi Chen, Fei Qin, Bo Hu, Xi Li, and Jiamin Liu**

**Abstract** In recent years, with the growing popularity of smart device, our daily life has come to revolve around with spectacularly successful mobile Internet services, which lead to the explosion of data traffic in mobile communication networks. The requirement on communication networks has become a critical issue. By 2020, the global mobile traffic volume will have about 1000 times growth compared to that of 2010. Recent research on 5G requirements indicates that the traffic density in crowded city or hotspot area will reach 20 Tbps/Km$^2$. Ultra dense network(UDN) has been introduced to meet the traffic capacity requirement of 5G. as a most promising method. Challenges, network architectures, key technologies will be discussed in this section.

## 1 Introduction

In recent years, with the growing popularity of smart device, our daily life has come to revolve around with spectacularly successful mobile Internet services, which lead to the explosion of data traffic in mobile communication networks. The requirement on communication networks has become a critical issue. By 2020, the global mobile traffic volume will have about 1000 times growth compared to that of 2010 [1, 2]. According to the history of 4G and the former generations, the next generation of mobile network, i.e. 5G which is towards 2020 and beyond, would reach 10 Gbps, which is 10 times compared to the peak data rate of 4G. Recent research on 5G requirements indicates that the traffic density in crowded city or hotspot area will

This work was supported in parts by the National Natural Science Foundation of China for Distinguished Young Scholar under Grant 61425012.

S. Chen (✉) • F. Qin • J. Liu
China Academy of Telecommunications Technology (CATT) and Datang Telecom Technology & Industry Group, Beijing, China
e-mail: chensz@datanggroup.cn

B. Hu • X. Li
Beijing University of Posts and Telecommunications (BUPT), Beijing, China



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_15

reach 20 Tbps/km$^2$ [3]. In addition, other requirements such as lower latency, higher spectral efficiency and energy efficiency are also included.

There are usually three ways to improve throughput in wireless system: (1) Improving spectral efficiency through new coding and modulation technologies; (2) Increasing spectrum bandwidth; (3) cell splitting to improve spectrum reuse density. In [4], it is observed that, by 2008, the wireless capacity has increased 1 million times from 1957. Among these gains, 25 times improvements came from wider spectrum and 25 times improvements were contributed by advanced air interface design, and 1600 times improvements were due to the reduction in cell sizes and transmit distance. From technique evolving viewpoint, the potential gain derived from the advance of wireless modulation and multi-antenna technologies, is approaching to the upper bound. In LTE-A, the theoretical peak spectral efficiency has reached 30 bps/Hz through 8-layer spatial multiplexing. This value nearly reaches the maximum of typical wireless transmitting technology. Secondly, the spectrum resource is limited. The continuous development of radio services such as satellite, broadcasting, scientist, fixed and mobile terminal result that spectrum become very rare resources. The International Mobile Telecommunication (IMT) forecast that requirement for spectrum in year 2020 will reach 1490–1810 MHz in China [5]. It is a big challenge to identify and allocate enough spectrum resource to IMT system.

Based upon the above analysis, it is hard to meet 5G traffic capacity requirement through increasing spectral efficiency and allocating more spectrum bands. Therefore, we can conclude that the increasing of the wireless access point (AP) density with smaller coverage is the most efficient way to improve the system capacity, especially in hotspot scenarios [6]. China promoted LTE-Hi [7] to increase the system throughput for hotspot areas. LTE-Hi is a Long Term Evolution (LTE) based hotspot solution, which aims to provide higher performance in hotspot scenarios by exploiting wider bandwidth in higher frequency. In ITU-R Report M.2320 [8], Ultra Dense Network (UDN) is promoted as one of the technology trends to meet the high throughput requirements of 5G. And METIS [9] also take the Ultra Density Network as one of the most important topics towards the mobile system for 2020 and beyond. On the other hand, Since Bell Lab proposed the concept of cellular network in 1970s, network architecture has been adopted in the successive generations of mobile networks. For architecture evolution perspective, authors in [10] point out that the macro cellular and local small cell coexistence architecture in 5G will replace the macro cellular-dominated architecture.

Typical scenarios of UDN include: office, apartment, open-air gathering, stadium, subways and railway station [3]. The common requirements in these scenarios are huge numbers of connections, high density network traffic and high data rate. To meet these requirements, the AP(s) need to be deployed densely, with minimum Inter-Site Distances (ISDs) of tens meters or below, i.e. one or more APs per room for indoor scenarios, and one AP on each lamp post for outdoor scenarios.

# 2 Concept and Challenges of UDN

## *2.1 What Is UDN*

The main difference between UDN and traditional cellular network is the APs density. In UDN, the radius of AP coverage is about 10 m and there are thousands of APs in 1 $km^2$. But in traditional cellular network, cell range is more than 500 m and usually less than 3–5 base stations (BSs) in 1 $km^2$. Correspondingly, only one or several terminals are connected to one UDN AP, whereas hundreds or even thousands of active users are resident in one macro cell. Table 1 gives the comparison between the UDN and traditional cellular network.

Another key point is that the type of APs in UDN is diversified. Small cell station, relay station, distributed Remote Radio Head (RRH) and user equipment (UE) itself can act as AP in UDN. However the macro BS in traditional cellular network is the dominate access point.

Besides the above features, higher spectrum and wider band, heterogeneous and irregular deployment, flexible backhaul, lower user mobility are also differences between UDN and traditional cellular networks.

In [9], METIS defines UDN as a stand-alone system that will be optimized for hotspot areas where the highest traffic increase will be observed. The concept of UDN includes radio access technologies, small cell integration/interaction and wireless backhauling. The authors in [11] suggested that UDN is a major technology to meet the requirements of ultra-high traffic volume density.

In general, UDN is a new wireless network solution for hotspot scenarios, to provide higher throughput and better user experience. In UDN, the AP density is comparable to or even higher than the user density. Different types of APs will tightly cooperate to achieve higher spectrum efficiency, lower power consumption and seamless mobility.

**Table 1** Comparison between UDN and the traditional cellular network

| ITEM | UDN | Traditional cellular network |
|---|---|---|
| Deployment Scenarios | Indoor, Hotspot | Wide coverage |
| AP density | More than 1000/$km^2$ | 3–5/$km^2$ |
| AP Coverage | Around 10 m | Several hundred meters and more |
| AP types | Small-Cell, Pico, Femto, UE relay, Relay | Macro/Micro BS |
| AP Backhaul | Ideal/Non-ideal, Wired/Wireless | Ideal, Wired |
| User Density | High | Low/Medium |
| User Mobility | Low mobility | High mobility |
| Traffic density | High | Low/Medium |
| Deployment | Heterogeneous, irregular | Single layer, regular cell |
| System Bandwidth | Hundreds of MHz | Tens of MHz |
| Spectrum | >3 GHz (up to mm Wave) | <3 GHz |

## 2.2 Challenges and Technical Directions of UDN

In order to provide very high traffic density and better user experience in UDN, there are many new challenges including network architecture, mobility management, interference management and others.

### 2.2.1 Challenges of Network Architecture

Traditional cellular network architectures are designed for wide area seamless coverage. Due to the differences of UDN and traditional cellular network discussed in Table 1, there are many problems to utilize traditional cellular network architecture (e.g. 4G architecture) in UDN.

- Signaling overhead and lengthy data path
  Too many functions such as service control and mobility control are centralized in Core Network (CN), which include Mobility Management Entity (MME) and Packet Data Network Gateway (PGW)/Serving Gateway (SGW). It is not efficient for UDN with high traffic throughput and ultra-dense APs deployment, and it leads to signaling overhead and lengthy data transmission path between APs and CN.
- Frequent handover
  The tight coupling of user plane and control plane over one air interface will result in frequent handover when the AP coverage is very small. It is not efficient and flexible in heterogeneous networking within both macro and UDN AP coverage.
- All distributed functions
  Higher layer process, Radio Resource Management (RRM), mobility management functions are distributed on each AP independently. In order to better support advanced interference management and resource management for UDN, the functions on each distributed AP needs to be centralized.
- Better user experience
  UDN is target to provide smooth handover and very high data rate for each user with ultra dense AP. The simple data gathering and transmission function of Local GateWay (LGW) cannot support better user experience. More functions are needed in LGW.

Therefore, a new architecture of UDN is needed to support high density AP deployment and flexible network management. In this new architecture, a local centralized user service center is necessary for knowing and measuring user's radio environment. Besides, the RRM and user service control center much closer to the user are required to provide better joint processing and Quality of Service (QoS) control, and a lower mobility anchor is also needed. Meanwhile, the CN functions should be simplified to provide only high level service to user.

### 2.2.2 Challenges of Mobility Management

With very small coverage and irregular network topology, mobility management in UDN is quite different from traditional cellular network.

In the scenarios that ultra-dense APs are deployed, the coverage radius of these APs is extremely small, i.e. only a few meters to tens meters. According to traditional handover judgment method, frequent handover may occur and the interruption probability of the user's experience of high data rate may increase. Moreover, from the perspective of the network side, it also means a high handover signaling overhead. How to solve the problem is essential for UDN [6, 12].

### 2.2.3 Challenges of Interference Management

Huge amount of APs may bring much higher throughput and better user experience, but may also lead problems. Interference management may directly impact the system performance [13]. Along with the multiplexing of resources for more access opportunities, interference increases as well and becomes more complex related to traditional cellular network. We need to solve the following problems:

- The ultra-density environment results in more interference sources. For example, in the crowd subway trains, lots of terminals and APs exist, and signals may have more reflected and scattering paths. Then transmission power should increase, which turns out more interference than before, however.
- Decreasing interference and increasing the resource utility become two parts in the contradiction and we need to find the proper tradeoff point.
- The existing parameters to evaluate interference impact such as interference threshold may not reflect the overall performance of the networks, and then more suitable parameters should be used to give a better indication between interference managing results and throughput, energy efficiency and other system level parameters.

Therefore, we should set up suitable interference models, analyze the typical wireless transmission scenario, and then propose effective interference control approach for UDN.

### 2.2.4 Challenges of Flexible Networking

With extreme densification and complex heterogeneous deployment, network planning and optimization will become difficult for UDN. It is critical to further enhance Self Organization Network (SON) to support flexible networking.

Huge amounts of APs in UDN make it more complex to realize self-configuration, self-optimization and self-healing.

Ultra high throughput, ultra low latency, ultra high reliability, massive connections are required to be provided in UDN. On another side, UDN is a very

complex network to cover both indoor and outdoor scenarios, with ideal backhaul AP deployment and non-ideal backhaul AP deployment. From view of the access technologies, there are 5G new access technologies, LTE access technologies, WLAN access technology working together. So, flexible network architecture with intelligent network sensing and automotive network optimization is very important to flexible networking and increasing spectrum efficiency in UDN.

Based on the above analysis, new network architecture for UDN is an important direction. Key technologies towards flexible networking, self-backhauling, multi-RAT coordination, advanced interference management, advanced mobility management and radio resource management are also essential for UDN.

In addition, due to different radio scenarios and higher spectrum and wider bandwidth, new transmission technologies and new air interfaces design are also need to further enhance the system performance for UDN. Millimeter wave communication is a very attractive direction for UDN [14], which can provide Gbps user experience and tens of Gbps throughput per AP. And it's also very suitable for wireless backhauling for UDN APs. Details of millimeter wave communication are discussed in other chapters of this book.

# 3 Network Architecture of UDN

## *3.1 Overview*

The topology of UDN toward 5G can be summarized with following key characteristics: (1) high density of mobile APs distributed randomly without network planning [15]; (2) various kinds of APs with different backhaul or fronthaul capabilities [16]; (3) heterogeneous networks with different overage ranges, different spectrum band and multiple radio access technologies (RATs) [14, 17].

Based on above characteristics, in order to provide high throughput and better user experience, some new principles and methods for network architecture design are introduced:

### 3.1.1 Localization and Flatter

It is a promising cost effective method to offloading the 1000 times of traffic within local UDN area. A flatter architecture is needed to support localization and decrease the cost of transmission. Localization of data path and controlling function are introduced by METIS and NGMN [18, 19] toward 5G UDN.

### 3.1.2 U/C Separation

New principle in [7] is the user plane and control plane separation. It's an effective way to provide user high data rate through density small cells but without mobility and connectivity experience degradation. This feature is supported in 3GPP Release 12 through dual/multi-node connections.

### 3.1.3 User-Centric

Toward 5G, UDN is quite different from traditional cellular network. It is very important to introduce user-centric principle [6], the network will organize the dense APs dynamically to serve the user to make it feel like always in the centre of a cell. User-centric ultra-dense network (UUDN) [6], including user-centric mobility management, user-centric clustering for data plane [20] and user-specific network controlling, can provide better user experience and higher spectrum efficiency.

### 3.1.4 C-RAN, Distributed-RAN and Flexible-RAN

Centralized Radio Access Network (RAN) and distributed RAN architectures for UDN are discussed in [15]. For centralized architecture, it can provide better joint processing and get higher spectrum efficiency, but need very high backhaul/fronthaul capability. For distributed architecture, it is more flexible for network deployment, but interference management is difficult and has lower spectrum efficiency. So a more flexible architecture is preferred to adaptively connect APs with different backhaul/front haul capability [18].

### 3.1.5 SDN & NFV

The new trends of Software Defined Network (SDN) and Network Function Virtualization (NFV) are also impacts the design of the network architecture for 5G [18, 21, 22]. With the decoupling of controlling and forwarding, software and hardware, the network architecture can be redesigned to define the logical functions and their interfaces but not the entities and their interfaces. This will also enabling even more flexible architecture towards UDN 5G.

In summary, the trends of UDN architecture are becoming more flatter, localized, U/C separation, user-centric, flexible and intelligent. Some examples toward small coverage density network are given in following subsections.

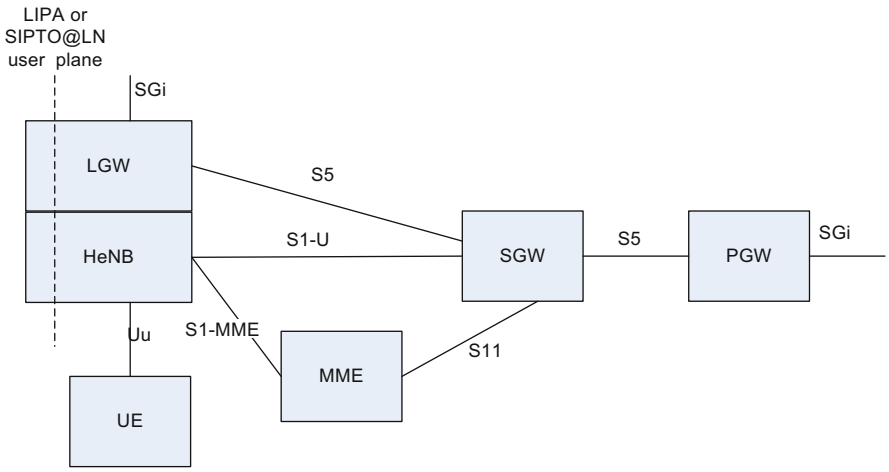


**Fig. 1** 4G LTE/SAE network architecture

## *3.2 GPP HeNB Architecture*

Home enhanced Node B (HeNB) is target for dense small coverage for hotspot area. A typical 3GPP architecture for HeNB under the 4G LTE/SAE framework, is shown in Fig. 1 [23, 24]:

In this architecture, the network side mainly includes five entities: HeNB, LGW MME, SGW and PGW. The function of each entity can be found in [15]. The function of HeNB is almost same as traditional macro BS except smaller coverage. And the user side control plane and data plane are managed by CN. This architecture also derives the localized features, which enable the function of Local Internet Packet Access (LIPA) and Selected IP Traffic Offload (SIPTO) for data offloading.

The HeNB architecture is easy for equipment development at early stage deployment with lower density of HeNBs. But when APs density increased, challenges of the architecture are as following:

(1) Many functions such as service control and mobility management are centralized in CN. It is not efficient with high traffic throughput and dense APs deployment. This will lead to extra signaling overhead and longer data transmission latency between APs and CN.
(2) The tight coupling of user plane and control plane at air interface will result in frequency handover for ultra dense network.
(3) The data transmission function of LGW cannot support user-centric service experience. More functions are needed to be collected and supported in LGW.

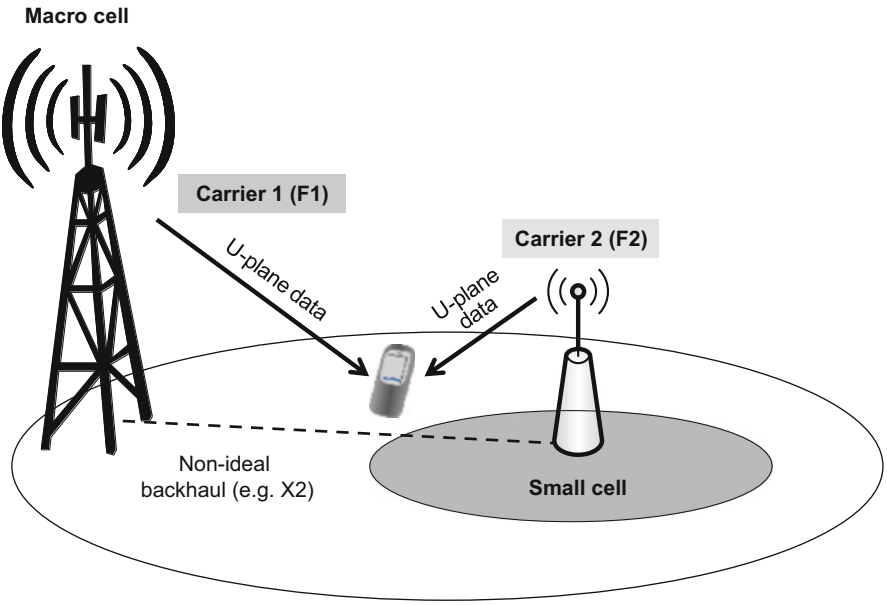


**Fig. 2** Inter-node radio resource aggregation

## *3.3 Small Cell Enhancement (SCE) Architecture*

To further support the heterogeneous network deployment scenarios, 3GPP developed the SCE architecture in release 12 [25].

A "dual connectivity" is used to refer to operation where a given UE consumes radio resources provided by at least two different network points connected with non-ideal backhaul. Dual connection can be done by aggregating radio resources in more than one eNB for user plane data transmission as illustrated in Fig. 2. By this solution, signaling overhead towards the CN can be potentially saved by keeping the mobility anchor in the macro cell.

In dual connectivity operation, a UE has only one RRC connection. With this principle, only the Master eNB (MeNB) generates the final RRC messages to be sent towards the UE. The MeNB and Secondary eNB (SeNB) will tightly coordinated to provide a UE dual User plane connections as illustrated in Fig. 3. The UE RRC entity sees all messages coming only from one entity (in the MeNB) and the UE only replies back to that entity.

Within these characteristics, the mobile anchor is moved to macro station, handover signaling overhead is reduced when the user is moving among small cells. But there are still many RRC signaling between high speed UEs and macro station during the movement for the SeNB updating procedure.

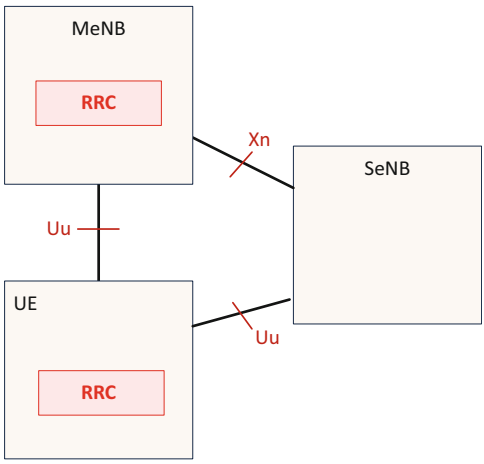


**Fig. 3** Radio interface C-plane architecture for dual connectivity

## *3.4 METIS UDN Architecture*

UDN is identified as one of the five Horizon Topics (HTs) of METIS project toward 5G, and also one of the four main technical enablers [7, 18]. Considering the trends of SDN and NFV, in order to support wide range of scenarios and use cases for 5G, METIS designed a general 5G architecture with flexible manners as shown in Fig. 4-1 in [18]. The architecture can be used for UDN by different function selection at different network domain as shown in Fig. 4-3 in [18].

Based on the separation of user plane and control plane, six functions are defined in this architecture. Synchronize radio control plane functions (SRC) and synchronize radio user plane functions (SRU) are deployed very close to end users. It supports Multi-APs and RRHs with ideal backhaul with centralized PHY/MAC, like C-RAN technologies.

Asynchronous radio control plane functions (ARC) and asynchronous radio user plane (ARU) together with CN user plane (Core-U) are located RAN side, which is much closer to user than traditional 3GPP SAE architecture. They can provide higher layer control functions like mobility management, QoS control, inter-RAT coordination, and also user plane function including local breakout.

CN control plane function (Core-C) and CN user plane functions (Core-U) are located in aggregation and CN side to provide high level management like Authentication, Authorization and Accounting (AAA) and roaming.

METIS UDN architecture based on the C-RAN, SDN, NFV technologies with principle of localized, U/C split concept, it's a flexible and efficient 5G UDN architecture.

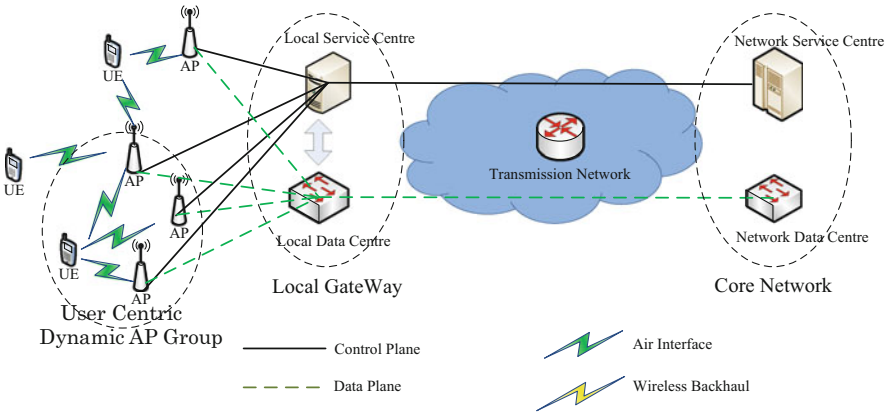


**Fig. 4** User centric architecture

## *3.5 User-Centric Architecture for UDN*

In [6], user-centric UDN (UUDN) architecture with the principles of user-centric, localized, U/C separation. It can be easily and flexible deployed with SDN and NFV manner as Fig. 4.

In this architecture, there are no more "cells" logically and physically from user's perspective. Dense APs in one area will be organized intelligently to follow user's movement and provide data transmission on-demand.

Four functional entities are introduced to provide user-centric services. At the radio side, local service center (LSC) and local data center (LDC) are introduced to provide the logical decoupling of control plane and user plane. All kinds of APs are connected to the LSC and LDC by various backhaul (idea/no-ideal, wire/wireless). At the core network side, network service center (NSC) and network data center (NDC) are introduced to provide the control and transmit functions. The LSC and LDC can be integrated into one Local Gateway as a physical entity, and NSC and NDC may also be integrated into one CN entity.

The functions and interfaces of each entity in UUDN are as following [6]:

(1) AP is the radio access channel for UE including the data plane and the control plane. APs can be built with RF, PHY, MAC and IP layer functions or combinations of them based on the backhaul capacity. If the AP has only RF, the PHY to IP layers will be centralized into LDC. With this architecture, LDC can provide joint processing in PHY layer. The advanced signal process can thus be used to avoid the inter-AP interference.
(2) LSC is the control service center to organize a dynamic APs Group (APG) to serve one user. It will have the new functions of user-centric RRM, multi-RAT coordination, effective QoS control, user-centric mobility management, and local radio link control.

**Table 2** Characteristics of the network architectures toward UDN

| | 3GPP HeNB architecture | 3GPP SCE architecture | METIS UDN architecture [18] | UUDN architecture [5] |
|---|---|---|---|---|
| Localization and flatter | LIPA & SIPTO | LIPA & SIPTO | Localized (User Plane and Part of Control Plane) | Localized (User Plane and Most of the Control Plane) |
| U/C separation | No | Yes | Yes | Yes |
| SDN & NFV based | No | No | Yes | Yes |
| Flexible Backhauling | No | No | Yes | Yes |
| User-Centric | No | No | No | Yes |

(3) LDC is the local data center to dealing with user data transmission. It will provide the user plane functions including higher layer process, dynamic AP channel processing. It also has the functions of Multi-AP coordination based and multi radio bear convergence for users.
(4) NSC is the network data center to provide the functions of user policy control, AAA, and high level mobility (roaming, inter NSC handover) etc.
(5) NDC is worked as packet data gateway at the network side.

With this architecture, LSC and LDC are very close to the location of APs, so that it is easy to provide user-centric service functions, advanced resource management, and interference management. It is more flexible for UUDN deployment with the decoupling of user plane and control plane, and decentralizing the CN functions to LSC and LDC. And signaling overhead and backhaul overhead can be greatly reduced for UDN scenarios.

### *3.6 Summary of Architecture*

The trends of UDN architecture are becoming more flatter, localized, U/C separation, user-centric, flexible and intelligent. A good UDN architecture will better support very dense APs work efficiently at complicated deployment scenarios. The characteristics of some typical architectures toward UDN are summarized in the following Table 2.

## 4 Key Technologies of UDN

Based on the new architecture, many key technologies can be introduced to provide high QoE, high area spectrum efficiency and low cost. The promising technologies are summarized below.

## *4.1 Flexible Networking*

Flexible networking technologies, as SON in 4G [26], are promising technologies to enable UDN in real deployment. With extreme densification, UDN will face huge challenges on self-installation, self-configuration of physical layer Identifier (ID), Automatic Neighbour Relations (ANR), Mobile Load Balancing (MLB) and etc.

### 4.1.1 Self-Installation

In super dense deployment of the network, the number of APs is very large. It is difficult for operators to configure and initialize each device manually, and then require that each AP can be plug and play [26]. Using the mechanism, the BSs automatically install in LTE, APs automatically start and stop initialization, including automatically get IP address, automatically establish the connection with the Operation Administration and Maintenance (OAM) server, automatically download the software, initial the configuration of wireless and transmission.

### 4.1.2 Physical Layer Cell ID and Self-Configuration

In UDN, hundreds or even thousands of APs will be deployed in the same area. The amount of AP IDs may not be enough. The types of AP are diverse, i.e. including mobile AP. Such AP needs to support self-configuration ID. In order to reduce the collision and confusion probability of AP IDs, there are two methods [26]. One is to provide sufficient number of IDs in the design of physical layer; the other is that adjacent AP can automatically avoid using the same ID through signal detection and coordination.

Method 1 is to provide enough number of the physical IDs (for example, thousands of IDs) in 5G system design. But increasing the number of physical IDs is proportionate to the complexity of the physical layer design of the system. It will lead to higher request numbers to the signal decoding ability of the terminal and increases equipment cost. Therefore, to design physical layer IDs for UDN system, trade off between ID amount and system complexity need to be considered.

When the number of physical ID increased, it is more difficult to allocate ID to each AP Manually. Therefore, physical ID self-configuration and self-adjustment mechanism are needed to avoid potential collision or confusion. Method 2 requires that each AP can listen to the signal on the radio to obtain physical layer IDs of neighbour APs to avoid the configuration of these IDs.

Due to the limitation of the transmitting power and coverage of AP, signal listening could not detect neighbour AP which is out the coverage range. Then there may be physical layer ID confusion. This can be coordinated through the signaling between APs, similar to the mechanism of LTE.

### 4.1.3 Automatic Neighbour Relations

In UDN, AP number is huge and changing and the neighbor relationship is more complex compared to LTE, which makes it unimaginable to establish and maintain AP's neighbour list manually. Therefore, it is essential to set up automatic self-configuration function to support neighbouring topological relations.

- Centralized mode
  In scenario with the macro BS coverage, the AP can adopt LTE ANR mechanism. APs can establish the neighbour relations and network interfaces through reading macro BS broadcast information, For APs under the coverage of the macro BS, the macro BS can use beam forming technology to receive and measure the specific pilot signal of each AP, which can calculate the direction angle and the approximate distance with respect to macro BS. Based on this information, macro BS can draw the network topology map of APs in its coverage region. Then based on the network topology map, macro BS can specify the neighbor list for each AP automatically.
- Distributed mode
  In distributed mode, each AP can receive the wireless signal of the neighbour APs directly, try to read the system broadcast information of them, get their global identity information and automatically create a neighbour relations between APs. In order to discovery more neighbor cells, new discovery signal design is need.

In other semi-distributed mode, dynamic cluster structure between APs is adopted. Each AP cluster will have a cluster head node. Neighbour relation map is maintained by the cluster head node.

### 4.1.4 Load Balancing

In UDN, the coverage of AP is smaller than traditional eNB. UE's serving AP will change fast. Load balancing among APs, for example resource load balancing, is very important in order to achieve the purpose of improving the system efficiency and user experience.

- Scenario with macro BS coverage
  In UDN with dual-connectivity mechanism, the control plane of the UE is carried by macro BS. According to the load condition of the APs, the RRM algorithms of macro BS can timely update UE's serving APs and achieve load balancing between APs.
- Scenario without macro BS coverage
  In localized structure, a centralized RRM can gather the users' measurements information offload conditions of each AP. And then it will select the appropriate AP resources to provide services for the UE and achieve load balancing between APs globally.

## *4.2 Backhauling*

Deployment scenario that all the APs have ideal wired backhaul linked to the network will not always exist. Hence the APs with wireless backhaul cannot be avoided in UDN [27].

In 3GPP Release 10, solutions related to relay node has been standardized [28]. But the solution of relay's wireless backhaul has some limitations. Firstly, 3GPP relay can only support one hop transmission, which means one AP can only select one BS with wired backhaul as donor eNB (DeNB). But in UDN scenarios, with the AP density increasing and coverage decreasing, one point is that it is not easy to find donor AP with idea backhaul in one hop coverage range [29]. Hence **multi-hop backhaul** can further enhance the flexibility of self-backhauling. Another point is that one AP can discover more than one UDN AP, sometimes macro BS, around it can also provide relay links. So **multi-connectivity backhauling** can provide much higher throughput to relay links. Secondly, 3GPP relay can only serve as transparent backhaul links for APs without wired link.

Meanwhile, more tight coordination between APs is preferred to provide joint transmission and joint beamforming. This requires more frequent signaling interaction and data forwarding among coordination APs group. **Wireless interfaces between APs** is a effective way to directly exchange information on the air and reduce interface delay from current delay (i.e. 20 ms for X2) to 1 ms level in UDN scenario [30].

### 4.2.1 Multi-Hop Backhaul Mechanisms

Enabling multi-hop makes the wireless backhaul mechanism more complexity. It is not easy to assign the donor APs for each wireless backhauling AP manually. Donor AP discovery and adaptive backhaul path detection and updating in UDN are the key issues.

The process of selecting the DeNB is usually completed in the OAM configuration with the process of measurement and access by the relay as a UE. The above pre-configuration process is not suitable for the APs in flexible deployment, and donor selection method of the current relay is so simple that it cannot cover all the wireless APs in UDN.

For the wireless path selection after the wireless APs are switched on, two kinds of the modes are included from the overall structures that are centralized mode and distributed mode.

Furthermore, due to the need of the nomadic deployment, the demand for energy saving in UDN, complex network environment and mobility of access node, the APs may be frequently switched on and off or removed, which results that the established path can no longer continue to provide services. Due to the changing of the traffic demand and link status, it is possible that the original wireless path cannot meet the requirements. So it is necessary to consider and design the process of backhaul path detection and updating.

- Donor AP discovery
  When an AP without wired link switch on, it need to firstly work as UE to discovery it's neighbor AP which can be served as it's donor AP. Discovery signal should be properly designed and assigned on the time and spectrum domain. It need to support enough coverage range but do not take too much overhead. After discover the AP, more system information and measurement information are also needed to help the wireless AP to select donor AP. These information including Reference Signal Received Power (RSRP), backhaul capacity, system load and so on can be broadcasted on the air.
- Wireless backhaul path detection
  Wireless backhaul path detection refers to the measurement of the link and transmission conditions for the current wireless backhaul path, and the gathering and processing of the measurement results, in order to form the assessment of the state of the current wireless link. When a wireless backhaul link cannot meet the requirements of the transmission or cannot provide services, it should be reported and feedback as soon as possible and then start the subsequent updating process.
- Wireless backhaul path updating
  The basic process of the wireless backhaul path updating is to destroy the old path and establish a new path, in which the method of the new path establishment is basically same as the process of the path establishment.

### 4.2.2 Multi-Connectivity Backhaul Mechanisms

Compared to legacy network, multi-connectivity backhaul can provide more abundant routing path selection [31]. When the quality of one path become poor, user data and signaling can be transmitted throughput other path. Furthermore, the total amount of traffic rate is the sum of all the paths when multiple paths are used at the same time. Multi-connectivity backhaul can enhance the reliability and increase the data rate of backhaul. In order to facilitate the management and coordination in a number of connections, it needs to have a primary connection, and the others are the secondary connections. The method of the primary connection decision is usually that the macro BS serves for the primary connection and the rest of the APs serve for the secondary connections, or the first access AP serves for the primary connection and the rest of the APs serve for the secondary connections. The primary connection will have most of the controlling functions and be easy for centralized control and management. For convenience, the secondary connections can also retain some control functions in order to control itself related transmission.

In the multi-path backhaul connections of the wireless APs, the air interface backhaul bearer of the wireless APs are complex, which are a aggregation of bearers of all UEs served by the wireless AP, i.e., the multiple UE bearers with the same characteristics converge into an air interface backhaul bearer of the wireless AP. Based on such complex mapping relations, when to split the wireless APs bearers, it needs to consider the unified processing in access network and to avoid the complexity in CN.

### 4.2.3 Wireless Interfaces Between APs

In the solution of 3GPP relay, relay node is equivalent to a cell belonging to its DeNB from the view of the neighbour nodes. The DeNB acts as a proxy gateway of the relay node and processes all kinds of network interface as an agent. This idea can still be used in UDN wireless backhaul scenarios with multi-connection/multi-hop backhaul link. But it needs to enhance the design due to the change of the network architecture and the increasing of the wireless backhaul connections. Furthermore, the reduced distances between APs in UDN also make the AP can transfer some of dynamic coordination messages each other through the air interface to enhance the whole system transmission efficiency.

## *4.3 Multi-RATs Coordination*

Along with the development of UDN, multi-access technologies coexistence is a long-term trend, especially 4G, 5G and WLAN. How to run and maintain multiple networks effectively, reduce maintenance costs and save energy, is a key issue to be resolved. Multi-RAT tight integration will become an important direction in the future wireless communication standardization fields.

How to coordinately use the various wireless access technologies and how to further improve the overall operational efficiency and user experience are a problem that need to be solved by multi-RAT integration technology [18]. Shown below for multi-RAT integration logic function schematic figure, multi-RAT, i.e. 5G/4G/WiFi, is converged through a centralized wireless network control entity.

In 3GPP Release-12 and Release-13, one of network integration technologies has been studied and standardized, which are heterogeneous network integration based on WLAN-LTE interworking [32]. The aim of WLAN-LTE interworking is to solve the problem of UE's mobility and how a UE to simultaneously use LTE cell's resource and WLAN resource. As shown in the following figure of a possible multi-layer coverage network environment, UE can work simultaneously in eNB and WLAN AP. In this scenario, eNB is responsible for the mobility management and control of UE. WLAN APs only participate in data transmission.

In addition, the existing technology can also support the close integration between LTE and WLAN. As shown in Fig. 5 of a joint deployment scenario of WLAN AP(s) and LTE eNB(s), similarly, UE can work simultaneously in 4G eNB and WLAN AP while a new Xw interface is used between AP and eNB. eNB can transfer part or all of UE's data to WLAN AP to obtain the services provided by WiFi and realize resources aggregation and users' throughput improvement.

Future potential research areas include:

(1) Access technology intelligently selection: Constructing a user's non aware access technology selection mechanism. According to the real-time network state, the wireless environment, combining with the intelligent service sensing

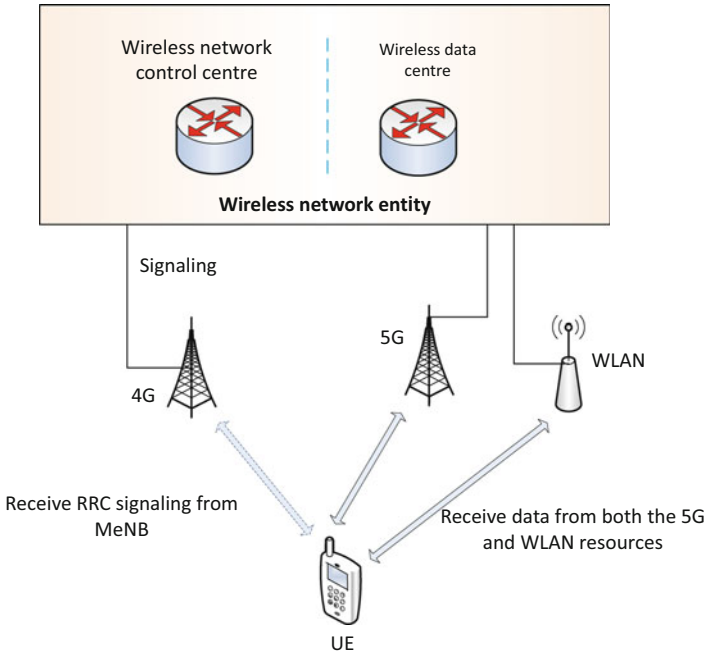


**Fig. 5** Multi-RAT coordination architecture

technology, the different services will be mapped to the most suitable access technology.

(2) Multi-RAT multi-connection technology: Terminals can simultaneously establish connections with different RAT network nodes and realize multiple flow parallel transmission to improve the throughput, enhance the user experience, achieve services dynamic splitting and convergence between the different access network technologies.

## *4.4 Mobility Management*

In UDN, mobility support and association decision is important because the small cell size causes frequent handovers, which introduce frequent short term service disruptions to AP and then significantly affect the user experience [33]. Meanwhile, the mobility management should be joint-optimized with the resource management and interference coordination in UDN. As a consequence, an effective mobile management mechanism can provide not only higher mobility performance but also an opportunity to meet higher throughput requirements for users [34].

The UDN is a multi-layer and multi-RAT network, network functions are built with these multiple levels of hierarchy. The centralized mobility anchors naturally

match with the hierarchical architecture, which will lead to more signaling load to the core network and suboptimal paths transmission. The advantage of centralized mobility management (CMM) mechanism resides in its simplicity, because the central anchor can follow user movements by simply rerouting the packets over tunnels created with the access router where the UE is currently connected. However, the single-handed mobility anchor thus easily becomes a single point of failure. Therefore, more flat and flexible mobility architectures and mobility management approaches are expected to be proposed [35].

### 4.4.1 Mobility Management Architecture in UDN

Today's trend points to the deployment of UDN in order to provide ubiquitous connectivity at high data rates. However, this is hard to couple with the current mobile networks' architecture, which is heavily centralized, posing difficult challenges when coping with high-frequent measurement, intolerable handover failure. Additionally, future 5G networks will exhibit disparate types of services, posing different connectivity requirement. The authors in [36] propose that distributed mobility management (DMM) is emerging as valid framework to design future mobile network architectures, taking into account the requirements for high-frequent handover and the rise of extremely dense wireless access networks.

In extremely dense wireless access networks, it is important to design a more flat and flexible mobility management architecture. The authors in [12] propose a systematic and unified methodology for designing mobility management architecture. The design philosophy is applied in the proposed function reference model and protocol reference model of Mobility-Driven Networks (MDN) to abstract the mobility support functions, entities, and principles.

Consider a Wireless Local Area Network consisting of Access Points that can provide only a limited cone of coverage, which leads to frequent changes of serving APs. The network as a whole, is, however, required to provide full coverage of a relatively large area, compared with the coverage of a single Network AP. At this point, the authors in [37] illustrate the structure of a Blanket Wireless Local Area Network (BWLAN), which serves a number of End Nodes (ENs). The BWLAN structure also contains centralized services and databases, as certain network operations, e.g., admission control, are best handled via a centralized gateway. In this architecture, each EN communicates with a set of NAPs through the wireless medium, and these NAPs will be referred to as the EN's "blanket." The notion of a blanket allows the network to locate each EN and transmit to the EN only from the NAPs that can reach it.

Existing cellular networks are insufficient in meeting future mobile users' demand, in part due to inflexible and expensive equipment as well as complex

and non-agile control plane. Software defined networking is emerging as a natural solution for next generation cellular networks as it enables further network programmability. An all-SDN network is described in [38] and followed by a unified approach to mobility, handover and connectivity management. It has proposed an alternative realization is to keep the handover control at the BS and RAN controllers but in close coordination with the routing control that is realized at the network controller, thus decreasing the signaling load to CN.

### 4.4.2 User-Centric Mobility Management

UDN is envisioned as a key deployment to address coverage and capacity issues in emerging wireless systems. In this case, the number of APs is likely comparable to the number of users.

Location areas are statically configured in existing network, but the boundary of a traditional location area in UDN becomes unclear. Thus, the location management mode will be redesigned, from "static AP planning" to "dynamic AP cooperation". In [39], an overview of challenges and requirements for mobility management in user-centric networks is given, and a new distributed and dynamic per-application mobility management solution is presented. Therefore, the authors in [6] propose a Dynamical APs Grouping method (DAPGing), which is considered a user-centered mobility management approach. With this method, every registered user has a unique APG with a unique APG-ID. APG context will be stored in a local serve center (LSC), and most of the DAPGing processes will be executed by LSC. Some high level processes such as authentication, handover are managed by network serve center (NSC). While a user is moving, its APG will be dynamically adjusted to support its movement, which is quite different from traditional mobility management and handover process. In traditional cellular network, users are handover from one cell to another. While in their proposed de-cellular ultra-dense network structure, the network will follow user's movement. The DAPGing makes mobility management functionality different. There are three following mobility scenarios and related approaches of user-centric mobility management:

(1) For Intra LSC or Inter LSC mobility, only APG members may perform refreshment for the terminal that has radio bearers, the APG-ID is not changed.
(2) For Inter NSC mobility, the APG should be possibly moved from one LSC to another LSC that connected to different NSC, assuming that the APG-ID is unique within one NSC.
(3) For Inter NSC mobility, or the mobility between UUDN and the traditional cellular network, the APG handover procedure is applied, e.g. from on one NSC to another NSC, or from one NSC to another traditional cellular network.

Furthermore, mobility management in UDN needs to be considered with interference coordination and resource management.

## 4.5 Interference Management

Interference management is very important in UDN to meet the high throughput requirement with short inter site distance. The resource multiplexing is adopted to increase the access opportunities, and then also brings more challenges for interference control. Traditional methods include interference randomization, interference cancellation, and inter-cell interference coordination (ICIC), interference alignment and etc. It needs further discussion to use these techniques in UDN. Meanwhile, new technologies such as millimeter-wave introduced into UDN may also present new challenges [14].

The research on interference management may include, but not limited to, the following:

(1) Channel model and capacity analysis. The wireless transmission environments in UDN become very complex with densified APs in multi-layer and multi-RATs condition. Therefore, the effective channel models need to be setup for various scenarios. Meanwhile, the channel capacity should be studied. The research in UDN just begins and needs further investigation. It is worth to note that this research direction relates to not only interference evaluation, but also other key technologies, such as coding method, antenna technique, etc.

(2) Interference model based on the effective evaluation methods. The ultra-dense environment results in more interference sources. For example, in crowd subway trains, lots of terminals and APs exist, therefore, signals may have more reflecting and scattering paths. Proper model should be set up to describe the interference level. Meanwhile, the existing parameters to measure and evaluate the impact of interference, such as interference temperature and interference threshold, may not reflect the overall interference measure and performance control of the networks. More suitable parameters should be discussed to give a better indication between interference managing results and throughput, associated with energy efficiency and other system level parameters [5].

(3) Interference management technologies with proper complexity. Interference management technologies have been studied with many research results [13] in traditional cellular networks. Considering the constraint in UDN, these techniques should be optimized with proper modification.

- Interference cancellation. It provides methods to regenerate the interfering signals through various coding methods and then subtract them from the desired signal [40]. It requires for other interfering users' information. Considering the complexity, it is usually used in the BSs. In UDN, APs' functions may be simplified. Therefore, interference cancellation should be modified.
- Inter-cell interference coordination. Fractional frequency reuse (FFR) and soft frequency reuse (SFR) have been proposed to control the neighbour cells interference from spectrum planning [41, 42]. Dynamic ICIC (D-ICIC) has the advantage of flexibility with many existing works [43]. The frequency

and time domain radio resources could be reused under certain limitation in multihop cellular networks [44]. These methods need coordination among BSs, therefore the signaling overhead generated in the process of exchanging information may impact the network performance.

- Coordinated multipoint transmission and reception (CoMP). One classic scheme to achieve CoMP is joint processing (JP)/joint transmission (JT), which is regarded as an advanced downlink solution and mainly focuses on achieving spectral efficiency in LTE-A. It is also possible to use the idea in UND, under the effective cooperation among APs.
- Interference alignment (IA). Each user is capable of reaching higher Degrees of Freedom (DoF) by using IA. DoF known as capacity prolong or multiplexing gain [45], is a good characterization of the approximate capacity because it becomes increasingly accurate in the high signal noise ratio (SNR) regime. IA regulates the interference falling into a particular signal subspace and leaves the residual subspace interference free. In UDN, appropriate pre-encoder could be designed at the transmitter with the purposes of preprocessing signals.

(4) Interference problems brought by new technologies. Advanced technologies have been introduced into UDN, to provide wider usable spectrum, higher throughput and better user experience. Millimeter-wave and non-orthogonal multiple access technology are considered as some of the most promising candidates and will be used in wireless access and backhauling. In millimeter-wave communication, high-gain beam forming is adopted to mitigate path loss and ensure low interference [14]. Besides, other traditional interference management methods should also be used to improve the overall performance [46].

The interference management needs to be jointly considered with resource management, mobility management and network deployment [5]. During the design of relative algorithms, interference control should be taken into consideration. Therefore, the high throughput requirements may be fulfilled in UDN.

## *4.6 Radio Resource Management*

As mentioned before, RRM in UDN faces new challenges from the complex communication environment and skyrocketing throughput requirements. More services will be developed with various QoS demands, requiring shorter setup time and delay, as well as reduced signaling overhead and energy consumption [47]. This means more flexible and efficient resource management schemes. Furthermore, the heterogeneous networks coexisting topology brings multi-RATs and multi-layer scenarios. How to jointly allocate resources in different networks to maximize the overall utility efficiency and optimize the system performance is an important and interesting problem for UDN.

The research on RRM of UDN may include, but not limited to, the following directions:

(1) Terminal association problem and corresponding resource allocation method. The neighbour APs usually have very short ISD with each other. This means the user may be under several APs' coverage, and then enjoy accessing services from one or more APs. Depending on the association condition, the resource assigned to serve the user may vary. The impacting factors may include the connected APs number of the terminal, the available resources, the required data rate and other QoS parameters, etc. Some researchers point out that the possible frequency band for UDN may extend from GHz to millimeter-wave, with wider bandwidths, higher flexibility in use of uplink/downlink resources based on time division duplexing (TDD) access, and physical layer waveform based on the modified METIS proposals [48]. This indicates wider choices for the scarce spectrum resources, as well as more complicated allocating algorithms.
(2) Cooperation and load balance among APs. When multiple APs serve for a terminal, these APs need close cooperation with each other to provide better services. The cooperation may come from PHY layer and MAC layer. Then the assigned resources for each AP should comply with the cooperation method. Meanwhile, considering the different APs' capability and access condition, the load balance and fairness among the APs should be considered. Besides, the necessary signaling overhead for the cooperation and load balance also need to be considered. In traditional cellular networks, the interaction among cells may be impacted in practical constraints due to backhaul delay, backhaul capacity and user mobility [49]. Whereas in UDN, these signaling messages may be dealt with in the local control unit and then save the processing cost.
(3) Power allocation and energy-saving scheme. In UDN, densified Aps indicate the possibility of higher overall energy consumption. Therefore, the green communication technologies are essential for the network. The power allocation should be considered to secure the overlapping of the wireless signals and QoS requirements, under effective interference control methods. Also, in the future mobile networks, the traffic load distribution will not only be impacted by the geographic area, but also relate with time, such as the night-day human behavior and daily commuting between offices and residential areas [44]. In order to save the energy consumption of idle APs, many researchers have proposed switch on/off methods for UDN. This will directly impact the resource allocation and access algorithms.
(4) Scenario of heterogeneous networks. Multiple networks coexisting with each other will become a typical deployment condition in the future communication networks. It means that UDN should also consider the coexistence and cooperation with other traditional cellular networks. Under this heterogeneous manner and multi-RATs interfaces, RRM has more challenges. The operations of different RATs have been independently defined by respective standards, but may lead to suboptimal usage of the wireless resources [50]. Multiple networks should jointly manage the radio resource and then improve the overall system

performance. The access control in heterogeneous networks needs further discussion when considering the joint optimization of RRM. The load balance among multiple networks, cooperation transmission and wireless backhaul will also have more flexible and complex schemes.

It has been pointed out that the RRM in UDN should be jointly considered, not only within the aforementioned research directions, but also with interference control, mobility management, multi-RATs, backhaul and other respects [5]. Although many researchers have paid attention to the relative fields, there is no well recognized resolution yet. Based on the cognition of transmission environment, the limited resources should be flexibly allocated and adjusted. From AP side or network side, two different approaches may be considered [5].

(1) When there are sufficient available resources and necessary function from the terminal, RRM may be fully distributed. The AP side will be responsible for the resource allocation process. It requires the APs should be capable of environment sensing on idle channels, interference levels, and neighbour APs' conditions. The advantages of this approach include the high efficiency of local resource managing and real-time fulfilling for the users requirements. Yet its disadvantages are also obvious. Negotiation process should be introduced to settle the conflict among APs. Then the processing complexity will increase. Also the requirements will be very high for environment cognitive ability of APs.
(2) The second approach needs the network side to participate and to offer center control in certain degree. A local control unit maintains the resource list in the neighbour areas, such as an available resource pool. It also controls the resource allocation process for the local APs. The APs may get the resource information by periodical broadcast or dedicated request. Once there is need for transmission, AP would then select the proper resource from the pool, the resource list kept by the local control unit may change accordingly. By this approach, RRM would achieve optimization by avoiding resource contention and decreasing interferences. The disadvantage is that it is hard to determine the neighbour area of the resource list in real cases.

The design of RRM algorithms would become hot research fields for UDN. Feasible and effective schemes will greatly improve the overall performance of the network and fulfill the users' requirements.

## 5 Conclusion

UDN is a technology to meet the mobile traffic requirements of 5G. It can provide better user experience even in extremely user density scenarios. Ultra density of APs is the key feature of UDN. Higher spectrum efficiency and energy efficiency, flexible networking, lower cost are the most important targets of UDN. User-centric

and localization are the new concept for UDN architecture design. Based on the new architecture, key technologies such as mobility management, interference management, SON and multi-RAT coordination need to be further enhanced.

**Acknowledgment** The authors would like to give special thanks to Dr. Ming Ai and Mr. Zhonglin Chen of CATT for their kind reviews and revisions.

## References

1. ITU-R M.2243, Assessment of the global mobile broadband deployments and forecasts for International Mobile Telecommunications, March 2011
2. Cisco, Visual Networking Index: Global Mobile Data Traffic Forecast Update 2013–2018, February 2014
3. IMT-2020 (5G) Promotion Group, 5G Vision and Requirement, May 2014
4. V. Chandrasekhar, J.G. Andrews et al., Femtocell networks: a survey. IEEE Commun. Mag. **46**(9), 59–67 (2008)
5. ITU-R M.2290, Future spectrum requirements estimate for terrestrial IMT, 2014
6. S. Chen, F. Qin et al., User-centric ultra-dense networks (UUDN) for 5G: challenges, methodologies and directions. IEEE Wirel. Commun. Mag. **23**(2), 78–85 (2016)
7. S. Chen, Y. Wang et al., LTE-Hi: a new solution to future wireless mobile broadband challenges and requirements. IEEE Wirel. Commun. Mag. **21**(3), 70–78 (2014)
8. ITU-R M.2320, Future technology trends of terrestrial IMT systems, November 2014
9. P. Popovski, V. Braun et al., Initial report on horizontal topics, first results and 5G system concept, METIS Deliverable D6.2, 2014
10. S. Chen, J. Zhao, The requirements, challenges and technologies for 5G of terrestrial mobile telecommunication. IEEE Commun. Mag. **52**(5), 36–43 (2014)
11. IMT-2020 (5G) Promotion Group, 5G Wireless Technology Architecture, May 2015
12. S. Chen, Y. Shi et al., Mobility-driven networks (MDN): from evolutions to visions of mobility management. IEEE Netw. Mag. **28**(4), 66–73 (2014)
13. H. Zhang, S. Chen et al., Interference management for heterogeneous network with spectral efficiency improvement. IEEE Wirel. Commun. Mag. **22**(2), 101–107 (2015)
14. R. Baldemair, T. Irnich et al., Ultra-dense networks in millimeter-wave frequencies. IEEE Commun. Mag. **53**(1), 202–208 (2015)
15. H. Wang, X. Zhou et al., Coverage and throughput analysis with a non-uniform small cell deployment. IEEE Trans. Wirel. Commun. **13**(4), 2047–2059 (2014)
16. P. Rost, C.J. Bernardos et al., Cloud technologies for flexible 5G radio access networks. IEEE Commun. Mag. **52**(5), 68–76 (2014)
17. I. Hwang, B. Song et al., A holistic view on hyper-dense heterogeneous and small cell networks. IEEE Commun. Mag. **51**(6), 20–27 (2013)
18. U. Doetsch, N. Bayer et al., Final Report on Architecture, METIS Deliverable D6.4, 2015
19. R. EI Hattachi, J. Erfanian et al., NGMN 5G Initiative White Paper, February 2015
20. V. Garcia, Y.Q. Zhou et al., Coordinated multipoint transmission in dense cellular networks with user-centric adaptive clustering. IEEE Trans. Wirel. Commun. **13**(8), 4297–4308 (2014)
21. H. Ali-Ahmad, C. Cicconetti et al., CROWD: an SDN approach for DenseNets, in *2013 Second European Workshop on Software Defined Networks (EWSDN)*, pp. 25–31, October 2013
22. H. Wang, S. Chen et al., SoftNet: a software defined decentralized mobile network architecture toward 5G. IEEE Netw. Mag. **29**(2), 16–22 (2015)
23. 3GPP TS 23.401 v12.3.0, GPRS Enhancements for E-UTRAN Access, December 2013
24. 3GPP TS 33.401, 3GPP System Architecture Evolution (SAE); Security architecture (Release 12), October 2014

25. 3GPP TR36.842 v12.0.0, Study on Small Cell enhancements for E-UTRA and E-UTRAN: Higher layer aspects, January 2014
26. 3GPP TR 36.902, Evolved Universal Terrestrial Radio Access Network (E-UTRAN): Self-configuring and self-optimizing network (SON) use cases and solutions, 2009
27. S. Tombaz, P. Monti et al., Is backhaul becoming a bottleneck for green wireless access networks? in *2014 IEEE International Conference on Communications (ICC2014)*, pp. 4029–4035, June 2014
28. 3GPP TS 36.216, Evolved Universal Terrestrial Radio Access (E-UTRA): Physical layer for relaying operation, 2014
29. X. Ge, H. Cheng et al., 5G wireless backhaul networks: challenges and research advances. IEEE Netw. Mag. **28**(6), 6–11 (2014)
30. X. Su, K. Hi Chang, A comparative study on wireless backhaul solutions for beyond 4G network, in *2013 International Conference on Information Networking (ICOIN2013)*, pp. 505–510, January 2013
31. J. Núñez Martínez, Self-organized backpressure routing for the wireless mesh backhaul of small cells, Universitat Politècnica De Catalunya, 2014
32. 3GPP TS 36.331, Evolved Universal Terrestrial Radio Access (E-UTRA): Radio Resource Control (RRC), 2014
33. J. Zhang, J. Feng et al., Mobility enhancement and performance evaluation for 5G ultra dense networks, in *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1793–1798, March 2015
34. P. Kela, J. Turkka et al., Borderless Mobility in 5G Outdoor Ultra-Dense Networks, the journal for rapid open access publishing, pp. 1462–1476, August 2015
35. F. Giust, L. Cominardi, C.J. Bernardos, Distributed mobility management for future 5G networks: overview and analysis of existing approaches. IEEE Commun. Mag. **53**, 141–148 (2015)
36. J.C. Zuniga, C.J. Bernardos et al., Distributed mobility management: a standards landscape. IEEE Commun. Mag. **51**(3), 80–87 (2013)
37. A. Reznik, C. Ye et al., Mobility Management for Dense Networks, vol. 34, pp. 1–4, May 2011
38. V. Yazici, U.C. Kozat et al., A new control plane for 5G network architecture with a case study on unified handoff, mobility, and routing management. IEEE Commun. Mag. **52**(11), 76–85 (2014)
39. F. Sivrikaya, S. Salsano, M. Bonola, M. Trenca, *Mobility Support in User-Centric Networks, Part IV of User-Centric Networking* (Springer International Publishing, 2014), pp. 269–288
40. B. Kaufman, E. Erkip et al., Femtocells in cellular radio networks with successive interference cancellation, in *2011 IEEE International Conference on Communications Workshops (ICC)*, pp. 1–5, June 2011
41. Z. Xu, L. Geoffrey Ye et al., Throughput and optimal threshold for FFR schemes in OFDMA cellular networks. IEEE Trans. Wirel. Commun. **11**(8), 2776–2785 (2012)
42. S.-P. Chung, Y.-W. Chen, Performance analysis of call admission control in SFR-based LTE systems. IEEE Commun. Lett. **16**(7), 1014–1017 (2012)
43. K. Zheng, Y. Wang et al., Graph-based interference coordination scheme in orthogonal frequency-division multiplexing access Femtocell networks. IET Commun. **5**(7), 2533–2541 (2011)
44. K. Zheng, B. Fan, J. Liu, Y. Lin, W. Wang, Interference coordination for OFDM-based multihop LTE-advanced networks. IEEE Wirel. Commun. **18**(7), 54–63 (2011)
45. H.-M. Anders, N. Aria, The multiplexing gain of wireless networks, in *2005 International Symposium on Information Theory (ISIT2005)*, pp. 2065–2069, September 2005
46. K. Zheng, Y. Wang, W. Wang, M. Dohler, J. Wang, Energy-efficient wireless in-home: the need for interference-controlled Femtocells. IEEE Wirel. Commun. **18**(6), 36–44 (2011)
47. R. Baldemair, E. Dahlman et al., Evolving wireless communications: addressing the challenges and expectations of the future. IEEE Veh. Technol. Mag. **8**(1), 24–30 (2013)
48. H. Tullberg, Z. Li et al., Towards the METIS 5G concept: first view on horizontal topics concepts, in *2014 European Conference on Networks and Communications (EuCNC2014)*, pp. 1–5, October 2014

49. R. Wang, H. Hu et al., Potentials and challenges of C-RAN supporting multi-RATs toward 5G mobile networks. IEEE Access **2**, 1187–1195 (2014)
50. Y.-N. Li, J. Li et al., Energy efficient small cell operation under ultra dense cloud radio access networks, in *2014 Globecom Workshops (GC Workshops)*, pp. 1120–1125, December 2014

# 5G RAN Architecture: C-RAN with NGFI

**Chih-Lin I, Jinri Huang, Yannan Yuan, and Shijia Ma**

**Abstract** This chapter discusses Cloud Radio Access Networks (C-RAN), which has been viewed as one of the key RAN architectures for 5G networks. First, the basic concept and the initially defined architecture of C-RAN are recalled, including the major benefits in terms of acceleration of network deployment, cost reduction and facilitation of 5G technologies. Then the major challenges of C-RAN realization are analyzed. One of the key challenges lies in fronthaul (FH) transportation which may limit C-RAN implementation in 5G. To address the issue, a new FH interface called Next Generation Fronthaul Interface (NGFI) is proposed. The C-RAN architecture itself evolves with the NGFI interface. The design principles for NGFI are presented including decoupling the FH bandwidth from the number of antennas, decoupling cell and user equipment processing and focusing on high-performance-gain collaborative technologies. NGFI claims the advantages of reduced bandwidth as well as improved transmission efficiency by exploiting the tidal wave effect of mobile network traffic. The transmission of NGFI is Ethernet-based to enjoy the benefits of flexibility and reliability. The major impact, challenges and potential solutions of Ethernet-based FH network are also analyzed. In addition, a prototype is developed and presented to verify NGFI.

## 1 Introduction

C-RAN which stands for Centralized, Collaborative, Cloud and Clean RAN was first proposed by China Mobile Research Institute in 2009 [1, 2]. The original motivation for the proposal of the new concept comes from practical needs in network deployment. Traditionally, network deployment requires a separate room per site with supporting facilities such as air conditioning to accommodate the base station (BS) or baseband unit (BBU). Every site consumes an equipment room. As the number of sites increases greatly, which is particularly the case for LTE, the number of equipment rooms required increases too. Traditional methods of deployment are

Chih-Lin I • J. Huang (✉) • Y. Yuan • S. Ma
Green Communication Technology Research Center, China Mobile Research Institute, Beijing, China
e-mail: huangjinri@chinamobile.com



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_16

becoming increasingly difficult since available real estate is becoming scarcer and rental costs are increasing. In this case, centralization of different BSs or BBUs into one location seems to be a natural solution for network deployment.

However, C-RAN does not stop at the centralization stage. As 5G recently became the research focus, it was realized that C-RAN serves as an ideal framework in support of 5G technologies. C-RAN is supposed to be able to accommodate and facilitate several 5G technologies such as Large Scale Antenna Systems (LSAS), full duplex, ultra-dense networks cooperative communication, mainly thanks to its inherent centralization nature as well as the flexibility and scalability of a cloud-based implementation [1]. In this sense C-RAN has become an essential element for future 5G systems.

In the following of this chapter, first the C-RAN concept will be recalled briefly and then an evolved C-RAN architecture which is based on a newly defined FH interface called NGFI will be proposed. The definition and design principles of NGFI will be described in detail with the advantages and challenges analyzed. Furthermore, different function split solutions will be analyzed, using LTE as an example which is critical to NGFI realization. The FH network design will also be analyzed from the perspectives of synchronization, jitter and delay. In addition a Proof-of-Concept to verify the NGFI will be described, followed by the conclusion.

## 2 C-RAN Basics

This section gives a basic introduction of the C-RAN concept, features and advantages.

### *2.1 The Concept of C-RAN*

With distributed BSs as the basic component, a C-RAN system centralizes different baseband processing resources together to form a pool in which the resources could be managed and dynamically allocated on demand on a pool level. Figure 1 shows the C-RAN architecture, which consists of three parts [2].

- Baseband resource pool: A baseband pool is a pool of resources from which resources can be instantiated to a “soft” BBU. In C-RAN the BBU is soft in the sense that the processing resources and capabilities are dynamically allocated and could be reconfigured based on real-time conditions (e.g. traffic status).
- Remote Radio Head (RRH) networks: RRHs are the same as in traditional systems to provide basic wireless signal coverage.
- Transport networks: A transport network provides a connection between a BBU instance in a pool and the remote RRHs. It could be of different forms depending on the scenario. Some examples include direct fiber connection via dark fiber, microwave transmission and fiber transport networks.

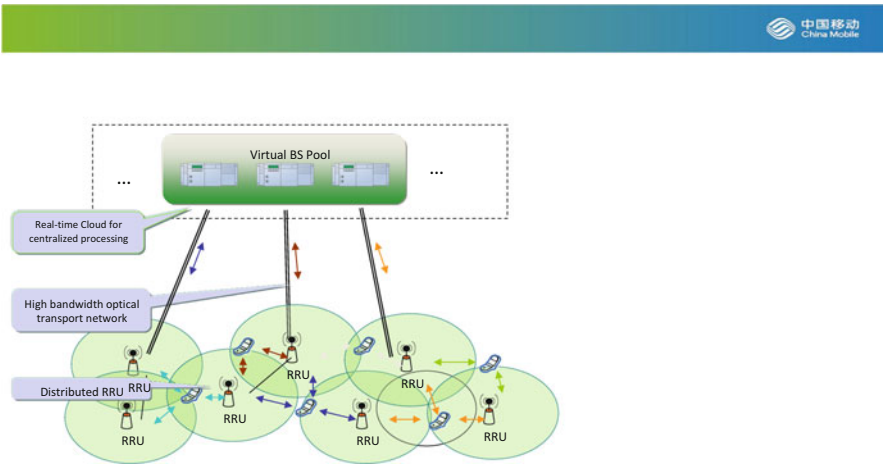


**Fig. 1** C-RAN architecture

## *2.2 Features and Advantages of C-RAN*

Aggregating the computing resources into one central location, centralization is the foundational feature and the first step toward a complete C-RAN. Centralization allows the aggregated computation nodes to share the same facilities, for instance air-conditioning, which greatly reduces the power consumption and, therefore, operating expense (OPEX). In a trial in China Mobile's commercial networks, it was demonstrated that power consumption could be reduced by up to 70 % with C-RAN deployment compared with traditional deployment methods [2]. In addition, centralization makes it easier to find a smaller number of central offices to accommodate the equipment, which in turn can speed up network construction.

Generalization of platform is another essential feature of C-RANs. The use of general purpose platforms (GPPs) not only reduces the procurement cost for operators but more importantly lays down the basis for the implementation of virtualization technology.

Based on the GPP, virtualization technology can be implemented to bring various benefits including improved resource utility efficiency, flexibility, scalability and energy efficiency, as they have already achieved and demonstrated in modern data centers in IT industry. In particular, with virtualization, the resource can be allocated on demand for different scenarios, traffic patterns, applications and radio access technologies (RATs). C-RAN in this sense can be called both "soft" and "green".

C-RAN is also "open" in the sense that C-RAN is designed to provide a set of standard application programming interfaces (APIs) to outside parties to encourage new service development on the edge and to provide interoperability. For example, dozens of radio parameters on user or cell level, e.g. the channel quality of a user or the number of users in a cell can be exposed so that the edge applications can adjust

parameters such as uplink/downlink (UL/DL) audio/video bit rates, picture and text compression rates accordingly. In this way the user experience can be improved. With its openness, C-RAN aims to push forward the prosperity of the ecosystem.

## *2.3 Support of 5G Technologies*

C-RAN may serve as an ideal framework for many potential technologies in various 5G aspects, such as duplex mode, network optimization, multi-RAT coexistence and spectrum sharing. For example, massive multiple-input multiple-output (MIMO) is identified as a key enabler for 5G air interfaces to dramatically increase both energy efficiency and spectral efficiency of the system via multiuser beamforming with a large number of antennas. Among many challenging issues, the inter-cell pilot contamination severely degrades the performance of massive MIMO, which mandates joint design of UL pilots and joint optimization of multiuser beamforming in adjacent cells. In addition, from the aspect of multiple access control, non-orthogonal multiple access (NOMA) supports more than two users on the same resource, thereby requiring more efficient and flexible means of system optimization. As to 5G duplex mode design, full duplex is regarded as an important direction of evolution. Full duplex base stations are capable of transmitting and receiving simultaneously on the same time and frequency resource, thus promising a maximum 100 % capacity improvement with well-scheduled DL and UL transmissions. The major issue for both NOMA and full duplex lies on the much more severe intra- and inter-cell interference. Especially in full duplex systems, on top of the conventional DL to DL and UL to UL interferences, there exist many complicated interferences between DL and UL. These interferences must be mitigated efficiently. In addition, one important network-level consideration of 5G is the extremely high base station density required to meet the ever-increasing mobile traffic requirements. Naturally, the high-density network design involves joint consideration of many issues, like the interference mitigation in a heterogeneous network, control and data decoupling, DL and UL decoupling, carrier aggregation and Coordinated Multi-point (CoMP) etc. As to the multi-RAT coordination, 5G networks need to coexist harmoniously all the existing 3G, 4G and WiFi networks. The information sharing between these different networks and corresponding joint scheduling are essential to its success. Finally, due to the scarcity of wireless spectrum, spectrum sharing among all wireless networks (and operators) will fundamentally enhance resource utilization efficiency.

The performance improvements promised by the above technologies, however, require systematic information sharing and joint optimization, which is difficult to achieve with traditional distributed architecture. Fortunately, thanks to the inherent capability of central processing, such information sharing becomes feasible in the C-RAN context and therefore C-RAN is deemed to be able to facilitate the realization of those 5G key technologies.

## 2.4 Fronthaul Challenges and Existing Major Solutions

There are two major challenges for C-RAN realization. One is the FH challenge to enable efficient and scalable centralization and the other is resource cloudification. The ultimate goal of C-RAN is to realize the features of cloudification, i.e. processing resources can be dynamically allocated to form soft BBU entities. The solution to this is to implement virtualization technology which has been pervasive in IT industry for cloud applications. Thanks to the extremely strict real-time requirements for signal processing in wireless networks, the implementation of virtualization may require special design and optimization. Some challenges for this include optimization of operating systems and the hypervisor, optimization of virtualization management functions, I/O virtualization to improve the virtual machines (VM)'s I/O performance and its compatibility with live migration. Readers could find more information in [2, 3]. In this chapter we will mainly focus on the challenges on the FH side since centralization is the basis to realize other features of C-RAN.

FH is defined as the link between a BBU and a RRH. Typical examples of FH interfaces include Common Public Radio Interface (CPRI) and the Open Base Station Architecture Initiative (OBSAI) [4].

Centralization is the critical first step required in order to realize all the other features of C-RAN. Centralization aggregates different BBUs into one central office with shared facilities. The key challenge for centralization is that it requires a large number of fiber resources if using direct fiber connections. The issue can be illustrated by the following example. In a TD-LTE system with 20 MHz bandwidth and RRHs of 8 antennas, which is the most common scenario in CMCC's networks, the CPRI data rate between the BBU and the RRH for one TD-LTE carrier transmission is as high as 9.8 Gbps. When considering both UL and DL, 2 fiber connections would be required with 10 Gb/s optical modules. Since usually one site consists of three sectors with each supporting at least one carrier, the number of fiber connections for one site can be as high as 6, which is difficult to achieve for most operators due to limited fiber resources.

In order to overcome the fiber disadvantage in a centralized implementation, various solutions have been developed. Some are mature enough while others are still in the early stage.

One of the solutions is CPRI compression. There are various kinds of compression techniques such as non-linear quantization and IQ data compression [5]. So far, major vendors have been able to support 2:1 compression with lossless system performance. In [3], the authors have verified the functions in their commercial LTE C-RAN networks. Another solution is Single Fiber Bi-Direction (SFBD) which allows simultaneous transmission of UL and DL on the same fiber. SFBD could further reduce the usage of fiber by another 50 %. As a result, when combining SFBD with compression, fiber resources can be reduced 4-fold with lossless performance.

The most common and effective FH solution today is to use wavelength-division multiplexing (WDM) technology. With WDM dozens of carriers could be carried on a single fiber, which greatly conserve fiber resources. Although the introduction of new transport nodes in the network may also cause additional delay and noise jitter, it turns out that the expense is affordable. In [3] the authors have tested the mainstream WDM FH equipment in the commercial LTE C-RAN network and found that the additional delay and jitter contributed by the WDM equipment are small enough to be negligible. Even with the optical transport network (OTN) technology which is supposed to have much higher delay than WDM, the delay is still around 20–30 μs, small enough to have little impact on the system performance.

However there is one common disadvantage for all the solutions today, no matter whether they are WDM, CPRI compression, or SFBD, which is that the object that these transport solutions carry is CPRI data. Due to the inherent shortcomings of CPRI (see next section), the transport efficiency is low.

In order to better support C-RAN large-scale deployment and key 5G technologies, a new FH interface is proposed in the following sections.

## 3 NGFI: Next Generation Fronthaul Interface

In this section, the design of the fronthaul interface in order to better support 5G technologies, especially C-RAN will be introduced.

### *3.1 Motivation*

As mentioned in previous sections, on the road of C-RAN realization, the FH issue has been one of the biggest challenges. Due to the high FH data rate centralization requires consumption of a high number of fiber cores which are scarce and not easy to afford. Although other transport technologies such as WDM and OTN could save fiber consumption, the cost on the introduction of additional transport equipment makes the economic viability a concern to operators. Some operators are still not very convinced on the merits of C-RAN deployment. Therefore in 5G to enable large-scale C-RAN deployment it is necessary to reduce FH bandwidth.

Current FH interfaces could also raise new issues for C-RAN virtualization. C-RAN is supposed to run on a general-purpose processor (GPP) platform consisting of standard IT servers, storage and switches. GPPs do not provide a FH interface for telecom applications. To support FH, either a new interface should be created on the GPP platform or an adapter card is needed, both complicating the system and introducing additional costs. It would be desirable to base the 5G FH interface on existing GPP interfaces which could maximize efficiency and save costs.

In addition, scalability issues exist for today's FH technology to support widely discussed 5G technologies, including LSAS, CoMP processing etc. For example, it

is possible that a 5G LSAS RRH could be equipped with 64 or even 128 antennas. With LTE, the FH bandwidth will rise to the order of 100 Gb/s, which makes it unaffordable. It is clear that the impact of the number of antennas on FH should be minimized to the greatest extent possible.

In order to address the issues above, one possible idea is to change the FH interface fundamentally.

In the following chapter, CPRI will be used to represent traditional FH solutions since it is the most widely adopted in the current industry.

## *3.2 CPRI Re-Visiting*

The CPRI interface helps to separate the BBU and the RRH to enable the deployment of distributed base stations. Although in traditional networks the CPRI is mainly deployed over short distances, usually in the order of several meters or several hundred meters, it can support up to 40 km between the BBU and the RRH. CPRI has been working well for traditional mobile networks including 2G, 3G and 4G. There are three traits for the CPRI interface.

- The CPRI line rate is constant regardless of traffic;
- The mapping between the BBU and the RRH is fixed one-to-one correspondence and not flexible;
- The sampling I/Q data rate is dependent on the number of antennas.

With networks evolving to 5G, CPRI is becoming more and more unsuitable to accommodate the evolution. First, it is well known that mobile traffic varies in the temporal dimension, which is called tidal wave effect. For example, the data traffic in an office area is high in the daytime and plummets at midnight. For dense urban areas, the tidal wave effect is noticeable. However, the CPRI data stream is synchronous digital hierarchy-alike (SDH-alike), which means that it is constant regardless of the changes in traffic [4]. Even when there is no user traffic in the network, there are still CPRI streams running between the BBU and the RRH. This is a waste of bandwidth and leads to low utilization efficiency.

Second, with CPRI an RRH is in one-to-one correspondence with a BBU. The relationship is configured offline. It may cause a concern in the context of C-RAN. In C-RAN, the baseband units are centralized and virtualized in a pool. Reliability becomes extremely important as each pool takes care of thousands of users. Therefore for the sake of protection, it would be desirable if in C-RAN, one RRH could be automatically switched to another BBU pool. Current CPRI however, does not support such flexible and automatic re-mapping.

Finally, the CPRI bandwidth is dependent on the number of antennas. As the number of antennas increases, the CPRI data rate increases in proportion. This could become a major hindrance for CPRI's applicability in 5G as far as multiple antenna technologies are concerned.

Based on the analysis above, it is believed that the CPRI interface that is universally applied throughout the industry is primarily applicable to peer-to-peer connections. Due to the disadvantages such as low transmission efficiency, poor flexibility and poor scalability, and particularly due to the high cost of centralized deployment, it is difficult for current FH interfaces to meet the evolving needs for 5G-oriented FH networking.

To increase the efficient use of network resources and better support the evolution of wireless networks toward 5G systems, as well as to better support C-RAN-centralized deployment, a new BBU-RRH interface is proposed in the following sections.

### *3.3 Definition and Design Principles of NGFI*

The motivation for designing a new FH interface is to facilitate C-RAN deployment, to make it compatible with GPP platforms and to be scalable enough to support the evolution to 5G.

The new FH is defined as the Next Generation Fronthaul Interface (NGFI) between the BBU and the RRH with the following four features [6, 7]:

- Its data rate should be traffic-dependent and therefore support statistical multiplexing.
- The mapping between BBU and RRH should be of one-to-many correspondence and flexible;
- It should be independent of the number of antennas.
- It should be packet-based, i.e. the FH data could be packetized and transported via packet-switched networks.

With NGFI, several advantages are expected, including:

- The FH transmission efficiency could be improved. This is because NGFI is traffic-aware. Therefore it could fully leverage the tidal effect of mobile networks so that the FH data rate is variable with traffic change.
- With reduced average data rates, it could further reduce the pressure on FH transport networks for the sake of C-RAN deployment.
- NGFI could support 5G antenna technologies well thanks to the antenna-independent feature.
- The packet-based feature makes it possible to use Ethernet to transport FH data. The benefits would be several-fold. First, the Ethernet interface is the most common interface on standard IT servers and the use of Ethernet makes C-RAN virtualization easier and cheaper. The operation and maintenance are therefore straightforward. The Ethernet can fully make use of the dynamic nature of NGFI to realize statistical multiplexing. The flexible routing capability could also be used to realize multiple paths between the BBU pool and the RRH and to enable FH resources sharing.

The fundamental way to realize NGFI is to re-design the function partitioning between the BBU and the RRH. Furthermore, the NGFI interface will have a great impact on the design of the underlined transport networks for C-RAN deployment.

In the next section, the design and analysis of BBU-RRH function split as well as the packet-switched FH networks will be introduced.

## *3.4 NGFI-Based C-RAN Architecture*

NGFI requires the baseband (BB) function to split between the BBU and the RRH, i.e. moving partial BB functions from the local BBU to the remote RRH. C-RAN architecture evolved with NGFI. The NGFI-based C-RAN architecture is shown in Fig. 2 which consists of the following components:

- Radio Aggregation Unit (RAU): The moved partial BB functions formed a new entity which is called radio aggregation unit. RAU is a logical concept and its realization depends on implementation solutions. For example, RAU could be integrated to the RRH to form a new-type RRH. Alternatively, it could also be an independent hardware entity.
- Remote Radio Systems (RRS): A RRS consists of an RAU and multiple RRHs. It is expected that collaboration could happen among different RRHs via the RAU within the same area coverage of a RRS. There could be multiple RRS in a C-RAN network.

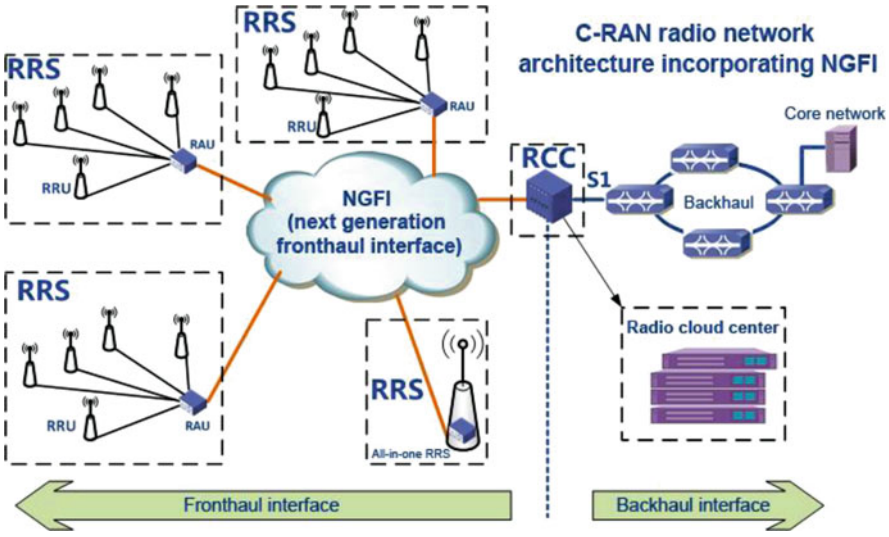


**Fig. 2** NGFI-based C-RAN architecture

- Radio Cloud Center (RCC): The remaining BB functions, together with high-layer functionalities constitute a RCC. RCC is the place where all the processing resources are pooled into a cloud with virtualization technology.

## 4 Analysis of RCC-RRS Function Split

Traditionally, the baseband-related functions are processed by the BBU while the RRH processes radio frequency-related functions. It is this simple partitioning that leads to the shortcomings of CPRI, as mentioned above. Therefore, the NGFI design should start with a paradigm shift by rethinking and redesigning the function split between the RCC and the RRS. Moreover, the function split between RCC and RRS may be different for different scenarios. For example, generally the number of the antennas of each radio head is 2 in indoor distributed systems, in which fewer functions may be moved from RCC to RRS. On the contrary, for outdoor scenarios, the number of the antennas of a massive MIMO radio head could be 128 or even larger. In this case more functions need be moved from RCC to RRS to reduce the FH bandwidth.

### *4.1 RCC-RRS Function Split Principles*

Function split is the key way to achieve NGFI and different split options would lead to different features for fronthaul interface. In this section the principles for desirable function split schemes will be discussed.

#### 4.1.1 Decoupling the FH Bandwidth from the Number of Antennas

The air interface bandwidth per carrier on 2G, TD-SCDMA and TD-LTE are 0.2 MHz, 1.6 MHz and 20 MHz, respectively. Correspondingly, FH transport bandwidth per carrier is 30 Mbps, 400 Mbps and 10 Gb/s, respectively. At the same time, FH is facing a bandwidth explosion, considering the rapid traffic growth in 5G (potentially 1000× by 2020) and C-RAN-centralized deployment. Compared with the air interface bandwidth, the existing FH interface transportation efficiency is low. One of the most important reasons is that FH bandwidth is proportional to the number of antennas. In order to increase transport efficiency, a RCC/RRS function split principle should enable NGFI to decouple FH bandwidth from the number of antennas.

Taking TD-LTE as an example, an 8-antenna TD-LTE carrier FH bandwidth is 10 Gb/s based on the current BBU/RRH function split. For a C-RAN baseband pool with 100 TD-LTE carriers, the overall FH bandwidth would be as high as 1 Tb/s. If it is a 128-antenna LSAS system, a TD-LTE carrier FH bandwidth will

increase to 160 Gb/s. Then the overall FH bandwidth of the C-RAN baseband pool will increase to 16 Tb/s. Thus, the existing FH blocks the C-RAN centralized deployment and is facing a big challenge with the increase of antenna numbers. In order to reduce the bandwidth, one potential idea is to redesign the RCC/RRS function split. In particular, in order to phase out the effect of antenna numbers, it is proposed that antenna related functions should be moved from the RCC to the RRS. For LTE, the antenna related functions include DL antenna mapping, FFT/iFFT, channel estimation and equalization. It is shown that an LTE carrier FH bandwidth may decrease in the order of 100 Mb/s no matter how many antennas are used. Therefore, FH bandwidth will decrease significantly, if the RCC/RRS function split can decouple non-antenna related processing and antenna related processing.

### 4.1.2 Decoupling Cell/UE Processing

Dynamic variation is a major feature of wireless traffic. The tidal wave effect is obvious in many wireless deployment scenarios such as residential, office and commercial districts. Moreover, the traffic load of most areas is usually in the valley between late night and early morning. However, the current constant-rate FH does not match the mobile traffic features, which results in a waste of resources.

In order to quantify wireless traffic features, a TD-LTE traffic investigation on a commercial LTE network was done via network monitoring systems. There were six base station sites in the investigation, each having at least three carriers. Two of them are indoor distributed systems while the others are outdoor macro base station sites. The investigation period was 7 days during which the traffic load was sampled and collected by network monitoring systems every 15 min.

In Figs. 3 and 4, the DL and UL traffic load distribution of the above six base station sites are shown. Site 1, 2, 3 and 4 are outdoor sites. Site 5 and 6 are indoor sites. Based on the statistical data, several traffic load features were observed as follows:

- The tidal wave effect is obvious in the test districts, where the traffic load is almost zero for 12 h.
- Even when the site is at “peak” status the resource block (RB) usage is not high. The duration of a site at peak status is short, usually not exceeding 30 min.
- The probability that different sites are simultaneously in peak status is almost zero.
- When the site is at “valley” status the RB usage is low and the duration is long, usually exceeding 2 h.
- Frequently, different sites are in valley status at the same time. Moreover, the overlapping time is long among different sites.
- Most of time the traffic load stays at the average level, which is low.

Based on the above observations, it is clear that constant-rate FH transport does not match the mobile traffic features, which results in a waste of re-sources. To address this issue, it is first observed that the existing base band processing can be divided into cell processing and user equipment (UE) processing. Cell processing

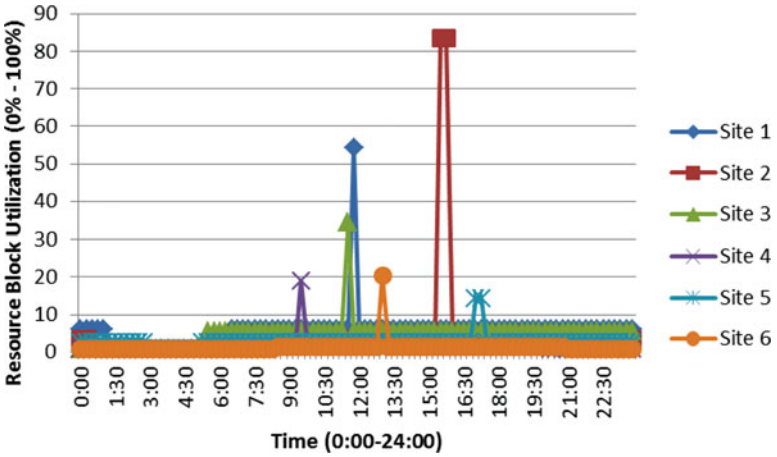


**Fig. 3** DL traffic load distribution

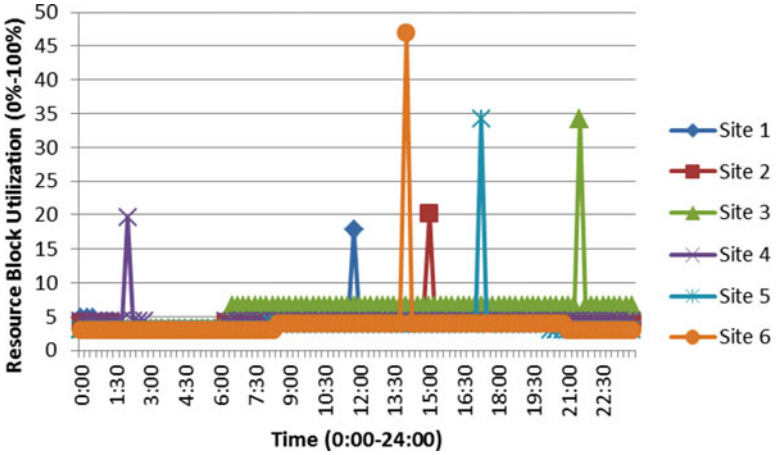


**Fig. 4** UL traffic load distribution

is irrelevant to traffic load and is fixed no matter how many UEs are active. Some examples of such processing units in LTE include inverse fast Fourier transform/fast Fourier transform (iFFT/FFT), cyclic prefix (CP) addition/removal, cell-specific reference signal/primary synchronization signal/secondary synchronization signal (CRS/PSS/SSS) generation and physical broadcast channel (PBCH) processing. It is therefore proposed to move these cell processing functions from the RCC to the RRS, i.e., decoupling the cell and UE processing.

If cell processing is moved from the RCC to the RRS, the FH bandwidth will be lower and load dependent. The load-dependent feature gives an opportunity to exploit the statistical multiplexing gain when it comes to FH transport network

design for C-RAN deployment. Although the peak throughput requirements for radio resources are very high and rising with the evolution of RAN, the statistical result of our commercial network shows that the duration of a site at peak status is short, usually not exceeding 30 min. Moreover, the peak throughput occurrence time is usually different among different sites. Therefore, every site can also reach peak throughput under FH network statistic multiplexing model. In order to not decrease the air interface performance, FH network statistic multiplexing must be designed suitably according to the business model of different scenarios. Thanks to statistical multiplexing, the bandwidth needed for transport of a number of FH links in C-RAN could be reduced greatly and therefore the cost is diminished.

Cell/UE processing decoupling can further help reduce power consumption and enhance the network reliability. This is because cell basic cover-age signal processing is a kind of cell processing. Therefore, cell basic coverage will be provided by the RRS if cell processing functions are moved from the RCC to the RRS. On the one hand, RCC software can be switched to a dormant state to save power when there is no active UE. On the other hand, RRS is able to provide continuous air interface coverage, even when RCC breaks down. This way, it provides sufficient time for RCC fault processing.

### 4.1.3 Focusing on High-Performance-Gain Collaborative Technologies

CoMP has been viewed as one of the important 5G technology candidates to improve system performance, which can be divided into two classes: MAC layer coordination and physical layer coordination. For example, collaborative scheduling (CS) is one of the MAC layer coordinated mechanisms. Joint reception (JR) and joint transmission (JT) are physical layer coordinated technologies. The design of NGFI should take into account support for CoMP. The above two principles lead to a low bandwidth and traffic dependent FH. In the meantime, some physical-layer-coordinated technologies are difficult to implement in RCC since some collaborative information has be processed and terminated by RRS. Therefore, it is proposed that data exchange among RRS can be supported on FH network. However, data exchange among RRS is not necessary because it is found that the performance gain of JR/JT decreases significantly as the number of antennas increases [8]. Moreover, it is also found that MAC-level collaborative technologies can bring comparable performance gains with lower complexity, easier implementation and fewer constraints.

In order to verify this, a CoMP field trial was conducted in 2014 in which two CoMP schemes including JT and CS were examined and compared. The testing zone was a central business district including around 7000 active UEs, which is a typical CoMP test scenario. It was covered by 35 base stations with different antenna heights in which inter-cell interference is serious because of the high ratio of overlap area.

Test results show that cell edge UE throughput increased by $127.45\,\% \sim 173.65\,\%$ when the serving cell reference signal received power varied from $-88$ dBm to

**Table 1** Utilization factor of JT and CS

| Network load | JT/CoMP | CS/CoMP | CoMP/Total |
|---|---|---|---|
| 20 % | 20.04 % | 79.96 % | 42.02 % |
| 50 % | 18.09 % | 81.91 % | 34.43 % |
| 70 % | 20.89 % | 79.11 % | 54.33 % |
| 100 % | 24.89 % | 75.11 % | 47.31 % |

−106 dBm. Table 1 shows the utilization factor of JT, CS and the CoMP application. In Table 1, the "CoMP/Total" means the utilization factor of CoMP which is defined as the ratio of the number of CoMP transmission time intervals (TTIs) to the total number of test TTIs. Similarly, the "JT/CoMP" and "CS/CoMP" are defined as the ratio of the number of JT and CS TTIs to that of CoMP TTIs, respectively.

From Table 1, it is found that the network load growth results in a small reduction of CS usage. For example, when the network load increases to 100 %, the usage ratio of CS is still around 75 %, similar to other cases. The usage of JT is only around 25 %, much lower than CS. It is therefore fair to say that most of the performance gain is contributed by CS. Compared with JT, CS does not need complex matrix computing. It is easier to implement CS with contemporary base station equipment. On the contrary, JT performance is influenced by antenna calibration accuracy, channel estimation accuracy and channel variation speed, all requiring high FH bandwidth.

The test results demonstrated that MAC-level collaborative technologies could mitigate most network interference. Therefore, RCC should focus on high performance gain collaborative technologies rather than all the collaborative technologies. This principle provides guidance on how to make a tradeoff between wireless and FH performance.

## *4.2 Analysis and Comparison of LTE Function Split Solutions*

As shown in Fig. 5, taking LTE as an example, potential function split solutions for the RCC-RRS interface are discussed in this section. Baseband processing functions can be divided into UE processing (related to traffic level) and cell processing (unrelated to traffic level). The green blocks in the figure are traffic-dependent UE processing modules, and the yellow blocks are traffic-independent cell processing modules. The blue block, channel estimation and equalization, is also UE processing. The complexity of channel estimation and equalization is correlated with the number of receiving antennas.

In order to estimate the FH bandwidth demands of different RCC-RRS function split solutions, the following assumptions are made:

1) Single 20 MHz LTE carrier
2) 2 antenna ports
3) 8 antennas

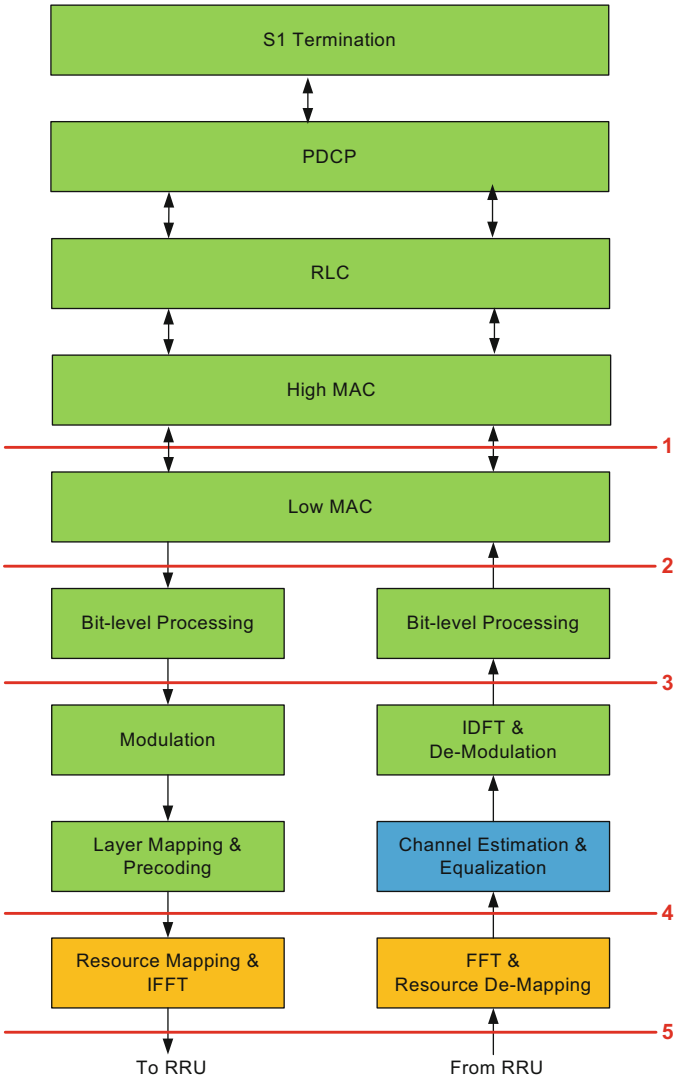


**Fig. 5** Potential RCC-RRS function split solutions in LTE

4) The highest modulation type of DL is 64 quadrature amplitude modulation (QAM). The highest modulation type of UL is 16 QAM
5) Maximum number of users: 100

As shown in Table 2, "DL/UL" means DL/UL bandwidth requirement per carrier which does not include synchronization data or Ethernet headers. In the table, the bandwidth of solution 1, 2, 3, and 4 is the maximum bandwidth, and the bandwidth of existing solution 5 is constant. Delay is shown not in the exact

**Table 2** Comparison of different split schemes

| | Solution 1 | Solution 2 | Solution 3 | Solution 4 | Solution 5 |
|---|---|---|---|---|---|
| DL | 174 Mb/s | 179.2 Mb/s | 125.2 Mb/s | 498 Mb/s | 9830.4 M b/s |
| UL | 99 Mb/s | 78.6 Mb/s | 464.6 Mb/s | 2689.2 Mb/s | 9830.4 M b/s |
| Delay | 1 ms | 0.1 ms | 0.1 ms | 0.1 ms | 0.1 ms |

numerical value but in the order of magnitude. It is found that the FH bandwidth requirement of solution 1 & 2 is approximately equal with that of backhaul. All the function split solution 1, 2 and 3 meet the RCC-RRS function split principles, as described in the previous subsections. Further research on each potential solution is still needed to determine which function split solution is the most suitable for NGFI.

Based on the analysis of the above potential solutions, NGFI is a compromise solution that takes into account FH bandwidth, cost, delay, mobile traffic characteristics, support for potential 5G technologies, upgrading and maintenance challenges, support for cooperative algorithms and other factors. With the evolution of C-RAN, NGFI also should be evolving to meet the FH requirements of next-generation wireless technologies (e.g. 5G, 6G).

## *4.3 Impact of Function Split on the RRH Side*

Compared with the traditional RRH, RCC-RRS function split increases the complexity on the RRH side to some extent, mainly in terms of the following three aspects:

- Partial baseband function processing is added on the RRH side.
- Clock synchronization modules are added to meet the time and frequency synchronization requirements.
- The existing BBU-RRH peer-to-peer connection is expanded to multipoint-to-multipoint connection. Therefore, RRS should be viewed as an independent network element, and RRS management functions should be also added accordingly.

Moreover, there are two ways to replace the partial baseband functions to be moved from the BBU side. One is to integrate them into the RRH, leading to a new type of RRH. The other is to add dedicated equipment to deal with the partial baseband function processing, which could be co-located with the original RRH. For the integrated RRH, adding partial baseband function processing will increase the power dissipation, which brings some impact on the volume and weight. The volume, weight, power dissipation and transmission power level of the RRH are closely interrelated. According to the estimate of the current RRH cooling capacity, at the upper limit of temperature, the heat volume is approximately 15 W/L.

Compared with the existing RRH, the volume of the integrated RRH will increase by 1 L for each additional 15 W of power consumption, at the same time with a corresponding 1 kg weight increase. Currently, based on the commercial network power dissipation test, it was shown that the power dissipation of the RRH is in the order of 100 W and that of BBU is in the order of 10 W. Therefore, the power dissipation of partial baseband processing should be in the order of 1 W. Preliminary predictions are that adding partial baseband processing on the RRH side will only bring a slight increase in power dissipation. However, the exact impact of the power dissipation and cost needs to be evaluated quantitatively in more details in the future.

## 5 NGFI FH Network Design

The principle of NGFI points out that a desirable FH interface should be packet-based, which makes it easily transmitted by packet-switched networks, especially Ethernet. This could make full use of the advantages of Ethernet to achieve multi-point to multi-point connection, statistical multiplexing, and flexible routing. In the meantime, the adoption of Ethernet also brings new challenges. In this section, such challenges will be analyzed and potential solutions will be given based on wireless data packetization.

### *5.1 Wireless Data Packetization*

Wireless data packetization refers to using Ethernet packets to transmit wireless data between the RCC and the RRS. Compared with time-division-multiplexing (TDM) packaging in traditional CPRI interface, data packetization could provide greater flexibility and scalability, meet the need for high-layer wireless data transmission, be capable of identifying user traffic and achieve efficient transmission of wireless user loads and related control loads in packet-switched networks. To meet the high requirements of wireless data transmission and retain the advantages of packetization, achieving packetized transmission based on wireless data characteristics is one of the problems that must be considered in realizing NGFI interfaces in the transport field.

Given that wireless payloads are very sensitive to synchronization, jitter and delay, it may be necessary to introduce specific packet headers to deliver in-band control information such as timestamps or sequential numbers to facilitate coordination between the receiving and transmitting ends. The process, i.e. introduction of packet headers into wireless loads is called encapsulation. Encapsulation could be “structure aware” or “structure agnostic”. The design of encapsulation should take into account several factors such as the application scenarios, overheads and other requirements. In addition, synchronization technologies necessary for wireless data transportation are another research focus. The encapsulation process

could be irrelevant to either the baseband function split schemes or the specific radio technologies in order to have better forward compatibility and technological neutrality.

Packetized wireless payloads should be transparent with respect to the transportation networks and could be carried by different transmission technologies, including Ethernet, Multiprotocol Label Switching-Transport Profile (MPLS-TP), IP routing and L2/L3 MPLS switching. Different transportation technologies carry the wireless payloads in different ways based on their own characteristics and capabilities. At the same time, to distinguish such loads from other existing protocol or packet types, it may be necessary to introduce new encapsulation or protocol identifiers for the wireless loads. Different transportation technologies vary in terms of the carrying and forwarding efficiency, network/traffic scalability and adaptability, operation and maintenance management capabilities and ecosystem. Appropriate choices should be made based on the overall strategy of the transmission networks, together with the technical characteristics of each transmission technology.

In Fig. 6, an example of FH topology of a C-RAN system is shown. In the example function split between the RCC and the RRS is not considered. Therefore the RCC is the same as BBU pool while the RRS is the same as the RRH. As shown in the figure, all the RRHs are connected to a BBU pool through a ring Ethernet

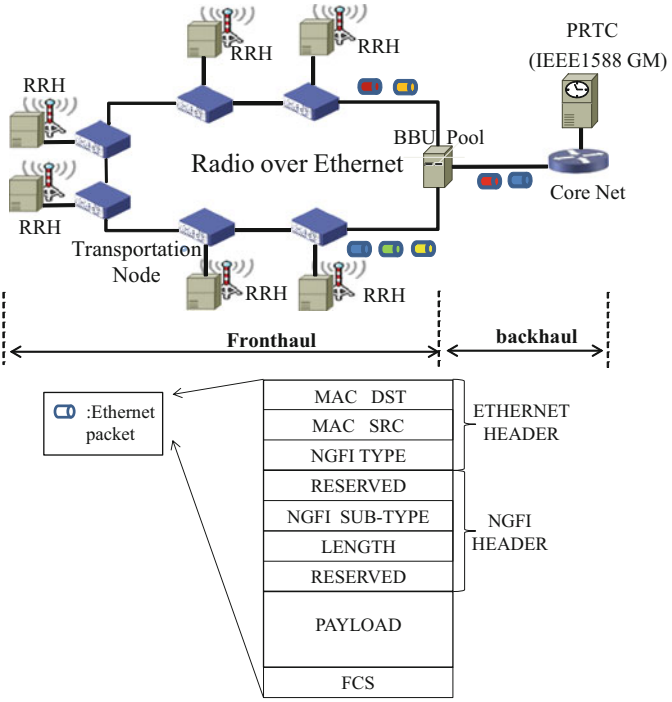


**Fig. 6** A FH topology example and illustrative Ethernet packet format to support NGFI

network. There are multiple routes between BBU and RRH to help to enhance network reliability. When one of routes fails, FH packets can be transported through another route.

FH DL data is encapsulated with Ethernet headers in the BBU and de-encapsulated from Ethernet packets in the RRH and vice versa for UL. In Fig. 6, the structure of an NGFI-supporting Ethernet packet is proposed. It includes traditional Ethernet header, NGFI header and payload. Source Media Access Control (MAC) address, destination MAC address and packet type are fill in the traditional Ethernet header. The packet type here is the NGFI type which is new to distinguish NGFI packets from other packets. The NGFI header consists of NGFI packet sub-type, packet length and reserved field for protocol extension. There could be at least two NGFI packet sub-types. One is for wireless data while the other is for control & management data. For control and management type, it may include link delay test, link status monitoring, RRH configuration and RRH status reporting.

## *5.2 Main Challenges of NGFI Transmission*

### 5.2.1 NGFI Synchronization Issues and Potential Solutions

TDM systems require strict synchronization which includes two aspects—frequency and time (or phase). For Time Division-Synchronous Code Division Multiple Access (TD-SDMA) and TD-LTE, the accuracy of frequency synchronization should be in the range of $\pm 0.05$ ppm while the accuracy of time synchronization should be in the range of $\pm 1.5$ μs [9].

In CPRI three types of data including wireless protocol data, synchronization data and control & management data are packaged together and transmitted in TDM mode. Upon receiving the CPRI frames, the Clock and Data Recovery (CDR) circuit of the RRH can extract frequency information to achieve frequency synchronization. Meanwhile, CPRI transport time is nearly constant and can be measured by the BBU. Based on the measurement, timing between the BBU and RRH can be configured in advance. With timing information extracted from CPRI frames, time synchronization at the RRH could be achieved.

For Ethernet-based NGFI, contrary to CPRI, the transportation time of FH data is no longer constant due to the packet-switched nature of Ethernet. As a result, frequency and time synchronization between the RCC and the RRS becomes potentially difficult.

The potential NGFI synchronization scheme includes the introduction of Global Position System (GPS) or the BeiDou system on the RAU side or synchronization via FH networks. If the synchronization scheme adopts GPS or the BeiDou system on the RAU side as the synchronization source, the impact of the NGFI will be relatively small. If it adopts clock synchronization from the FH network, it is necessary to carefully consider the requirements of the synchronization clock with respect to the FH network. Synchronization accuracy is related to the scale of the

FH network and performance of the transmission equipment. Synchronous Ethernet (Sync-E) and 1588v2 are potential mechanisms to achieve synchronization by FH networks.

Sync-E could be a promising solution in NGFI to achieve frequency synchronization. In fact, traditional backhaul network have already adopted Sync-E technology to achieve frequency synchronization. In LTE, the macro station frequency synchronization index is defined as 0.05 ppm to meet the frequency synchronization error requirements of wireless air interfaces. The working principle of SYNC-E is similar with CPRI as both of them use an 8B/10B encoder in the physical layer. The performance of Sync-E could be found in ITU-T G.8262 [10]. In this reference document, it is shown that under free-running conditions, the EEC output frequency accuracy should not be greater than 4.6 ppm with regard to a reference traceable to a clock define in ITU-T G.811 [11]. What's more, the document of G.811 shows if the slave clock has tracked the Primary Reference Clock (PRC), then the maximum allowable fractional frequency offset for observation times greater than one week is 1 part in $10^{11}$, over all applicable operational conditions. Based on these observations, Sync-E can meet demands for frequency synchronization. Therefore, in the NGFI environment, RRSs can acquire frequency synchronization via Sync-E solution in the FH network.

As for the time synchronization, backhaul network with traditional BBUs must fulfill the need for air interface synchronization of the Time-Division-Duplex (TDD) system. According to the 3GPP definition, air interface accuracy should be within $\pm 1.5$ μs. To achieve time synchronization, a potential solution is to use 1588v2 which is a high-accuracy time synchronization protocol based on packetized networks. To implement 1588v2, 1588v2 modules should be added in both the RAU and the RRS. The performance of 1588v2 could be found in the document of G.8273.2 [12], which states that the maximum absolute time error for the clock of class A is 100 ns and for T-TSC Class B is 70 ns. So the accuracy requirement of time synchronization for $\pm 1.5$ μs can be achieved with 1588v2. In fact, the performance of 1588v2 has been verified in China Mobile's commercial networks. Considering the additional accuracy loss by BBU data processing, China Mobile even requires much stricter time synchronization accuracy requirement which is that the time error introduced by Packet Transport Network (PTN) equipment after 30 hops cannot exceed 1 μs [13].

However, when it comes to the time synchronization in the FH network, there are two major issues needed to be solved. The first issue is time hopping. When the RRS obtains time offset information from 1588v2, it needs to compensate this offset. The usual way is to adjust this offset directly, but it brings a vertical step change, which is called time hopping. Time hopping can result in discontinuous transmission in the RRS, which is intolerable for mobile communication. One potential solution is to use frequency-adjustable oscillators to calibrate time in RRS, which can adjust the oscillator frequency gradually to ensure continuous time variation.

Another issue is that cooperative technologies on the PHY level in LTE require higher demands on the time synchronization. For example, for CoMP, synchronization errors between different antennas of different RRH (RRH in the same site

or RRH in different sites) must be less than 130 ns in order to support DL joint delivery functionality between multiple RRHs [14]. This imposes a big challenge for Ethernet-based NGFI even when 1588v2 is leveraged. It cannot be achieved by 1588v2 currently. Therefore, more efforts are needed to figure out how to meet the stricter time synchronization requirement such as considering GPS solutions on the BBU, improving the accuracy of timestamps, increasing clock frequency and optimizing the deviation-adjusting algorithms.

### 5.2.2 NGFI Transmission Jitter Issues and Potential Solutions

NGFI interface transmission jitter is defined as the fluctuation in the time taken for data to be transmitted from the RCC network interface to the RRS network interface. Transmission jitter is mainly reflected in two aspects. First, the wireless equipment on both sides needs to implement data caching to compensate for jitter. The greater the jitter is, the bigger the data buffer and the more complex the equipment will be. Second, the baseband's actual processing time sequence must be designed according to the latest data arrival time. The greater the jitter is, the shorter the available baseband processing time will be.

In CPRI, jitter is negligible while it is common and unavoidable in Ethernet-based NGFI since all the packets are processed in every network node based on the store-and-forward pattern and the presence of transit nodes further increases traffic message jitter, which leads to transportation latency fluctuation. To guarantee wireless performance, transmission devices must be carefully designed or optimized to reduce the jitter and meet the jitter requirements of the NGFI. For example, an appropriate circular buffering may be needed in RRS. On the one hand, data packets can be sorted in order. On the other hand, data is continuously sent to air interface because transportation jitter can be isolated by the buffer.

### 5.2.3 NGFI Transport Delay Issues and Potential Solutions

NGFI interface RCC-RRS data transmission delay in NGFI is defined as the total time required for transmission and switching of data from RCC network interfaces to RRS network interfaces. Compared with traditional CPRI interfaces, new NGFI interfaces will create additional delay losses, mainly due to the additional switching delay in the switched networks. Taking China mobile's current PTN as an example, for PTN equipment, the processing time of one hop is 50 μs. A typical PTN ring consists of 6 nodes (hops) and has 20 km length. The transmission time of fiber for 20 km is 100 μs. In the case of 6 hops, the total delay is $100 + 50*6 = 400$ μs [13]. The maximum transportation delay is another big factor which influences FH transportation performance. This is because that current LTE protocols require the strict interaction time of HARQ on both the UE side and the system side. It is specified by 3GPP that in LTE when the RRH receives a frame from UE, it must respond by transmitting the responding DL frame within 3 ms. The 3 ms

time budget is consumed by the BBU processing, the RRH processing and FH transportation which includes the transmission latency on fiber and processing latency by FH nodes. The more time that the BBU and RRH processing occupy, the lower the allocable budget for FH transportation. Therefore, if RCC-RRS functional partitioning points are placed within the HARQ process, this will influence both data transmission delay and data processing delay. If the transmission delay is extended, it will then be necessary to shorten the processing delay, which will place higher demands on RCC chip processing capacity. If RCC-RRS functional partitioning points are placed outside the HARQ process, there will be too many functions preceding the remote location, which will affect multi-carrier cooperative performance. Therefore, transmission delay is a key factor in NGFI design. And the maximum allowable transport delay for FH networks requires co-design from both wireless and transportation perspectives.

## 6 NGFI PoC

A Proof-of-Concept (PoC) was developed in order to verify the NGFI concept and the feasibility, as shown in the Fig. 7. In this demo packet-based FH networks are targeted, i.e. Ethernet FH networks. The verification of RCC-RRS function split schemes are not carried out here. In this demo two specific Field Programmable Gate Array (FPGA) cards were developed on the BBU and the RRH respectively, which are used to transform Ethernet packets to and from CPRI frames. Taking the DL transmission as an example, the FPGA board on the BBU side encapsulates the

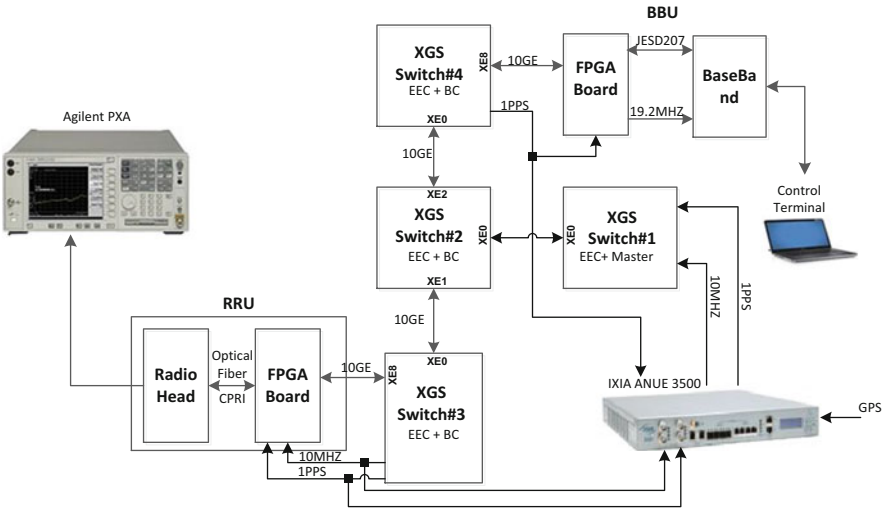


**Fig. 7** NGFI demo architecture

baseband frames into Ethernet packets, and then transmits them to the FH networks. As for the NGFI networks, it consists of three Ethernet switches connected directly. All of them support 1588v2 and Sync-E mechanisms to provide the necessary time and frequency synchronization information between the BBU and the RRH.

In this demo, a synchronization tester was used with a GPS signal as the source input for timing. Next, the time and frequency information is injected to the Ethernet FH networks by a switch device, which served as the master time and frequency server for the NGFI networks. On the RRH site, the FPGA card can recover the time and frequency information and use this to extract and assemble the CPRI frames from the Ethernet packets. In addition, the FPGA card could feed the time and frequency information back to the tester to calculate the difference.

To evaluate the synchronization performance, the Maximum Time Interval Error (MTIE) which is a key indicator for time and frequency synchronization performance, was tested. It was found that the MTIE for time synchronization in this demo is around 15 ns through 3 hops. The frequency deviation is around 0.01 ppb, much lower than the CPRI requirement of 2 ppb. From the results, it can be seen that the Ethernet-based FH networks with specific design could meet the synchronization requirements of CPRI cases.

The Error Vector Magnitude (EVM) which is a key parameter to indicate the wireless performance was also tested. For comparison, a distributed BS with CPRI connection was set up, as shown in Fig. 8. E-UTRA Test Model (ETM) is used to test the wireless performance. The results are shown in Table 3. It is found that the EVM of this PoC is only slightly higher than that of traditional systems. For example, with ETM2 (64QAM) configuration, the EVM of the NGFI demo was 1.17 %, a bit higher than that of CPRI-based distributed BS which was 1.07 %. Both of them meet 3GPP requirements. Therefore, it could be concluded that the adoption of packet-based FH networks has no impact on wireless system performance. The feasibility of the NGFI concept was thus initially verified.

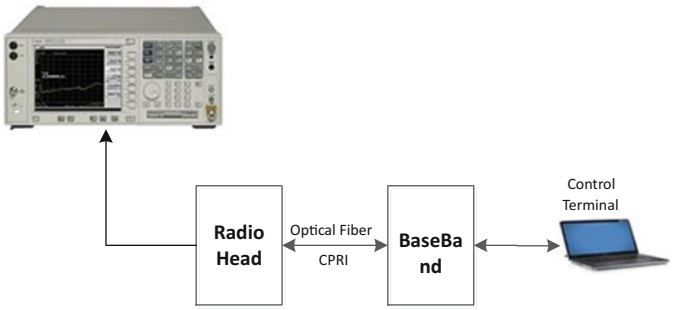


**Fig. 8** Distributed BS with direct CPRI connection

**Table 3** Wireless performance in terms of ETM

| | Traditional BS with CPRI | NGFI with 3 hops | 3GPP requirement |
|---|---|---|---|
| ETM2 (64QAM) | 1.07 % | 1.17 % | 9 % |
| ETM3.1 (64QAM) | 2.13 % | 2.19 % | 9 % |
| ETM3.2 (16QAM) | 3.14 % | 3.21 % | 13.5 % |
| ETM3.2 (QPSK) | 4.50 % | 4.79 % | 18.5 % |

## 7 Conclusions

C-RAN has been viewed as an essential technology for 5G. In this chapter, first the basic concepts, features, benefits and challenges of C-RAN are briefly recalled. One of the major challenges for C-RAN lies on the fronthaul transportation. To address the challenge, a new FH interface called Next Generation Fronthaul Interface is proposed. The proposal of NGFI aims to address the shortcomings of traditional FH interfaces such as low transmission efficiency, poor scalability and ultimately to facilitate large-scale deployment of C-RAN and to support other 5G technologies. The design principles are described and the major impact, challenges and potential solutions on FH transport networks are analyzed. NGFI requires re-design of the BBU-RRH function split and packetization of FH data. By decoupling the FH bandwidth from the antenna number, NGFI can better support large antenna technologies. In addition, the cell-processing functions should be decoupled from the UE-processing functions to make NGFI traffic-aware, which can enable exploiting the statistical multiplexing gain to improve efficiency and further reduce power consumption. It is also suggested that the function split schemes for DL and UL could be different to improve the flexibility and efficiency. The use of Ethernet for NGFI transmission brings the benefits of improved reliability and flexibility due to the packet-switching nature of Ethernet. While Sync-E and 1588v2 could be introduced to address the time and frequency synchronization issues, they still need careful design in order to support collaborative technologies. In the meantime, jitter and latency remain the other key difficulties to overcome to finally realize NGFI. Finally a PoC was introduced with the conclusion that the feasibility of adopting Ethernet as FH networks was initially verified.

In addition to the challenges analyzed in this chapter, in the future there remains a lot of work to do to deeply understand NGFI. For example, the control, data and management channels that are transported via NGFI should be analyzed and carefully designed to make NGFI a better fit in different 5G architectures. In addition, the coexistence of NGFI data and other traffic on the same FH networks needs further examination.

**Acknowledgments** We would like to express our sincere gratitude to our partner, Broadcom for development of the NGFI PoC. We also owe special thanks to all C-RAN team members in China Mobile for helpful discussion and valuable comments.

## References


1. I. Chih-Lin, C. Rowell, S. Han, Z. Xu, G. Li, Z. Pan, Toward green and soft: a 5G perspective. IEEE Commun. Mag. **52**(2), 66–73 (2014)
2. C. M. R. Institute, C-RAN White Paper 3.0: the road towards green ran. [Online]. Available: labs.chinamobile.com/cran, 2014
3. I. Chih-Lin, J. Huang, R. Duan, C. Cui, J.X. Jiang, L. Li, Recent progress on C-RAN centralization and cloudification. IEEE Access **2**, 1030–1039 (2014)
4. CPRI, Common Public Radio Interface (CPRI) Specification (V6.0), Tech. Rep. Aug. 2013. Online: http://www.cpri.info
5. Light Radio Portfolio: Technical Overview. Alcatel Lucent, Feb. 2011
6. Chih-Lin I, Y. Yuan, J. Huang, S. Ma, R. Duan, C. Cui, Rethink Fronthaul for Soft RAN. IEEE Commun. Mag. **53**(9), 82–88 (2015)
7. C. M. R. Institute, White Paper of Next Generation Fronthaul Interface. [Online]. Available: labs.chinamobile.com/cran, 2015
8. A. Davydov, G. Morozov, I. Bolotin, A. Papathanassiou, Evaluation of joint transmission CoMP in C-RAN based LTE-A HetNets with large coordination areas, in *IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013
9. 3GPP TS 36.101, User Equipment (UE) radio transmission and reception (Release 9), Version 8.9.0, Mar. 2009
10. ITU-T Recommendation G.8262/Y.1362, Timing characteristics of synchronous Ethernet equipment slave clock (EEC), Aug. 2007
11. ITU-T Recommendation G.811, Timing characteristics of primary reference clocks, Sep. 2009
12. ITU-T Recommendation G.8273.2/Y.1368.2, Timing characteristics of telecom boundary clocks and telecom time slave clocks, Apr. 2014
13. YD/T 1704-2007, General Technical Requirement for Intelligentized PSTN Network, 2007
14. 3GPP TS 36.104, Base Station (BS) radio transmission and reception (Release 11), Version 11.4.0, Mar. 2013

# User-Centric Wireless Network for 5G

**Yunlu Liu and Guangyi Liu**

**Abstract** In order to better meet the future requirements of mobile Internet and Internet of Things, to fulfill multiple user experience requirements, such as low latency, high data rate, high reliability, low energy consumption, is taken as a key goal of 5G system. Therefore, flatter network architecture, flexible functionality and topology, smart user and traffic awareness, high efficient network operation with lower cost, etc. to facilitate the user-centric wireless network becomes the elements of 5G eco-system access part design.

## 1 Introduction

In order to better meet the future requirements of mobile Internet and Internet of Things, efficiency and agility are crucial to 5G. Specifically, Mobile Internet is to provide advanced user experience, such as Augmented Reality, Virtual Reality, ultra high definition video, mobile cloud. The requirements of Mobile Internet include Gbit data rate, *ms*-level delay, ultra-high traffic density, ultra-high connection density, and user experience guarantee in the high mobility scenarios, such as high speed train, express way. While Internet of Things broadens the objective of mobile communication from people and people to things and things, such as mobile hospital, vehicle to vehicle communication, smart home, industry control. The requirements of Internet of Things include high volume connections, ultra-low delay, low cost, etc.

Based on the study of 5G requirements, the gap analysis between LTE technology and 5G requirements is studied as shown in Table 1. Here we list the most related requirements to wireless network design, i.e., challenging scenarios including connectivity density, ultra-high volume density, diverse services referring to huge difference between the requirements, "0" latency and QoE (Quality of Experience), easy operation including diverse deployment scenarios and low cost.

Connectivity density may be fulfilled by Long Term Evolution (LTE) via massive machine type communication (MTC) technology, as well as increasing the number

Y. Liu (✉) • G. Liu
Wireless Department, China Mobile Research Institute, Beijing, China
e-mail: lyunlu@gmail.com



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_17

**Table 1** Gap analysis and 5G guidelines

| Challenges | Requirements | LTE capabilities | 5G Guidelines |
|---|---|---|---|
| Challenging scenarios | Connectivity density (one million connections/km$^2$) | Might be fulfilled, but not economic | Cost-efficient operation |
| Diverse services | Ultra high volume density (tens of Tbps/km$^2$) | No core network becomes a bottleneck | Flatter architecture and cost-efficient operation |
| | Huge difference between the requirements | No similar processing | Flexible topology and smart RAN |
| Easy operation | "0" latency | No long routing path through core network | Flatter architecture |
| | QoE+ | No blind access network | Smart RAN |
| | Diverse deployment scenarios | No lack of flexibility of network topology | Flexible topology |
| | Lower cost (USD/bit) | Need to be improved | Cost-efficient operation |

of base stations supported by the technologies such as inter cell interference coordination (ICIC) and coordinated multi-point transmission (CoMP) to improve performance [1, 2]. However, due to the limitation of LTE architecture and air interface design, it is inefficient to realize one million connections/km$^2$ requirement in LTE. Thus, the cost-efficient operation is needed.

ICIC and CoMP technologies can help enhance volume density to some extent, while it is hard for the tens of Tbps/km$^2$ volume density 5G requirement. Besides, under the 4th generation network (4G) architecture, all data transferred through the core network, which makes core network a bottleneck to fulfill ultra high volume density requirement. It is possible to be solved by new flatter architecture design and cost-efficient operation.

There are different 5G requirements, which are hardly to be fulfilled by unified procedure as in 4G, e.g., low latency vs. high reliability, enhanced mobile broadband (eMBB) vs. massive MTC. In this circumstance, smart Radio Access Network (RAN) is required to provide traffic awareness and distinction to improve the user experience, which cannot be well provided in LTE RAN. Flexible topology supports self-adaptive topology and functionality to fulfill diverse requirements. Flatter architecture can facilitate ms-level latency.

Due to the fixed network topology, it is hard to realize diverse deployment scenarios. Moreover, the low cost is always one of the most important requirements of communication systems. Therefore, flexible topology and cost-efficient operation are important elements of 5G design.

Therefore, the flatter network architecture, flexible functionality and topology, smart user and traffic awareness, and highly efficient low cost network operation are expected to be key elements of user-centric 5G access network design.

In this chapter, several potential technologies are addressed in order to fulfill the requirements as analyzed above.

User centric 5G access network architecture design: a flatter network architecture, with some functions and entities shifted down to RAN side, can help meet the "0" latency requirement.

Flexible functionality and deployment: to match diverse scenarios in 5G, it is necessary to provide flexible access point deployment and functionality. It may include supporting diverse access points, plug and play access points, and functionality slicing, among other things.

Smart user and traffic awareness and management: to better meet diverse user requirements of the mobile Internet, the 5G system needs to adapt services and user requirements based on user behavior awareness, such as multi-dimension QCI design, personalized local content provisioning, user behavior awareness.

High efficient low cost network operation: in addition to a higher data rate and lower delay, the operation cost efficiency requires flexible and self-organized operations.

## 2 User Centric 5G Access Network Architecture Design

To improve end-to-end (E2E) delay and user experience, the architecture is considered to be much flatter than the earlier systems such as LTE, where some functions are to be shifted down to RAN side, such as multi-RAT coordination, service awareness, localized caching, data processing and forwarding. It helps shorten the traffic path, provide customized service and flexible core-agnostic multi-RAT access.

Besides, flexible spectrum deployment and multi-connectivity have to be taken into architecture design account, to fully use broad range of spectrum, support flexible ultra-dense deployment, and ultimately bring high throughput.

Figure 1 shows the E2E 5G architecture proposed in [3]. It comprises three layers including the infrastructure resource layer, the business enablement layer and the business application layer, an E2E management, as well as an orchestration entity.

The infrastructure resource layer consists of the physical resources of a fixed-mobile converged network, comprising 5G devices (in the form of (smart) phones, wearable devices, machine type modules and others), access nodes, cloud nodes for processing and storing resources [4, 5], networking nodes and associated links. 5G devices may have multiple configurable capabilities and act as a relay/hub or a computing/storage resource, depending on the context [6, 7]. Hence, 5G devices are also considered as part of the configurable infrastructure resource. The resources are exposed not only to higher layers but also to the end-to-end management and orchestration entity through relevant access points. Performance

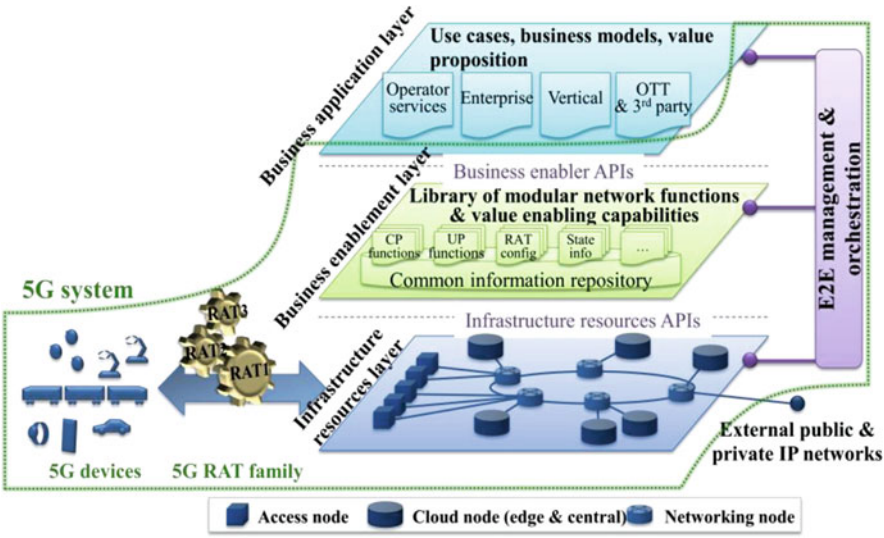


**Fig. 1** 5G Architecture [3]

and status monitoring as well as configurations are intrinsic part of such an access point. The business enablement layer is a library of all functions required within a converged network in the form of modular architecture building blocks. It includes functions realized by software modules that can be retrieved from the repository to the desired location, and a set of configuration parameters for certain parts of the network, e.g., radio access. The functions and capabilities are called upon request by the orchestration entity, through relevant access point. For certain functions, multiple variants might exist, e.g., different implementations of the same functionality which have different performance or characteristics. The different levels of performance and capabilities offered could be utilized to differentiate the network functionality much more than in today's networks (e.g., to offer as mobility function nomadic mobility, vehicular mobility, or aviation mobility, depending on specific needs). The business application layer contains specific applications and services of the operator, enterprise, verticals or third parties that utilize the 5G network. The interface to the end-to-end management and orchestration entity allows, for example, to build dedicated network slices for an application, or to map an application to existing network slices [3].

Based on the E2E architecture design, a paradigm of 5G access network is shown in Fig. 2. In the architecture, one logical module called "General Network Controller" is introduced to fulfill localized control function. The module works together with "Radio data centric" model as a black box to provide unified core-agnostic multi-RAT and diverse access style, and the shifted down functions.

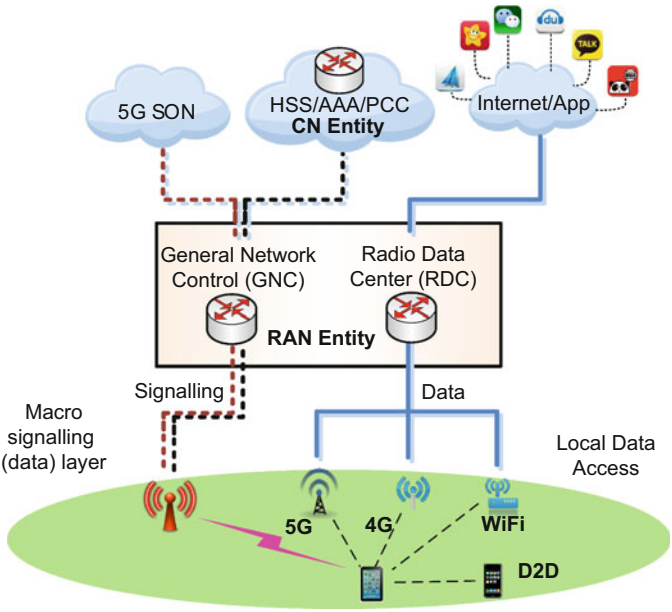


**Fig. 2** A paradigm of 5G access network

- Signaling and data decoupling
  To meet 5G "0" latency requirement, the traffic path is needed to be shorten. While from the point of control plane, it is better to keep a relative high anchor for seamless mobility compared to data plane. It is hard to meet both data and control plane requirements. Therefore, control and data plane decoupling is used to tackle this issue. A method to support control and data plane decoupling is to provide Macro signaling access point and data access point, i.e.,

  – Macro signaling access point
    The macro signaling access point is responsible for initial access and system information for data access point.
  – Data access point
    Data access point is activated as required. The transmission efficiency can be supported by several technologies such as: massive MIMO, beam forming, positioning.

    Usually, there are two evolutionary ways from LTE to 5G, i.e.,

    Option 1: At the first stage, signaling access point is upgraded from eNB, data access point is from 5G as shown in Fig. 2. Then, 5G system can work standalone at the second stage,, which means both signaling and data access points are based on 5G.

    Option 2: 5G system is standalone without co-existence with LTE.

    Although the technical routine may be different of the two ways, the objective of 5G design is kept the same, i.e., a flatter, flexible and user centric system.

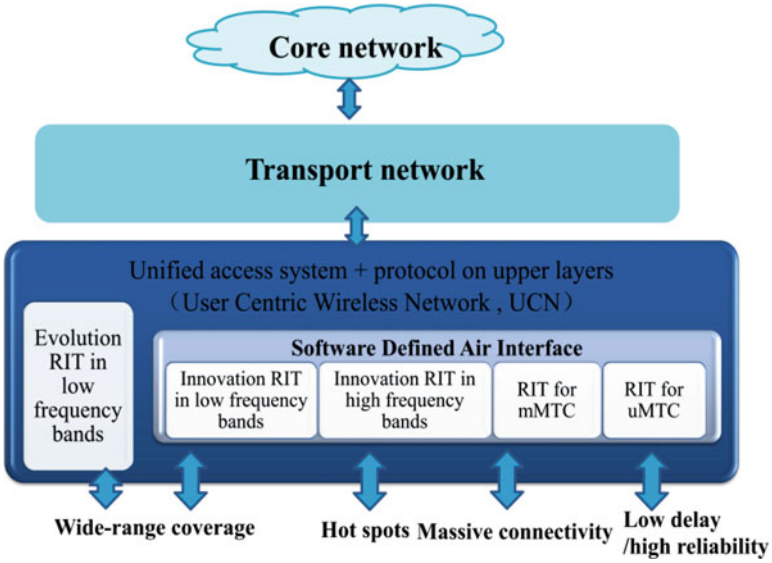


**Fig. 3** An example of Multi-RIT access networks

- Data plane functionality decentralization
  To meet the low latency and diverse 5G requirements, some data plane functionalities are to be put on the edge of the network to facilitate local breakout/switch/data provision.
- Multi-RIT (Radio Interface Technology) access
  In 5G, it is necessary to provide multiple RITs to support diverse application requirements. An example is shown in Fig. 3. The evolution RIT in low frequency bands, innovation RIT in low frequency bands, innovation RIT in high frequency bands, RIT for mMTC (massive machine type communication), and RIT for uMTC (ultra-reliable machine type communication) should be provided. Evolution RIT in low frequency bands and innovation RIT in low frequency bands are to support wide-range coverage. Innovation RIT in high frequency bands, RIT for mMTC is to support hot spot scenario. Massive connectivity is supported by RIT for mMTC. Moreover, RIT for uMTC provides low latency and high reliability access. Although multiple RITs are necessary to meet diverse requirements, unified access system via software defined air interface is to be supported to decrease system/UE (User Equipment) complexity and enhance user experience.

## 3 Flexible Functionality and Deployment

Network/device functions and RAT configuration have to be tailored for each use case, leveraging the Network Function Virtualization (NFV) and Software Defined Network (SDN) concepts. Thus, the network should support flexible composition of network functions, as well as, their flexible allocation and location. The network functions need to be scalable so that capacity is provided when and where needed. Even when particular functions or nodes become unavailable, e.g., due to disaster events, the system should support graceful degradation instead of service interruption. To improve such robustness, the state information should be split from functions and nodes so that contexts could be easily relocated and restored even in failure events. 5G aims to virtualize as many functions as possible, including the radio baseband processing. Although some functions may still run on non-virtualized platforms, e.g., to meet state-of-the-art performance targets, they should be programmable and configurable using C-plane functions according to Software Defined Network (SDN) principles [8].

### *3.1 Flexible Deployment*

As shown in Fig. 4, diverse low cost access points are to be supported in 5G. Based on the classification according to functionality, there are coverage access point and hotspot access point. The functionality of coverage access point is to guarantee the basic coverage and uniform access. It can be the anchor of mobility and also the signaling access point due to the relative larger coverage.

Hotspot access point is to fulfill the high throughput requirements in certain area. It needs to be plug and in equipment deployed in hotspot areas as required.

According to level support of protocol stack, there are full stack access point and semi-stack access point.

The former includes PHY, MAC, RLC, PDCP and RRC for control plane or IP for user plane. The latter inherits the protocol stack of eNB (evolved Node Base-station) in LTE to provide complete protocol treatment including IP, PDCP, RLC, MAC and PHY.

Semi-stack access point is proposed to simplify the design of access point and reduce cost. It is expected to be low cost and easy deployed.

### *3.2 Flexible Functionality*

Software and hardware decoupling of network equipment is the fundamental of flexible functionality. Thus, diverse network functionalities can be flexibly implemented as required. In this circumstance, the network provides multiple

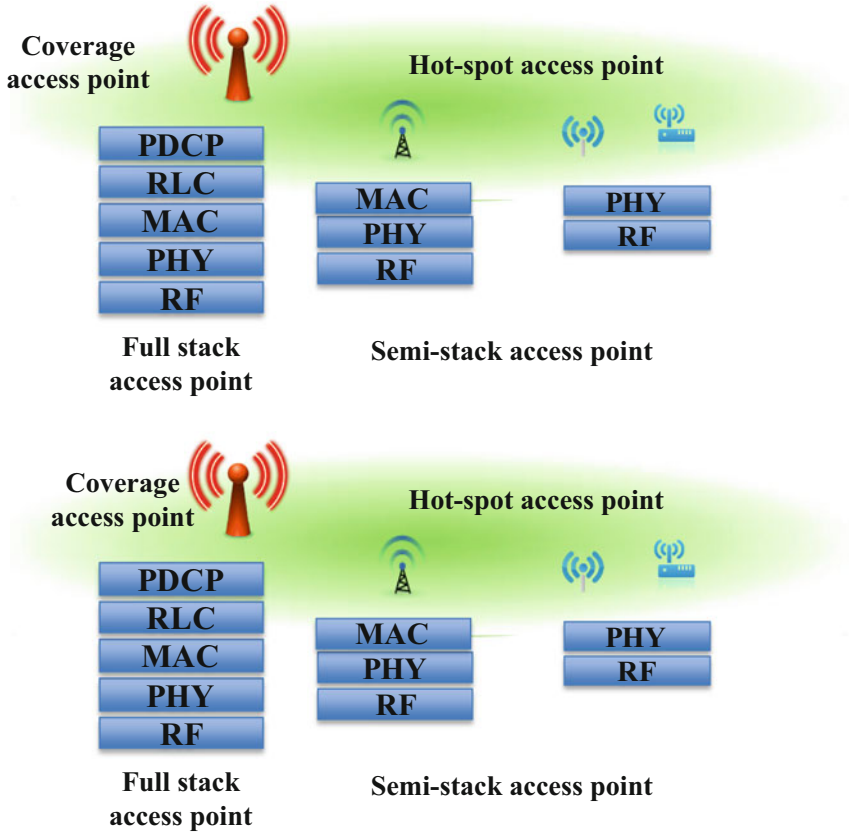


**Fig. 4** A paradigm of flexible network deployment

functionality modules to fulfill diverse requirements. A group of functionalities to meet certain requirements constitute a slice. Figure 5 presents an example, which has three different slices, i.e., eMBB (enhanced mobile broadband) slice for mobile Internet service, mMTC slice for massive machine type access, uMTC slice for traffic with low latency and high reliability requirement. The functionality modules for different slices may be different. As shown in Fig. 5, eMBB slice may use content and service module implemented at RAN side for content awareness/analysis/local caching for hot content, and forwarding function module at RAN side for local breakout for low latency service. For example, the regional user and service requirements are analyzed via big data algorithms with user and traffic information. Then, the corresponding network functionality module can be deployed according to the analysis e.g., to deploy Multimedia Broadcast Multicast Service (MBMS) functionality module in areas with the broadcasting requirement, to deploy device to device (D2D) functionality module in areas with the group communication requirement, and to deploy safety functionality module in areas with

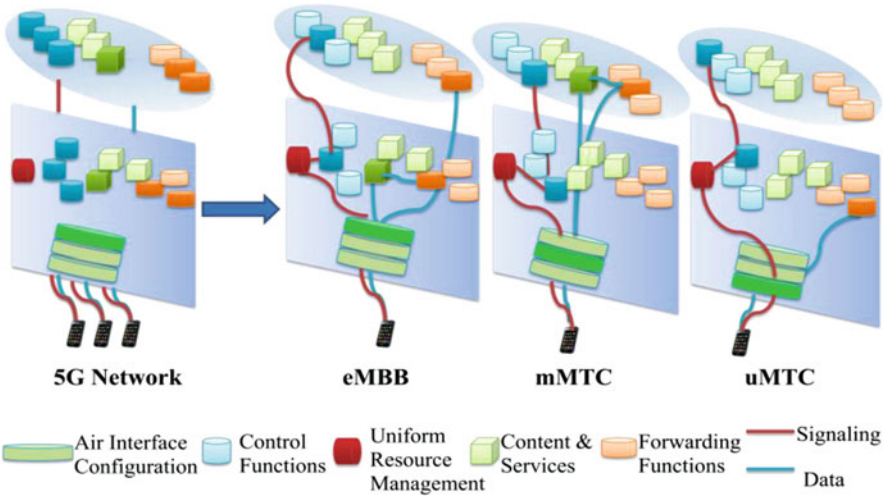


**Fig. 5** A paradigm of network slicing

the high requirement of security. mMTC slice can use content and service module implemented at core network side for data processing, and forwarding function module at core network side for latency insensitive traffic. uMTC slice directly uses forwarding function module at RAN side for local breakout to meet low latency requirement.

Regarding the control plane, all slices go through the uniform resource management module to provide unified radio access regardless of air interface technology.

## *3.3 Local Breakout*

As described in the flexible functionality section above, the traffic can be transmitted by a short path via local breakout in order to meet the low latency requirement. Figure 6 shows the architecture of local breakout. The data can be directly steered to local network or Internet on the edge. Signaling traffic can continue to traverse the mobile operator network. It can not only shorten the transmission time, but also release the traffic pressure at the core network. There are several potential solutions, e.g., LIPA (Local IP Access) and SIPTO (Selected IP Traffic Offload) are proposed in 3GPP [9]. The main idea of LIPA is to access local network through local gateway (L-GW). SIPTO is to steer part of the traffic to Internet directly through L-GW. There are usually two deployment manners, i.e., collocated L-GW which may have mobility issue among base stations, and non-collocated L-GW with mobility issue among L-GWs, each of which covers several base stations. There are still several open issues before commercial ready, such as the mobility issue described above, as well as UE and core network complexity. Therefore, studies on local breakout are

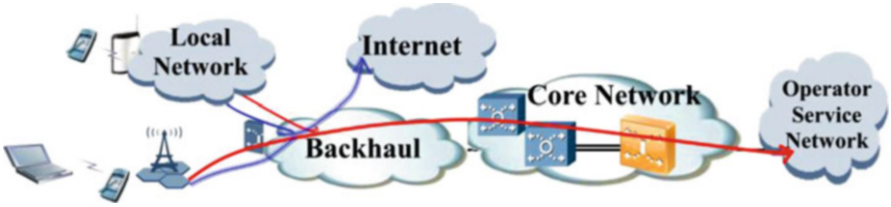


**Fig. 6** A paradigm of local breakout

not limited to this and go on broadly. No matter which solutions are to be put to use, local breakout is an inevitable branch of user-centric 5G network.

## 4 Smart User and Traffic Awareness and Management

Smart user and traffic awareness and management, together with user-centric 5G access network architecture and flexible deployment and functionality provide the foundation of user centric network.

### *4.1 Multi-Dimension QCI [QoS (Quality of Service) Class Indicator] Design*

QCI is introduced in LTE to meet the requirements of diverse traffic. Radio access network schedules and allocates wireless resource according to the indication from the core network via QCI. As shown in Fig. 7, the QoS management in LTE is a hierarchical E2E system, which can be divided into two parts, i.e., EPS bearer traffic and outer bearer traffic. The traffic of the same bearer has the same QoS guarantee.

There are several reasons that it can also not fulfill the requirements of diverse user and traffic addressed above.

1. There are only nine kinds of QCI in current standardization. It is not enough for much more diverse traffic requirements of 5G network.
2. The current Access network is “blind”, and knows nothing about users and traffic. The strategy of wireless resource schedule and allocation is from the QCI indication of the core network. It is relatively complicate and inefficient. 5G wireless network should have the capability of awareness and adaptation of the user and the traffic.
3. The difference of user requirements and behavior is not taken into consideration of network design before 5G. To better meet the diverse user requirements and improve user experience, the user requirements and behavior, e.g., daily mobility track, preferred traffic content, have to be considered in 5G network design.

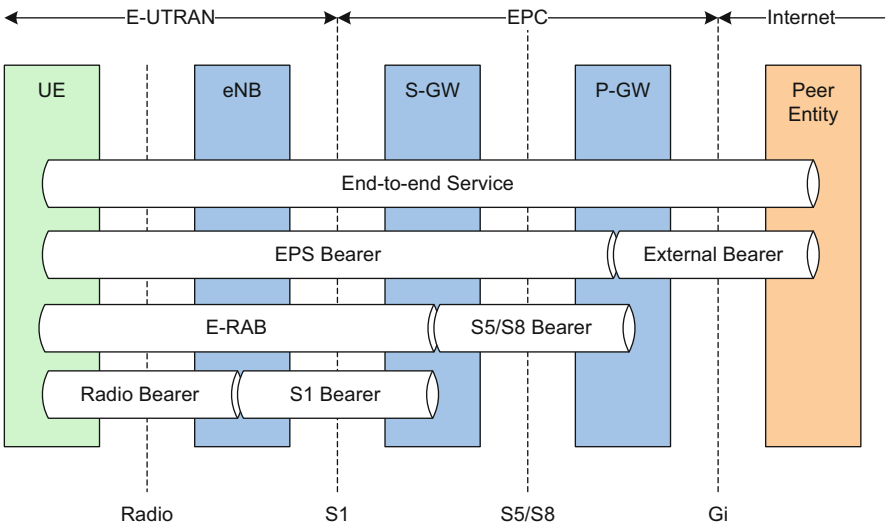


**Fig. 7** EPC bearer architecture in LTE systems

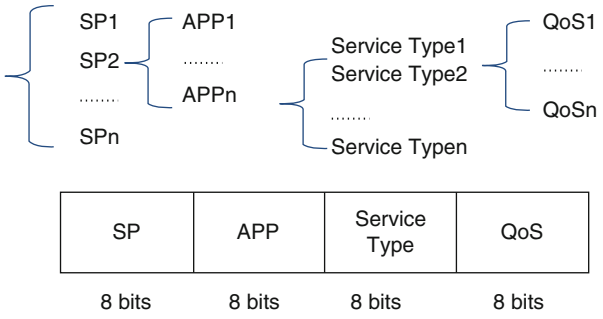


**Fig. 8** Hierarchical QoS system

4. The consistency of user experience is an important requirement of 5G network, especially for dense or high speed scenarios, such as stadium, high speed railway.

To better meet the requirements of diverse traffic in 5G network, new QoS guarantee mechanism is to be designed. A paradigm of new QoS guarantee mechanism is addressed in this chapter.

The mechanism is designed based on hierarchical bit-mapping. The information includes service provider, access point name, traffic type, QoS requirements, and so on. Let us take 32 bits QoS parameter as an example. As shown in Fig. 8, the first 8 bits represents service provider, the consequent 8 bits are used for application name, the next 8 bits is for traffic type, and the last 8 bits indicates QoS parameter. The information is helpful for wireless resource scheduling and allocation.

### 4.2 Personalized Local Content Provisioning

Personalized local content provisioning is an important target for the user-centric 5G network. It includes user and service awareness at the RAN side, local caching and management.

The access network used to be user and service agnostic, which is a "blind pipe" connecting the user to the core network. In order to meet 5G requirements such as low latency and user-centric requirements, the access network needs to be user and service aware. It may help facilitate local content provisioning. Big data analytics can predict user requirements with mining the data from the network [10–14]. Then, the requirements are checked locally if they can be met via local content.

Local caching and management are the basis to fulfill local content provisioning via monitoring and analyzes of source traffic, providing a local copy. It can reduce E2E latency and enhance user experience.

- User and traffic information acquisition
  The first step is to collect traffic attribute via packet analysis, which may include the application type (FTP, HTTP, etc.), server address, port number, traffic content, etc. Then, this data is analyzed comprehensively via clustering algorithm to label the traffic, e.g., as sports news, entertainment video, or romance/action movie.
- User requirement analysis and prediction
  The big data analysis algorithm, especially recommendation algorithms can recommend contents that the user may be interested in by considering traffic labels, user attributes, terminal types, etc.
- Local caching and content management
  The popular content is to be copied locally, as in the case of a downloaded movie, or to be downloaded from the application server and cached locally, as in the case of sports news, by matching the traffic labels with the content users may be interested in.
- Content provisioning
  When the user initiates an application request, the system will check whether the corresponding content is already locally cached and thus can be sent directly. On the other hand, according to the content recommendation via big data analysis, the system can check if it is cached locally and push to the user directly.

### 4.3 User Behavior Awareness

Wireless resources need to be optimized according to user and service requirements in 5G to improve efficiency and the user experience.

It is found that a 93 % potential predictability in user mobility (route and time) across the whole user base by studying the mobility patterns of anonymized mobile phone users [15]. As shown in Fig. 9, some passengers often get on the train

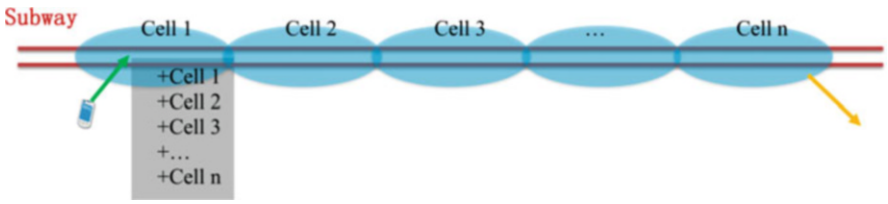


**Fig. 9** A paradigm of user behavior awareness

at a certain station via subway or high-speed rail, and get off at another station. The motion trajectory can be predicted via big data analysis to pre-configure the network.

The access points can record the historical access point list of each user, and directly upload it to a central processing module or to a target access point when the serving access point is changed. Then, the collected user motion data is analyzed to predict the motion trajectory via the big data algorithms. As shown in Fig. 9, when access point 1 (Cell 1) get the predication that the user motion trajectory is from Cell 1 to Cell n via Cell 2, 3, and n-1, it will send the configuration of Cell 2, 3, . . . ., n to the corresponding terminal to facilitate pre-configuration of these cells.

## 5 High Efficiency Network Operation

Operation efficiency is an important requirement in 5G network design. Self operation network (SON) techniques have been studied in LTE, including PCI optimization, self-configuration of neighbor cell, minimization of drive test (MDT), mobile load balancing, energy saving, etc. However, there is no mature technique to let the SON work well in practical systems so far due to the limitation of LTE network architecture. Therefore, SON should be taken into consideration at the stage of 5G network design.

The high efficiency network operation system includes traffic sensing, user capability sensing, centralized and personalized functionality management and distribution, unified wireless platform, among other modules.

The system is divided into two parts, i.e., the decision making domain and the implementation domain. The decision making domain is mainly responsible for the collection and management of user information, as well as service and terminal state, network upgrade, selection, configuration and unified management, etc. The implementation domain is responsible for the user, terminal and network status reporting, network configuration and dynamic deployment, etc.

Figure 10 shows a paradigm of big data based network operation. It includes user information collection and strategy making which belong to decision making domain, as well as UE context management and function configuration which compose implementation domain.

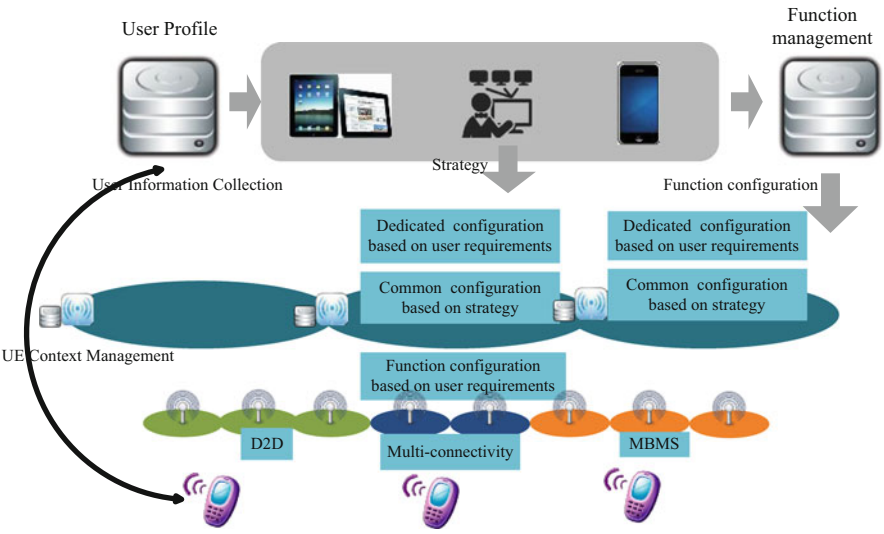


**Fig. 10** A paradigm of big data based network operation

The decision making domain provides the basic configuration to initialize the network. Then, it chooses and deploys the functionalities and access points which meet the user requirements. The implementation domain builds multi-connectivity bearers with terminals via dynamic access points, functionality and configuration according to the requirements after receiving the personalized configuration.

The efficiency of the big data based network operation can be maximized, through optimal resource allocation (frequency, time, antenna, power, etc) to each access point and to each user, such that the required service is provided from the optimal set of access points with optimally allocated resources. For example, in the 5G ultra dense deployment scenario, it may be a problem for users to have too many neighbor cells, which may be redundant and consume too much unnecessary power, and sometimes interfere with each other. To resolve this issue, big data algorithms can analyze the user and service requirements with the collection of network status. Then, the system can choose the corresponding configuration of neighboring cell according to the analysis. For example, it is possible that not all the access points support downlink 256QAM. If a service needs the downlink 256QAM functionality, only those access points with this functionality need to be configured as the neighboring cells, with remaining neighboring cells turned off for power saving. The decision making domain needs to have the whole picture of functionality distribution in the network and the requirements of users and services. The implementation domain configures the terminal according to the command from the decision making domain to fulfill user-centric neighboring cell configuration.

## 6 Chapter Summary

In this chapter, the concept of user-centric wireless network for 5G is addressed to fulfill multiple user experience requirements in 5G. Four key technical directions are studied based on the gap analysis between LTE technology and 5G requirements, i.e., user centric 5G access network architecture design, flexible functionality and deployment, smart user and traffic awareness and management, and high efficient low cost network operation. These key technologies work together with cross-layer and E2E solutions [16–19] to provide the user-centric 5G eco-system.

## A.1 Annex-User Experience Requirements Analysis [3]

To design a user-centric 5G network, it is better to study carefully of 5G user experience requirements. In this part, the analysis of user experience requirements, including consistent user experience, user experience data rate, latency, mobility, is provided from.

User experience requirements address the end users perception of technology when consuming one or more services. User experience will have to be managed in highly heterogeneous environment and under different user scenarios/contexts.

## A.2 Consistent User Experience

The 5G system should be able to deliver a consistent user experience over time for a given service everywhere the service is offered.

Consistent user experience is defined by service-dependent minimum KPIs (e.g. data rate, latency) being met over the service coverage area, with a level of variation configurable by the operator. These service-dependent KPIs are for further study. A consistent user experience across time and space depends obviously on the technology performance and capabilities, and on the operator deployment. The requirements address only the technology performance and capabilities.

## A.3 User Experienced Data Rate

Data rate requirements are expressed in terms of user experienced data rate, measured in bit/s at the application layer. The required user experienced data rate should be available to at least 95 % of the location and time in the considered environment

(i.e. at the "cell-edge"). The user experienced data rate requirement depends on the targeted application/use case. It is set as the minimum user experienced data rate required for the user to get a good experience of the targeted application/use case.

Use case specific user experienced data rates up to 1 Gb/s should be supported in some specific environments, like indoor offices, while at least 50 Mb/s shall be available everywhere at a reasonable cost.

## A.4 Latency

When considering latency requirements, the following metrics are considered:

- E2E Latency: Measures the duration between the transmission of a small data packet from the application layer at the source node and the successful reception at the application layer at the destination node and back.
- User Plane Latency: Measures the time it takes for a small data packet from user terminal to the L2/L3 interface of the 5G system destination node and back.

The E2E latency is the latency perceived by the end user. It accounts for the time needed for the data packet to cross all the nodes up to the application server and back, which includes nodes of the 5G system and nodes potentially outside the 5G system. In contrast, the user plane latency is limited to the 5G system only. Both latency metrics approximately coincide when the application server is located within the 5G system. In the latter case the latency is minimized when the application server is co-located with a radio node, e.g., the radio base station or another user terminal (for the case of device-to-device communication). As a result, the requirements on minimum latency are expressed in terms of E2E latency.

The 5G system should be able to provide 10 ms E2E latency in general and 1 ms E2E latency for the use cases which require extremely low latency. Note these latency targets assume the application layer processing time is negligible to the delay introduced by transport and switching.

The 5G system should also give the end user the perception of being always connected. The establishment of the initial access to the network (or status change from idle state to connected) should then be instantaneous from the human user perspective.

## A.5 Mobility

Mobility refers to the system's ability to provide seamless service experience to users that are moving. In addition to mobile users, 5G use cases identify that 5G networks will support an increasingly large segment of static and nomadic users/devices. 5G solutions therefore should not assume mobility support for all devices and services but rather provide mobility on demand only to those devices

and services that need it. In other words, mobility on-demand should be supported, ranging from very high mobility, such as high-speed trains/airplanes, to low mobility or stationary devices such as smart meters.

The mobility requirements are expressed in terms of the relative speed between receiver and transmitter at which consistent user experience should be ensured.

## References


1. K. Zheng, Y. Wang, W. Wang, M. Dohler, J. Wang, Energy-efficient wireless in-home: the need for interference-controlled femtocells. IEEE Wirel. Commun. **18**(6), 36–44 (2011)
2. M. Mukherjee, L. Shu, V. Kumar, P. Kumar, R. Matam, Reduced out-of-band radiation-based filter optimization for UFMC systems in 5G, in *Proceedings of the 11th International Wireless Communications and Mobile Computing Conference (IWCMC 2015)*, Dubrovnik, Croatia, 24–27 August 2015
3. NGMN 5G White Paper, NGMN Alliance, March 2015
4. W. Fang, Y. Li, H. Zhang, N. Xiong, J. Lai, A.V. Vasilakos, On the throughput-energy tradeoff for data transmission between cloud and mobile devices. Inform. Sci. **283**, 79–93 (2014)
5. W. Fang, X. Yin, Y. An, N. Xiong, Q. Guo, J. Li, Optimal scheduling for data transmission between mobile devices and cloud. Inform. Sci. **301**, 169–180 (2015)
6. K. Zheng, B. Fan, Z. Ma, G. Liu, X. Shen, W. Wang, Multihop cellular networks toward LTE-advanced. IEEE Veh. Technol. Mag. **4**(3), 40–47 (2009)
7. M. Mukherjee, L. Shu, Y. Zhang, Z. Zhou, K. Wang, Joint power and reduced spectral leakage based resource allocation for D2D communications in 5G, in *15th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP)*, Zhangjiajie, China, 18–20 November 2015
8. IMT-2020(5G) PG-white paper on 5G vision and requirements, China IMT-2020 Promotion Group, May 2014
9. 3GPP TR23.859 V12.0.1, Local IP access (LIPA) mobility and Selected IP Traffic Offload (SIPTO) at the local network (Release 12)
10. K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, W. Xiang, K. Yang, Big data driven optimization for mobile networks towards 5G. IEEE Netw. Mag. **30**(1), 44–51 (2016)
11. C.-L. I, Y. Liu, S. Han, S. Wang, G. Liu, On big data analytics for greener and softer RAN. IEEE Access **3**, 3068–3075 (2015)
12. J. Wu, H. Liu, H. Xiong, J. Cao, J. Chen, K-means-based consensus clustering: a unified view. IEEE Trans. Knowl. Data Eng. **27**(1), 155–169 (2015)
13. Y. Chen, L. Shu, A.M. Ortiz, N. Crespi, L. Lv, Locating in crowdsourcing-based dataspace: wireless indoor localization without special devices. Mob. Netw. Appl. **19**(4), 534–542 (2014)
14. Y. Chen, A.M. Ortiz, N. Crespi, L. Shu, L. Lv, Reality mining: digging the impact of friendship and location on crowd behavior, in *Mobile and Ubiquitous Systems: Computing, Networking (MobiQuitous), and Services*, 2013, pp. 142–154
15. C. Song, Z. Qu, N. Blumm et al., Limits of predictability in human mobility. Science. **327**(5968), 1018–1021 (2010)
16. J. Wang, J. Wen, Y. Han, J. Zhang, C. Li, Z. Xiong, CUBIC-FIT: a high performance and TCP CUBIC friendly congestion control algorithm. IEEE Commun. Lett. **17**(8), 1664–1667 (2013)
17. J. Wang, J. Wen, Y. Han, J. Zhang, C. Li, Z. Xiong, DC-Vegas: a delay-based TCP congestion control algorithm for datacenter applications. J. Netw. Comput. Appl. **53**, 103–114 (2015)
18. J. Wang, J. Wen, Y. Han, J. Zhang, C. Li, Z. Xiong, Achieving high throughput and TCP Reno fairness in delay-based TCP over large networks. Front. Comput. Sci. **8**(3), 426–439 (2014)
19. H. Wang, S. Chen, H. Xu, M. Ai, Y. Shi, SoftNet: a software defined decentralized mobile network architecture toward 5G. IEEE Netw. **29**(2), 16–22 (2015)

# Energy Harvesting Based Green Heterogeneous Wireless Access for 5G

**Hang Li, Chuan Huang, Fuad E. Alsaadi, Abdullah M. Dobaie, and Shuguang Cui**

**Abstract** A worldwide challenge for the design of future cellular systems is to meet the increasing energy demand, while, on the other hand, to lower the emission of greenhouse gases for achieving the environment sustainability. A feasible and efficient method to tackle this issue is to let the communication systems harvest energy from renewable energy sources instead of fossil fuels. However, by employing the energy harvesting (EH) technique, the instability of renewable energy resources introduce new challenges on the design of the upcoming 5G systems. In this chapter, we focus on uplink access schemes and power allocations for EH based heterogeneous networks. First, a heterogeneous access model incorporating EH based mobile users is proposed and followed by a throughput maximization framework. Then, by classifying transmission policies into two main categories (i.e., single-channel vs. multi-channel scenarios), the proposed framework is concretized under various practical conditions, including the availability of central control, causality of harvested energy, channel state information, and others. Finally, future research directions and open problems are discussed.

H. Li
Department of Electrical and Computer Engineering, Texas A&M University, College Station, USA
e-mail: david_lihang@tamu.edu

C. Huang
National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu, China
e-mail: huangch@uestc.edu.cn

F.E. Alsaadi • A.M. Dobaie
Department of Electrical and Computer Engineering, King Abdulaziz University, Jeddah, Saudi Arabia
e-mail: fuad.alsaadi@gmail.com; adobaie@kau.edu.sa

S. Cui (✉)
Department of Electrical and Computer Engineering, University of California, Davis, USA
e-mail: sgcui@ucdavis.edu



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_18

## 1 Introduction

As data service is prioritized by service operators, it now enjoys an unprecedented popularity across the world. Compared to the voice and text services, data service incurs more intensive communication usages at mobile users (MUs), like Internet surfing, picture uploading, video streaming, and many others, which could quickly deplete the energy stored in the battery. It is not surprising that the energy cost of the cellular system keeps growing [11]. Taking the electrical energy consumption of all cell phones as an example: a regular cell phone consumes about 4 kWh [27] annually, and the number of cell phones worldwide is about seven billion in 2015 [13], which approximately requires us to burn 15 million tons of coal at each year [35]. Although significant efforts have been made such as developing energy efficient transmission technologies to save energy and prolong battery lifetimes, the high energy cost and high carbon-oxygen emission issues have not been fundamentally solved. When further considering the energy consumption at the base stations (BSs) and data centers, such an energy challenge will inevitably draw both environmental and financial concerns for service operators and mobile users [14].

The 5G system, as a new generation of the cellular system, is expected to mitigate the above issues by becoming "green": lowering both the energy cost and greenhouse gas emission. The most effective way to achieve this is to shift the paradigm of energy sources from the conventional fossil fuels to renewable energy sources. Recently, with the rapid development of hardware technologies, energy harvesting (EH) techniques were developed, capable of efficiently converting the environmental energy (e.g., solar, wind, thermal, vibration, etc.) or man-made energy (e.g., vibration, RF) into electric energy. So far, many prototypes of EH based wireless sensors have been proposed [32] and various EH products have been commercialized [25]. As such, EH based wireless devices are expected to enjoy large-scale applications in 5G systems.

One important characteristic of 5G wireless systems is to achieve an aggregate data rate (characterized by bits/s per unit area) more than 1000× higher than that in 4G. The heterogeneous network (HetNet) architecture, which incorporates multiple radio access technologies (RATs), is a promising solution to realize such a capacity boosting. As such, the 5G system could support not only new standards (e.g., mmWave based ones with device-to-device (D2D) communications [2, 36]), but also the conventional 3G, 4G, and WiFi standards.

For the conventional battery or grid powered HetNet, the design of access and power allocation policies only accounts the limitations on the average and/or peak power consumptions, which are usually fixed over the whole operation period. When the transmitter is powered by energy harvesters, however, the design of the transmission policies faces a new type of constraints, named *EH constraints*, i.e., the power consumption is limited by the accumulated energy up to the current time. Besides, the ambient energy sources are usually random and intermittent. Thus, the

amount of harvested energy per unit time may vary, which brings extra unreliability to the system performance. In addition, hardware related limitations also need to be accounted into the transmission policy design for practical purposes, which include the battery storage efficiency, the full/half energy-duplex consideration, and the supporting circuit power overhead.

In this chapter, we consider the design of the transmission policy (i.e., the control of the channel access and transmission power) for EH-based heterogeneous networks, where the network provides either single-channel or multi-channel access for MUs, and each MU is equipped with an EH-based power supply. We propose an optimization framework to determine throughput-optimal transmission policies, where such a framework incorporates various practical considerations, such as finite time horizon, causality of the state information, and others. Following this framework, we also introduce the state-of-art transmission policies by describing their application scenarios and how they achieve efficient operations.

The rest of this chapter is organized as follows. In Sect. 2, we first present a heterogeneous wireless access model, and then consider a general optimization framework to maximize the average sum rate. In Sects. 3 and 4, we introduce transmission policies in single-channel and multi-channel scenarios, respectively. Finally, in Sect. 5, we summarize the existing access and power allocation policies, and discuss future research problems in EH based heterogeneous wireless access.

## 2 System Model and Optimization Framework

Given a targeted area, there are multiple EH-based mobile users (MUs) that are ready to transmit their data. To meet MUs' transmission requirements, different types of base stations (BSs) serve as wireless access points for the MUs. To fully and efficiently coordinate the BSs, a cloud-based radio access network (C-RAN) architecture has been proposed in [7], where a virtual BS pool performs the centralized management, including joint signal processing, inter-cell interference (ICI) mitigating, radio management, and MU scheduling. Here we adopt C-RAN in our framework for the heterogeneous access as shown in Fig. 1.

A BS can be a marco cell station, small cell station, or WiFi access point, and thus multiple RATs can be supported by multiple BSs handling the signals transmitted from MUs over certain radio frequencies. The signals propagate through a certain wireless medium, which could be divided into multiple parallel channels via various schemes [41], e.g., frequency division multiple access (FDMA), code division multiple access (CDMA), or orthogonal frequency division multiplexing (OFDM). In Fig. 1, there are $M$ parallel channels for $N$ MUs. A particular channel could be reused over space by different BSs. Note that if two BSs are too close, like BS 1 and BS 2, and both use channel 1, serious interference may be caused. Although parallel channels could be created in many ways, the total number of independent channel usages (the degree of freedom) is finite and usually smaller than the number of MUs in most practical multiple access scenarios. This requires

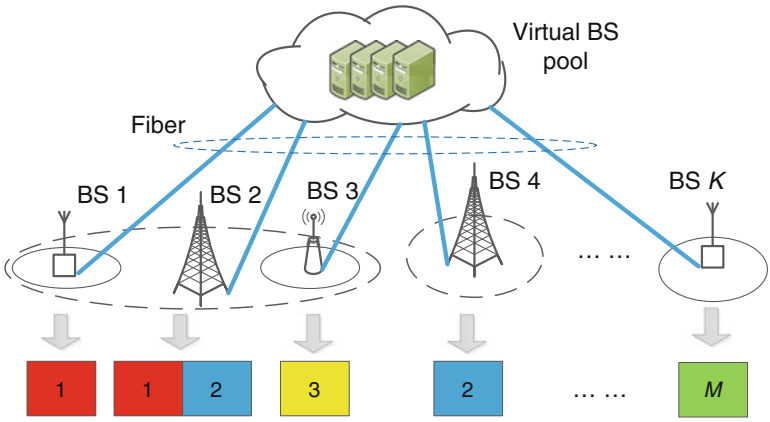


**Fig. 1** Network architecture of the C-RAN [7] with heterogeneous wireless access, where $K$ BSs are connected to the virtual BS pool via fiber links, and different BSs may use the different or the same wireless channels

the system designer to develop smart algorithms that can efficiently manage the channel resources so as to maximize the desired utility. Here, we only focus on the uplink wireless accesses of multiple MUs and assume that the C-RAN can perfectly deal with all the jobs at the receiver side.

## *2.1 System Model*

For the purpose of exposition, we tailor the system model of C-RAN and consider a simplified heterogeneous wireless access model as shown in Fig. 2, involving $M$ parallel synchronized channels for $N$ EH-based MUs. Each MU can each access more than one channels. The system is slotted over time with unit slot length. The considered time horizon is $T$, i.e., $1 \leq t \leq T$, where $t$ is the slot index. Moreover, $T$ can be either deterministic or random, which depends on the utility of interests. The deterministic case is commonly used in most studies, and $T$ is either finite or infinite. In some applications, $T$ can also be treated as a random variable, which will be shown later.

For the $n$th MU, the EH rate (i.e., energy harvested per slot) at time slot $t$ is denoted by $E_n(t)$, and the battery energy level at the beginning of time slot $t$ is denoted by $B_n(t)$. In general, the power consumption at the MU can be roughly divided into two parts: circuit power and transmission power. Note that the circuit power can be usually modeled as a constant value (see Sect. IX in [34]) that can be approximately treated as a constant offset in the total power consumption. Thus, to investigate the limit of the system performance, it is reasonable to first ignore

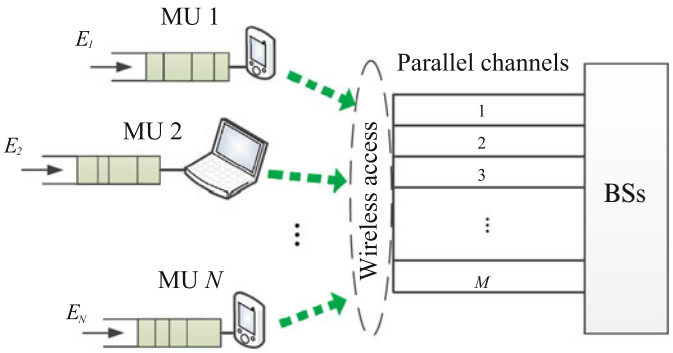


**Fig. 2** Access model for the heterogeneous network with EH-based MUs. Multiple BSs provide in total of $M$ parallel channels for the $N$ EH-based MUs, and each MU can access more than one channels

the circuit power consumption and purely focus on the transmission power control. Suppose that the total transmission power of MU $n$ during time slot $t$ is $P_n(t)$ (which is also the transmission energy given the unit slot length); the battery evolution can be expressed as

$$B_n(t+1) = \max\{0, \min\{B_n(t) + E_n(t) - P_n(t), B_{n,max}\}\}, \tag{1}$$

where $B_{n,max} > 0$ is the battery capacity.[1] We call $\{E_n(t), B_n(t)\}_{1\le n\le N}$ as energy state information (ESI) at time $t$. In general, for the $n$th MU, $\{E_n(t)\}_{t\ge 1}$ is a random process whose statistics is often viewed as a prior knowledge.

In our considered system, each MU could access more than one channels simultaneously. We let $s_{n,m}(t)$ be an indicator function such that

$$s_{n,m}(t) = \begin{cases} 1, \text{ the } n\text{th MU accesses channel } m \text{ in time slot } t; \\ 0, \text{ otherwise.} \end{cases} \tag{2}$$

We want to point out that $s_{n,m}(t)$ can be either deterministic or random depending on the adopted access scheme. For example, if a central controller was in charge of the access control and directly assigned channel $m$ to the $n$th MU at time $t$, there would be $s_{n,m}(t) = 1$, which is deterministic. However, if the controller randomly assigned a channel to the $n$th MU, $s_{n,m}(t)$ would become a random variable depending on the channel selection policy. A typical selection is to assign the channel with the highest channel power gain at time $t$ to the $n$th MU such that the instantaneous transmission

[1] Sometimes $B_{n,max}$ could be treated as positive infinity when the EH rate is relatively small compared to the batter capacity. For example, an AA-sized NiMH battery has a capacity of 7.7 kJ, which may require a couple of hours to be fully charged by some commercial solar panels (1 W $\sim$ 400 W) [21, 30], which generate relatively small $E_n(t)$.

rate is maximized. On the other hand, in a distributed access scenario, if the $n$th MU randomly accessed channel $m$, the indicator $s_{n,m}(t)$ would be a random variable. In short, access variables $\{s_{n,m}(t)\}$ are crucial in transmission policy design and we will further specify their characteristics later.

In general, the interference is caused by more than one MUs transmitting over the same channel. Suppose that the maximum number of allowed collisions in the $m$th channel is set as $N_m$, i.e.,

$$0 \leq \sum_{n=1}^{N} s_{n,m}(t) \leq N_m, \text{ for all } m = 1, 2, \ldots, M. \tag{3}$$

When $N_m \geq 1$, i.e., more than one MUs access the same channel and transmit simultaneously, their signals are mixed together, which may cause errors when the C-RAN decodes the information for each MU. Since here we do not focus on how C-RAN reconstructs information from the collided signals, we simply treat $N_m$ as a fixed system parameter. In some scenarios, only one MU is allowed to access one particular channel at each time slot, i.e., $N_m = 1$, in order to fully mitigate the interference, where many access schemes have been proposed under this setup [3, 9, 22].

From the MU point of view, we have

$$0 \leq \sum_{m=1}^{M} s_{n,m}(t) \leq A_n, \text{ for all } n = 1, 2, \ldots, N, \tag{4}$$

where $A_n$ is the number of channels that the $n$th MU could use, and is determined by hardware limitations and resource management policies. Since the total number of channels is $M$, it follows that $1 \leq A_n \leq M$. Thus, (3) and (4) together capture the characteristic of the proposed heterogeneous wireless access. Again, it is worth noticing that $\{s_{n,m}(t)\}$ could be determined in either a centralized or distributed manner.

Let $x_n(t)$ be the transmitted signal by MU $n$. Then, the received signal at the BSs (processed at C-RAN) over channel $m$ is given by

$$y_m(t) = \sum_{n=1}^{N} h_{n,m}(t) s_{n,m}(t) \sqrt{P_{n,m}(t)} x_n(t) + z_m(t), \tag{5}$$

where $h_{n,m}(t)$ is the channel coefficient treated as a constant during one time slot, $P_{n,m}(t)$ is the transmission power of the $n$th MU over channel $m$, and $z_m(t)$ is the circularly symmetric complex Gaussian (CSCG) noise with zero mean and variance $\sigma_m^2$. Note that the total transmission power at each MU is $\sum_{m=1}^{M} s_{n,m}(t) P_{n,m}(t) = P_n(t)$. For notation simplicity, we let $g_{n,m}(t) = \frac{|h_{n,m}(t)|^2}{\sigma_m^2}$ be the effective channel power gain, and call $\{g_{n,m}(t)\}$ as the channel state information (CSI). By applying

**Table 1** Key features for the system setup

| System features | Conditions |
|---|---|
| Time horizon | Finite ($T < \infty$) or infinite ($T \to \infty$) |
| Central controller | Available (centralized) or unavailable (distributed) |
| CSI | Non-causal or causal |
| ESI | Non-causal or causal |
| Channel accessibility | Single-channel or multi-channel |

the Shannon capacity formula over a multiple access channel [33], the sum rate of all $N$ MUs over the $m$th channel is given by $\log\left(1+\sum_{n=1}^{N} s_{n,m}(t)g_{n,m}(t)P_{n,m}(t)\right)$[2] for $1 \le m \le M$. Thus, the total instantaneous rate of the uplink is given by

$$\sum_{m=1}^{M} \log\left(1+\sum_{n=1}^{N} s_{n,m}(t)g_{n,m}(t)P_{n,m}(t)\right). \tag{6}$$

So far, we have presented the basic system parameters including: the environment variables, i.e., time horizon $T$, ESI and CSI; and the controlling variables, i.e., the access variables $\{s_{n,m}(t)\}$ and power allocations $\{P_{n,m}(t)\}$ for all $n$, $m$ and $t$. Note that these variables can be further specified according to the interest of system design. We list some further system specifications in Table 1, which will be discussed in details when we introduce transmission polices in Sects. 3 and 4.

## 2.2 Throughput Maximization

In this subsection, we present a general formulation framework for throughput maximization and introduce some useful techniques to find the solution structure of design variables $\{s_{n,m}(t)\}$ and $\{P_{n,m}(t)\}$.

For the considered system, the instantaneous rate of the uplink is given by (6). However, we do not directly consider the sum rate over the desired time horizon $T$ as the objective function. Instead, we introduce an auxiliary variable $\lambda$ as the objective function such that

$$\mathbf{E}\left[\sum_{t=1}^{T}\sum_{m=1}^{M} \log\left(1+\sum_{n=1}^{N} s_{n,m}(t)g_{n,m}(t)P_{n,m}(t)\right) - \lambda T\right] = 0, \tag{7}$$

[2]Note that the notation "log" could represent either the logarithm with base 2 or base $e$ according to the applications. In this chapter, we use this unified notation.

where $\mathbf{E}[\cdot]$ is the expectation operator over some dynamic statistics (if any). The initial conditions are $\{g_{n,m}(1)\}$ and $\{E_{n,m}(1), B_{n,m}(1)\}$. We employ $\lambda$ since condition (7) is able to provide us certain flexibility to define the utility of interest. If there is no specific regulations on $T$, an intuitive observation is that the variable $\lambda$ can represent the expected value of the average uplink throughput, i.e., condition (7) is equivalent to

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M}\log\left(1+\sum_{n=1}^{N}s_{n,m}(t)g_{n,m}(t)P_{n,m}(t)\right)\right]=\lambda. \tag{8}$$

In fact, as it will be shown later, $\lambda$ can also represent the average rate per transmission, like "rate-of-return," if we consider some particular dynamic optimization techniques.

Besides condition (7), the access variable $\{s_{n,m}(t)\}$ should satisfy constraints (3) and (4) for $1 \le t \le T$. Moreover, the total transmission power $P_n(t) = \sum_{m=1}^{M} s_{n,m}(t)P_{n,m}(t)$ is bounded and satisfies a new type of EH constraint, i.e.,

$$0 \le P_n(t) \le B_n(t), \text{ for all } n = 1, 2, \ldots, N, \tag{9}$$

where $B_n(t)$ is the battery energy level satisfying (1).

Thus, the throughput maximization problem can be formulated as

$$\begin{aligned} &\text{maximize} \quad \lambda \\ &\text{subject to} \quad (7), \\ &\qquad\qquad (3),\ (4) \text{ for } 1 \le t \le T, \\ &\qquad\qquad (9) \text{ for } 1 \le t \le T. \end{aligned} \tag{10}$$

Note that when $T \to \infty$ (or $T \to \infty$ almost surely when $T$ is random), all constraints in this framework still need to hold, and $\lambda$ in this case can represent the long term average throughput.

The solution to problem (10) is highly dependent on the system setups listed in Table 1. One such major condition is the causality of the system information (CSI or ESI). When both the CSI and ESI could be non-causally obtained, there is no randomness in problem (10) and the expectation operator could be dropped off. Then, problem (10) is solvable by applying some offline optimization tools, e.g., convex or mix-integer optimization.

When the knowledge about CSI or ESI is obtained from statistics, problem (10) could be solved by some dynamic optimization techniques, which usually require more definitions beyond the formulations described above. There are four main quantities in a dynamic optimization framework as follows:

1. **The space of system states**. In problem (10), the system state is represented by both CSI and ESI, i.e., $\{g_{n,m}(t)\}$ and $\{E_{n,m}(t), B_{n,m}(t)\}$. Thus, the space of the system states consists of all possible values of CSI and ESI.
2. **The space of actions**. This space is composed by all possible values of the design variables, i.e., the access control $\{s_{n,m}(t)\}$ and the transmission power levels $\{P_{n,m}(t)\}$. In each time slot $t$, the decision maker chooses a set of actions $\{s_{n,m}(t), P_{n,m}(t)\}_{1\leq n\leq N, 1\leq m\leq M}$ that must satisfy all constraints in problem (10). Note that in most existing literatures on EH-based wireless access design [19, 34], the spaces of system states and actions are assumed to be fixed over time.
3. **The reward**. For a given system state and action at the $t$th time slot, the corresponding instantaneous reward for problem (10) is instantaneous throughput (6).
4. **The transition of system states**. The transition of the system states is dependent on the system state and the actions at the current time slot. Particularly, the battery states $\{B_{n,m}(t)\}$ are influenced by both the actions and $\{E_{n,m}(t)\}$ according to the battery evolution in (1). The transition can be either deterministic or dynamic. In the dynamic case, the transition is often described by the state transition probability.

Now, solving problem (10) can be achieved by identifying the sequence of actions $\{s_{n,m}(t)\}$ and $\{P_{n,m}(t)\}$.

To derive the optimal solutions, popular dynamic frameworks have been applied in studying EH-based wireless communications, including the Markov decision process (MDP) [24, 34] (and the references therein), partially observable MDP (POMDP) [3, 12] (and the references therein), restless multi-arm bandit (RMAB) [3], and optimal stopping [20, 21]. The characteristics of these frameworks that reflect their applications are briefly discussed below.

**MDP and POMDP**: MDP is a controlled discrete-time stochastic process, where the decision maker can *perfectly know* the current state and the state transition probability of the system. Moreover, the system state at the next time slot is *conditionally independent* of all previous states and actions. The decision maker will choose an action from an action space according to the current reward, and this action will influence the evolution of the system state.

The POMDP is a generalization of an MDP, which permits certain *uncertainty* of the state information. Such uncertainty may be resulted from the indirect or partial observations over the system state. Thus, an additional observation model needs to be incorporated into the POMDP. For example, if the MU cannot perfectly obtain its CSI, a hypothesis testing process is commonly used to identify the channel status.

**RMAB and Optimal stopping**: RMAB makes decisions for multiple stochastic processes simultaneously, which fits the scenario of multiple MUs where the state of each MU is a stochastic process. The decision maker needs to deal with the tradeoff between allocating resources for a high current reward (*exploitation*) and acquiring new knowledge with a prospect of a better future reward (*exploration*). Moreover, the considered multiple stochastic processes keep the state evolution even if the decision maker does not make any actions, while in classic MAP, the processes are "frozen" if there are no actions. When the state transition of the system has Markov properties, RMAB is also an MDP (or POMDP if the state information cannot be obtained directly).

If the decision maker is required to decide when to stop the decision process (a binary decision), it becomes an optimal stopping problem. The main tradeoff is also the exploration vs. exploitation, while in optimal stopping, the exploitation also implies that the decision process is finished.

Despite the different formulations for the dynamic case, the solution under these frameworks, however, usually turns out to be nothing but dynamic programming. When the knowledge about the system (e.g., the reward or state transition probability) is unknown, the solution may turn to machine learning algorithms (e.g., reinforcement learning, Q-learning) [4, 28].

It should be pointed out that for either case of the causality over CSI/ESI, problem (10) has not been solved in general. Next, we only focus on several typical scenarios and discuss the state-of-art transmission polices. To highlight the flexibility of channel management, we introduce transmission policies in two main scenarios:

- Single-channel scenario, i.e., $M = 1$: This is a typical access model when the channel resources are limited, which is commonly used in ad hoc, sensor, and cellular network models.
- Multi-channel scenario, i.e., $M \geq 2$: In this scenario, there are more than one channels for MUs to access at each given time, while for data transmissions, an MU can use either only one channel, or more than one channels simultaneously. When the hardware of the MU supports multi-channel transmissions, possible multiple RATs, it can fulfil a transmission over multiple channels simultaneously.

Note that the channel accessibility sets limitations on channel access indicators $\{s_{n,m}(t)\}$, and thus constraints (3) and (4) need to be modified accordingly, which will be specified later.

## 3 Single-Channel Scenario

Consider a scenario where multiple MUs are sharing limited channel resources, and this can be modeled as the single-channel scenario, i.e., $M = 1$. For this case, transmission policies that solve problem (10) can be further classified according to their ability of mitigating the interference among different MUs:

1. Without collision: The access control is designed to remove the interference by guaranteeing that each MU solely uses the channel during its data transmission.
2. With collision: Note that the interference could be mitigated in different levels. If the receiver could handle interferences among MUs by adopting some advanced interference management techniques [2], the access control can allow multiple MUs to access one channel at the same time. In this case, the single-channel scenario degrades to the conventional multiple access channel.

### *3.1 Without Collision*

In this subsection, the access policy is required to manage the data transmissions of MUs one by one over time such that no transmission collisions happen. Hence, TDMA is one feasible scheduling scheme, where in each time slot the channel can only be assigned to one MU. CDMA with orthogonal codes could also support multiple MUs.

#### 3.1.1 Centralized Access and Power Control

When TDMA can be implemented in a centralized manner, it is optimal for the controller to pick the MU with the highest effective transmission rate according to its CSI and ESI in each time slot. That is, the selected MU should jointly have a good channel condition and sufficient energy in the battery.

The authors in [16] adopted the above idea and studied the case when all MUs have the same mean EH rate. The selection process in their proposed policy depends only on the channel condition. Specifically, the region of the channel power gain is divided into multiple interval units, based on some discrete-rate transmission schemes (e.g., M-QAM). Each channel state has an one-to-one mapping to a transmission rate, which determines the required transmission power. In order to meet the constraint of the average power consumption (i.e., the mean of EH rates), the length of interval unit must be designed accordingly. Then, at each time slot, the controller picks the MU with the highest ratio between the channel power gain and the length of the interval unit. The intuition is that the one with a higher ratio should attain a higher rate level. Note that if the channel statistics are identical, the length of the unit interval of the channel state will be the same, which is equivalent to pick the MU with the highest channel power gain. Due to the EH constraint, the selected MU

may fail to transmit at the desired rate. In this case, the selected MU can just keep silent. It is shown in [16] that by using such a policy, the average uplink throughput can achieve the similar performance as the system with conventional power supplies. In other words, the effect of EH constraints can approximately vanish.

The above centralized policy requires the knowledge of instantaneous CSI of all MUs, which may induce difficulties in implementations. Without the need of CSI, another TDMA-based transmission policy was investigated in [22], which lets the MUs access the channel and transmit one by one in a fixed order. Such a fixed TDMA scheme enjoys a lower computation complexity at the central controller compared to some aggressive selection algorithms that require the instantaneous status from each MU, like the one introduced in the previous paragraph. One transmission period is the total number of time slots, during which all MUs transmit for one time slot. As such, the total number of time slots for one transmission period equals $N$ since each MU only uses the channel once in each period. The channel access indicators $\{s_{n,1}(t)\}$ in our framework are preassigned by the controller. Suppose that the $n$th MU is assigned in the $t$th slot, and let $d_n(t)$ be the duration of the time slot $t$. Then, the total time of one transmission period is $\sum_{n=1}^{N}\sum_{t=1}^{N} d_n(t)$.

The controller manages the transmission time for each MU in order to maximize the expected sum rate under an additional constraint that the average transmission period cannot exceed a fixed number. The authors in [22] showed that the optimal transmission policy is the equal-power scheme, i.e., each MU should set its transmission power at the same level which is dependent on the energy information of all MUs. The optimal transmission time, on the other hand, is proportional to the mean EH rate such that a higher EH rate implies a longer transmission time. A simpler policy, equal-time scheme, was also studied in [22]. Unlike the equal-power scheme, the equal-time scheme simply divides one transmission period equally and gives all MUs the same transmission time. The power allocation uses the greedy algorithm. Comparing these two policies, the equal-power scheme outperforms the equal-time one when the battery has an infinite capacity. Whereas, they are similar when the battery has a finite capacity.

### 3.1.2 Distributed Access and Power Control

Centralized algorithms usually lead to a better throughput performance, while extensive computation complexity can be incurred at the central controller, especially in a large-scale system case. To relax the complexity, or even remove the necessity of the central controller, the investigation of distributed algorithms has drawn great attentions over the past years.

Commonly used distributed random access schemes, such as ALOHA and CSMA, are all applicable for problem (10). In this case, access variables $\{s_n(t)\}$

are determined locally[3] according to the chosen schemes. For example, if all $N$ MUs equally contend for the channel use with probability $1/N$ at each time slot, we obtain

$$s_n(t) = \begin{cases} 1, \text{ with probability (w.p.) } \frac{1}{N}; \\ 0, \text{ w.p. } 1 - \frac{1}{N}. \end{cases} \tag{11}$$

In this example, a successful channel contention happens when only one MU accesses the channel, i.e., $s_n(t) = 1$ and $s_i(t) = 0$ for $i \neq n$, where $1 \leq n, i \leq N$.

For the power control, an intuitive and feasible scheme is the best-effort delivery method, which is basically the same as the greedy algorithm such that whenever an MU occupies the channel, it will use up all available energy for data transmission without considering the channel condition and the battery status. This scheme is easy to implement, while the throughput performance may not be optimal.

An improved best-effort method, named distributed opportunistic framework, was designed in [21], which maximizes the long term average throughput of the network by jointly taking the advantages of the causal CSI and ESI. The proposed opportunistic framework operates in two stages: *channel probing* (CP) and *energy probing* (EP). In the first stage, all MUs keep contending for the channel use independently until there is only one occupying the channel. The successful MU obtains the CSI through some handshaking mechanism with the BS. This stage is called CP. After CP, the successful MU can possibly transmit over multiple time slots and then release the channel. The total number of time slots over which MU could transmit is called a transmission block. Moreover, the MU can chooses one of the following actions based on its local CSI and ESI right after CP:

(a) releases the channel (if CSI and ESI indicate that the transmission rate is lower than a given threshold) and let all links re-contend; or
(b) directly transmits until the end of the transmission block; or
(c) holds the channel, and starts the second stage EP.

It is worth noting that each MU keeps harvesting energy until it starts a transmission, and after each round of CP, only the successful MU makes a choice among three actions as listed above. When the successful MU decides not to take action (a) or (b) defined above, it starts the second stage EP, i.e., action (c), to obtain more energy. During this stage, the MU chooses to continue harvesting energy slot by slot, and then ends EP by action (a) or (b), i.e., either releasing the channel or transmitting over the rest of the transmission block.

Our proposed framework described in Sect. 2.2 can be used to optimize the two-stage probing strategy. Besides the constraints described in (10), we need to further specify $\{s_n(t)\}$ in the two-stage probing scenario. Note that $\{s_n(t)\}_{1 \leq n \leq N}$ are all random during the period of channel contentions. If some MU, say the $n$th MU, occupied the channel and transmitted over multiple time slots, $\{s_n(t)\}_{1 \leq n \leq N}$

[3]Here, we remove the channel index $m$ since we only consider the single-channel scenario.

would become determined during the period of the transmission, i.e., there must be $s_n(t) = 1$ and $s_i(t) = 0$ for $i \neq n$ during this time. Here, we let $Q$ denote the number of time slots of the transmission block.

The transmission power is determined by the remaining time within a transmission block (since the MU may perform EP) and the energy level when the MU starts a transmission. Here, we adopt a unform power allocation scheme. For example, if the number of remaining time slots was $\widehat{Q}$ and the transmission started at $t_0$, the transmission power would be $P_n(t) = \frac{B_n(t_0)}{\widehat{Q}}$ and keep the same during $\widehat{Q}$ slots.

We let $L_C$ and $L_E$ be the stopping rules of CP and EP, respectively. Stopping rule $L_C$ indicates the change point of $\{s_n(t)\}$ becoming deterministic, and stopping rule $L_E$ regulates the transmission power within the transmission block. Since each CP may take a random number of slots, we let $K_i$ be the number of slots for the $i$th CP. Then, the total number of slots to fulfil one transmission is given by

$$S(L_C, L_E) = \sum_{i=1}^{L_C - 1} L_{E,i} + \sum_{i=1}^{L_C} K_i + Q. \tag{12}$$

If we keep using the stopping rules $L_C$ and $L_E$ for $J$ times, the time horizon is given as

$$T = \sum_{j=1}^{J} S_j(L_C, L_E). \tag{13}$$

The infinite time horizon can be realized by letting $J \to \infty$, which leads to $T \to \infty$ a.s. In Fig. 3, we illustrate one realization of a successful transmission, which costs $L_C$ rounds of CP, and also $L_E$ slots of EP in the last CP.

Then, we can find the optimal stopping rules $L_C^*$ and $L_E^*$ by plugging (13) into the framework given in (10). More precisely, condition (7) becomes

$$\mathbf{E}\left[\sum_{j=1}^{J}\sum_{t=1}^{S_j(L_C,L_E)} \log\left(1 + s_n(t) g_n(t) P_n(t)\right) - \lambda \sum_{j=1}^{J} S_j(L_C, L_E)\right] = 0. \tag{14}$$

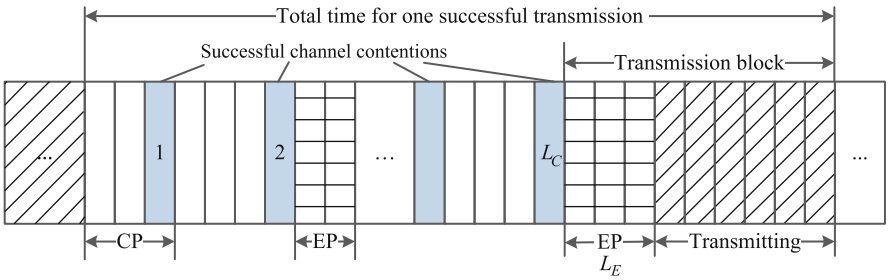


**Fig. 3** One realization of a successful transmission by applying the distributed opportunistic framework [21]

By letting $J \to \infty$, we obtain

$$\mathbf{E}\left[\sum_{t=1}^{S(L_C,L_E)} \log\left(1 + s_n(t)g_n(t)P_n(t)\right) - \lambda S(L_C, L_E)\right] = 0, \tag{15}$$

which is equivalent to

$$\lambda = \frac{\mathbf{E}\left[\sum_{t=1}^{S(L_C,L_E)} \log\left(1 + s_n(t)g_n(t)P_n(t)\right)\right]}{\mathbf{E}\left[S(L_C, L_E)\right]}. \tag{16}$$

Here, $\lambda$ can also be viewed as the "rate-of-return" in the sense that Eq. (16) describes the average rate per transmission.

According to the optimal stopping rules $L_C^*$ and $L_E^*$ derived in [21], the successful MU will transmit only when the CSI and ESI jointly indicate a higher transmission rate than a given threshold, which can be found via a simple one dimension searching algorithm. However, such an opportunistic framework does not guarantee the fairness among MUs. When applying this scheme, the transmission will be more likely to be fulfilled by those MUs with the better channel conditions and the higher EH rates.

Setting the performance of the best-effort delivery as the benchmark. In Fig. 4, it is shown how the throughput gains change as the number of MUs increases under the condition that the battery has a finite capacity. It is observed that the gain from EP decreases when the number of MUs increases. This is due to the fact that the more

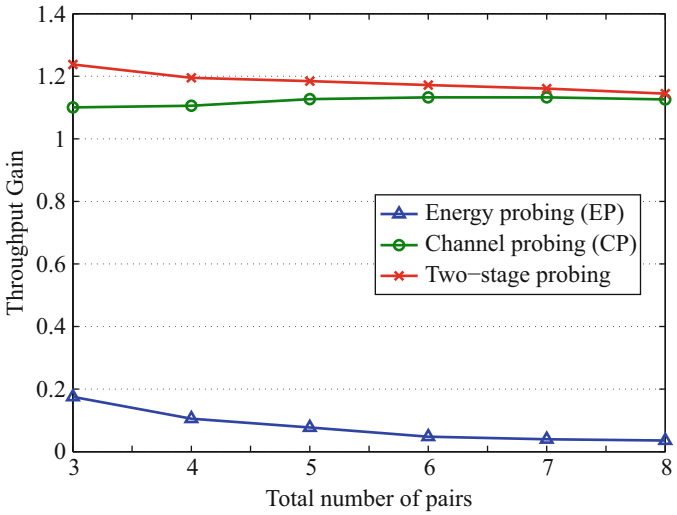


**Fig. 4** The throughput gain over the best-effort delivery method

MUs are present, the higher probability that CP will last a longer time. Therefore, more MUs would maintain a high energy level and EP is rarely triggered after CP. It is also observed that the throughput gain from CP increases with the increase in the number of MUs since CP could utilize the multi-user diversity of both channel gains and energy levels. The CP can offer a much higher gain than EP, and thus the gain from the two-stage probing is comparable to CP as the size of the network increases.

### *3.2 With Collision*

In this case, multiple MUs can transmit over the same channel simultaneously, which is just the *multiple access channel* under EH constraints [17, 37, 39]. Existing literatures have studied the sum rate maximization problem in some "ideal" scenarios. The algorithms introduced next are all centralized in the sense that controller needs to regulate the sum transmission power in order to achieve the optimality.

A typical case is Gaussian multiple access scenario with two MUs and the ESI is non-causally known [39]. One key observation is that if the goal is to maximize the sum rate only, the optimal sum power has a stair-case structure, which is the same as the single MU case (see Sect. III in [34]). Furthermore, if the fairness of MUs is accounted, the sequence of the optimal sum power allocation needs to be split into two subsequences in order to achieve the largest throughput region of two MUs. The optimal power allocation for each MU can be computed by applying the generalized iterative water-filling algorithm [39]. For the fading channel with non-causal CSI, authors in [37] proposed an efficient power allocation scheme by considering a finite battery capacity and an upper limit on the instant transmission power.

When ESI is causally known at MUs, an online algorithm is proposed for Gaussian multiple access in [17] with $T \to \infty$. One important observation is that the optimal power allocation should make the energy level of the battery stationary for a given EH rate process. Then, the optimal power allocation can be expressed by the stationary distribution of the battery and the distribution of the EH rate. For the infinite battery capacity case, the solution is the "constant" power strategy. Specifically, the transmission power level is equal to the mean of the EH rates. If the current energy is not sufficient enough, the MU just uses up the available energy. This is similar to the algorithm mentioned in [29] for point-to-point channels. For the finite battery capacity, it is shown that the optimal power allocation does not have a close-form expression and it can be solved numerically.

## 4 Multi-Channel Scenario

In this section, we consider the scenario when there are multiple channels for MUs to access, i.e., $M \geq 2$. Moreover, there is an additional condition in this case: whether or not the MU can transmit over more than one channels simultaneously.

- Transmit over one channel, i.e., $A_n = 1$ in constraint (4): This condition implies that although there are more than one channels to access, the MU is restricted to use at most one channel at each time. Such a restriction could equalize the opportunity of channel usages for different MUs to some extent, and also simplifies the design of access algorithms.
- Transmit over multiple channels, i.e., $A_n \geq 2$: Under this condition, the MU should first decide or be assigned with the available channels, and then apply the water-filling algorithm for the power allocation according to the current CSI and ESI. Obviously, the MU throughput could be potentially enhanced compared to the case with $A_n = 1$; but the hardware complexity and the total power consumption would be increased as well.

Next, we will introduce the transmission polices for problem (10) for two special scenarios according to the number of MUs: $N \geq 2$ and $N = 1$. In each subsection, we will further consider the realization of problem (10) for cases with $A_n = 1$ and $A_n \geq 2$.

### *4.1 Network Throughput Maximization with $N \geq 2$*

In this subsection, we mainly introduce the centralized transmission schemes. We also modify constraint (3) by letting $N_m = 1$ for all $m = 1, 2, \ldots, M$, i.e., each channel is only allowed to be accessed by at most one MU in each time slot. With this modified constraint, the designed access policy can avoid the collisions of data transmissions and improve the system reliability. Similar to the single-channel scenario, TDMA is one feasible scheduling scheme. The main difference is that for each time slot, a group of MUs need to be scheduled to multiple channels. Obviously, when the number of MUs exceeds the number of available channels, some MUs cannot be scheduled for a given time slot.

First, we consider the case $A_n = 1$ for all $1 \leq n \leq N$, i.e., the MU can only transmit over one channel. The average throughput of the network can be maximized by wisely choosing the set of scheduled MUs and controlling the transmission power according to the CSI and ESI of all MUs. However, the controller can hardly obtain the information of all MUs. To facilitate this, all MUs need to send their own ESI to the BSs, and also the BSs need to feedback the CSI to each MU. This process will introduce extra scheduling issues and may cost a long period when the number of MUs is large. As such, a reasonable condition is that the CSI and ESI of all MUs is causally and partially known by the controller. That is, the controller only obtains

the new CSI and ESI of the MUs who are scheduled in the current slot. Then, the controller can use all available CSI and ESI (may include the history) to optimize the current uplink throughput or to predict the future status of MUs.

Under this condition, the centralized schemes for problem (10) can be derived by applying some dynamic frameworks (e.g., MDP, PMDP, or RMAB). Here, we only introduce two centralized schemes: uniformizing random ordered policy (UROP) and myopic policy. Note that they may not be throughput-optimal since their throughput formulations are not exactly the same as (6). However, both are feasible and do not involve extensive computations.

- *UROP*: The authors in [9] studied the access policy for the packet level under the assumption of a constant energy consumption for delivering a packet. Since the battery status of each MU is random and is unknown at the receiver side, the problem here is how to schedule the accesses of MUs such that the expected average throughput is maximized. The solution proposed in [9] is a UROP: briefly, one round of scheduling starts with randomly ordering all $N$ MUs. Then, the controller picks the first $M$ MUs in the order list and let them access $M$ parallel channels, where each user can only access one channel for each time slot. In the next time slot, the MUs who did not transmit in the previous time slot will be replaced by those who have not been scheduled. The controller selects the unscheduled MUs according the order list. After such a replacement, the newly scheduled $M$ MUs will transmit for one time slot. This procedure will keep running until all MUs transmit at least once. Then, the controller will start a new round of scheduling by creating a new order list. It is shown in [9] that from the network point of view, when the average packet rate is smaller than the average delivering rate (i.e., the total number of channel usages divided by $N$), UROP leads to minimum waste energy.
- *Myopic Policy*: Given that the current states of MUs are unknown at the controller side in each time slot, the authors in [3] proposed an "myopic" policy that only maximizes the expected instantaneous sum rate. The power allocation in this policy follows the greedy algorithm that uses all available energy for data transmission. The scheduling in this policy is similar to UROP. The difference is that the order list is not created randomly. Instead, the controller orders MUs according to the length of the idle time after the previous transmission of each MU. Clearly, the longer the idle time, the more energy that can be possibly harvested. In each time slot, the first $M$ MUs in the order list can access channels. If the scheduled MU transmits, it will be added to the end of the order list in the next time slot since it uses up all energy. The MUs who have not transmitted (either scheduled or not) will go up in the order list. This myopic policy is proved in [3] to be throughput-optimal over the considered time horizon.

Next, we consider $A_n \geq 2$. In this case, the centralized scheduling is a mixed integer programming problem, which is non-convex and hard to solve. However, if we consider the centralized instantaneous scheduling (i.e., optimizing the throughput only in a given time slot), maximizing (6) is similar to the throughput maximization problem for a centralized uplink OFDM wireless system,

which has already been extensively investigated [18, 38]. It is worth noting that in OFDM systems, another key problem is the frequency bandwidth allocation, while in our considered system model, the frequency band is fixed and we only focus on the problem of channel assignment and power allocation. Thus, those algorithms concerning channel assignments and power allocations proposed for OFDM systems can be directly modified and applied in maximizing the throughput in our considered system.

The transmission polices designed for the case of $N_m = 1$ can be extended to the general case of $N_m \geq 2$ by reassigning the channels to the unscheduled MUs. Then, there are more MUs scheduled in each time slot. However, the general case also means that the access control does not avoid the interference among different MUs, which exerts more difficulties on the design of the optimal power control.

## *4.2 User Throughput Maximization with $N = 1$*

In this subsection, we focus on the scenario when the channel resources are relatively sufficient, e.g., late at night in the business area or working hours in the resident area. Such a scenario can be modeled as the single-MU scenario $N = 1$ with multiple channels. For the case $A_n = 1$, an intuitive access strategy for the MU is to pick the channel with the highest channel power gain among all the available channels. To obtain such information, the MU needs to probe channels with the cooperation of the BS. The access schemes implemented at MUs have been studied [15, 31], especially in cognitive radio networks. In general, the MU can perform one of the three main access mechanisms:

1. Parallel channel probing: Parallel probing can be realized by broadcasting a *query* message and receiving *reply* messages that contains the CSI of the probed channels. Then, the MU can access any one and start the transmission.
2. Sequential channel probing with recall: The MU probes channels one by one according to a certain order. When the MU stops probing, it can access any channel that has been probed.
3. Sequential channel probing without recall: The MU probes channels one by one according to a certain order. When the MU stops probing, it can only access the last probed channel.

Although the first two mechanisms may provide more options when the MU chooses which channel to access, both have apparent drawbacks that prevent from being implemented. Parallel channel probing requires a compatible hardware of broadcasting and also causes relatively high instantaneous power consumption. Sequential channel probing with recall needs the MU to memorize the state of the probed channels. However, the CSI may change when the MU accesses the channel once again.

On the contrary, sequential channel probing without recall does not carry the above shortcomings and enjoys much more attentions than the other two. Such

a probing mechanism has been extensively studied for cognitive radio networks [6, 31, 41]. However, for the EH-based heterogeneous access problem, the specific transmission polices accounting the energy information have not been well studied. As we discussed in Sect. 2.2, two dynamic frameworks, POMDP and optimal stopping, are ideal tools to formulate the sequential decision making process and can be used to derive the optimal transmission policies.

For the case $A_n \geq 2$, we apply a save-then-transmit protocol as the basic structure of the transmission policy. As its name implies, this protocol lets the MU accumulate energy in the first step, and then lets the MU use all available energy for data transmission. Such a protocol enjoys two major characteristics:

1. It captures the fundamental operation of EH-based systems that the energy must be accumulated in the first place, even in an energy full-duplex fashion. However, the time required for harvesting energy will induce some penalties.
2. It is simple, tractable, and applicable for any EH-based systems.

It is easy to see that the main design issue of the save-then-transmit protocol is to determine when to start data transmission such that the average throughput is maximized.[4] In the following, we first investigate the case when this protocol is used over a finite time horizon; and then, we extend the protocol to the infinite time horizon case.

### 4.2.1 Finite Time Horizon

The save-then-transmit protocol applied over a finite time horizon is depicted in Fig. 5, where the time horizon is normalized as one time slot. The MU spends a certain portion of time on harvesting energy, and then starts channel probing for CSI. Such a probing process is similar to sensing with recall [31]. After channel probing, the MU transmits over the selected channel(s).

Clearly, spending time on harvesting energy and channel probing will increase the instantaneous rate, while the effective transmission time will be decreased. To deal with this tradeoff, two main design variables need to be optimized for this protocol: the saving ratio (indicating the portion within one time slot for harvesting energy) and the number of channels to probe. Note that the first design variable is a continuous number, while the second one is an integer. Thus, maximizing the average throughput is also a mixed integer programming problem. An algorithm proposed in [40] can be used to solve this problem, which randomly selects many pairs of feasible solutions and chooses the best one.

[4] Note that this protocol can also be optimized to minimize the outage probability for a given time slot [23].

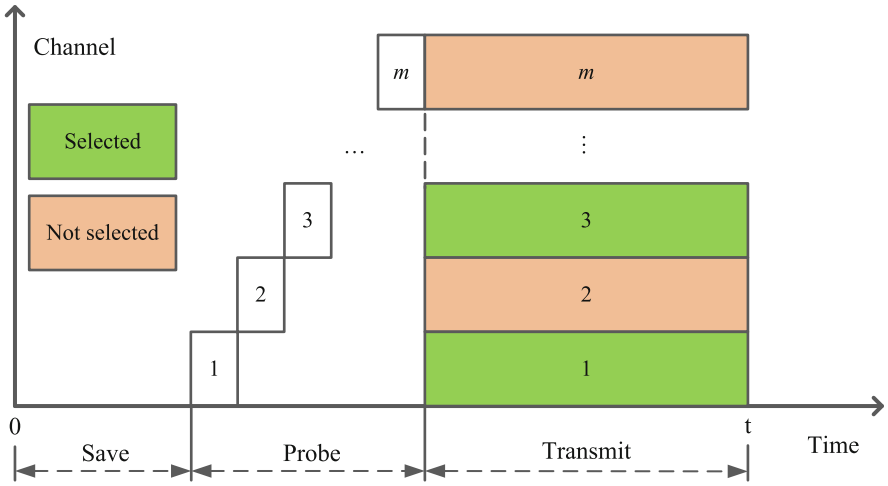


**Fig. 5** A realization of the save-then-transmit protocol within one time slot [40]

### 4.2.2 Infinite Time Horizon

Next, we extend the protocol in Fig. 5 to the infinite time horizon case (e.g., multiple time slots). In this case, the MU can save the energy for more than one time slot and then start the transmission. Besides the two design variables (saving ratio and the number of channels to probe) mentioned above, another important variable is the stopping rule: when to stop saving and to start the transmission. The stopping rule determines the action after the MU probes channels. The MU has the choice to transmit or skip the transmission in the current time slot according to its current CSI and ESI. The skipping is preferred when the CSI and ESI jointly indicate a low transmission rate.

The stopping rule can be reflected in the indicators $\{s_m(t)\}$, where the index $n$ for MUs is omitted since $N = 1$. Particularly, if we let $L$ be the stopping rule such that the MU transmits only at the time slot $L$, there must be $\{s_m(t) = 0\}_{1 \le m \le M}$ for $1 \le t < L$. At time slot $L$, $\{s_m(t)\}_{1 \le m \le M}$ are determined by the channel selection scheme. Thus, by applying the stopping rule $L$, the total transmission power $P(L)$ satisfies

$$0 \le P(L) \le B(L) = \min \left\{ \sum_{t=1}^{L-1} E(t), B_{max} \right\}, \tag{17}$$

If this stopping rule is used for $J$ times, condition (7) can be written as

$$\mathbf{E}\left[ \sum_{j=1}^{J} \sum_{t=1}^{L_j} \sum_{m=1}^{M} \log\left(1 + s_m(t) g_m(t) P_m(t)\right) - \lambda \sum_{j=1}^{J} L_j \right] = 0, \tag{18}$$

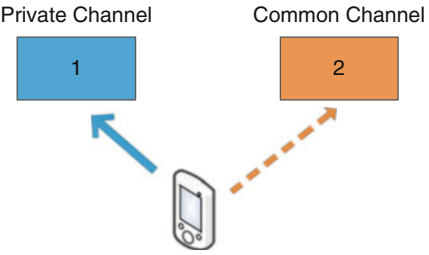


**Fig. 6** The hybrid access model

which leads to

$$\lambda = \frac{\mathbf{E}\left[\sum_{m=1}^{M} s_m(L) \log\left(1 + g_m(L)P_m(L)\right)\right]}{\mathbf{E}[L]}, \tag{19}$$

by letting $J \to \infty$, where $\sum_{m=1}^{M} s_m(L)P_m(L) = P(L)$. Then, we can use framework (10) to find the optimal stopping rule. Such a save-then-transmit protocol with optimal stopping rule can be named as a opportunistic transmission policy since it explores the state information over both time and space.

It is worth noticing that the term "opportunistic spectrum access" [41] in cognitive radio networks is quite different from the opportunistic transmission discussed above. The opportunistic spectrum access describes the spectrum access of the secondary users in a hierarchy network model. In such a network, the secondary user detects the availability of the licensed channel and then decides whether transmit or not. The term "opportunistic" is used in the sense that the secondary user makes a decision that can only achieve certain optimality in probability. Although the term "opportunistic" may represent different philosophies, some techniques proposed for spectrum access in EH-based cognitive radio networks [26] are also applicable in EH-based heterogeneous access.

To better illustrate the save-then-transmit protocol with optimal stopping rule, we set the saving ratio as zero and assume that the time for the MU to learn the CSI is negligible compared to the time slot length. That is, we only focus on the design of the stopping rule $L$. Particularly, we consider an interesting case of a hybrid access scenario as shown in Fig. 6, where the MU has both deterministic access via one private channel, and dynamic access with certain probability via one common channel. Such a common channel usually can be offered by a local BS, e.g., a WiFi router. The MU can only access the common channel with certain probability given that the common channel may be close temporarily. In this hybrid access scenario, the total number of parallel channels is $M = 2$. Following the notations in Sect. 2, we let $\{g_1(t)\}_{t\geq 1}$ and $\{g_2(t)\}_{t\geq 1}$ be the sequence of channel power gains for the private channel and the common channel, respectively. Then, the instantaneous rate is written as

$$R(t) = \sum_{m=1}^{2} \log\left(1 + s_m(t) g_m(t) P_m(t)\right).$$

We assume that $\{g_2(t)\}_{t\geq 1}$ are independent and identically distributed (i.i.d.) over time, given that the distance from the MU to the local BS is short and thus the common channel often experiences fast channel fading. Note that the private channel can always be accessed, i.e., we have $s_1(t) = 1$ for all $t$. The common channel can only be accessed randomly, and the MU can transmit via the common channel only when the MU secures it. Let $p_s$ denote the probability that the MU can secure the common channel, termed *securing probability*. Note that the securing probability may not necessarily equal the access probability, but their relations can be derived according to the system setups. Here, for notation simplicity, we assume that whenever the MU accesses the common channel, it can also secure the channel. Then, $s_2(t)$ is a random variable such that

$$s_2(t) = \begin{cases} 1, \text{ w.p. } p_s; \\ 0, \text{ w.p. } 1 - p_s. \end{cases}$$

If we let $L$ be the stopping rule, $\lambda$ in (19) can be rewritten as

$$\lambda = \frac{\mathbf{E}\left[R(L)\right]}{\mathbf{E}[L]}. \tag{20}$$

Thus, the design of the save-then-transmit protocol is now to find the optimal stopping rule $L^*$ that maximizes (20).

When $\{g_1(t), E(t)\}_{t\geq 1}$ are modeled as Markovian processes, the optimal stopping rule $L^*$ was proved to exist in [20] and have a state-dependent threshold-based structure under both finite and infinite battery capacity assumptions. The optimal throughput is proved to be strictly increasing over the securing probability.

When $\{g_1(t), E(t)\}_{t\geq 1}$ are modeled as i.i.d. processes, respectively, the corresponding optimal stopping rule is proved to be a pure-threshold policy, i.e., the threshold does not change over time, which could be found via one-dimension search. We plot the variation of the average throughput as the threshold changes in Fig. 7, and observe that the average throughput could be optimized by adjusting the threshold.

In Fig. 8, we compare the throughput performance attained by the opportunistic transmission policy (with the optimal stopping rule $L^*$) and the best-effort delivery method. Note that the optimal power allocation with traditional power supply was investigated and shown to outperform the opportunistic policy [20]. Such observation may be predictable given that the EH constraint sets a hard cap on the instant transmission power, while the conventional average power constraint does not.

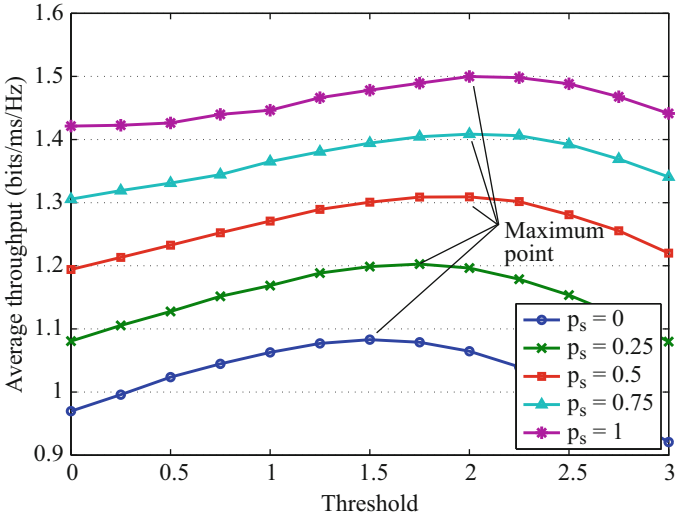


**Fig. 7** The average throughput over the threshold

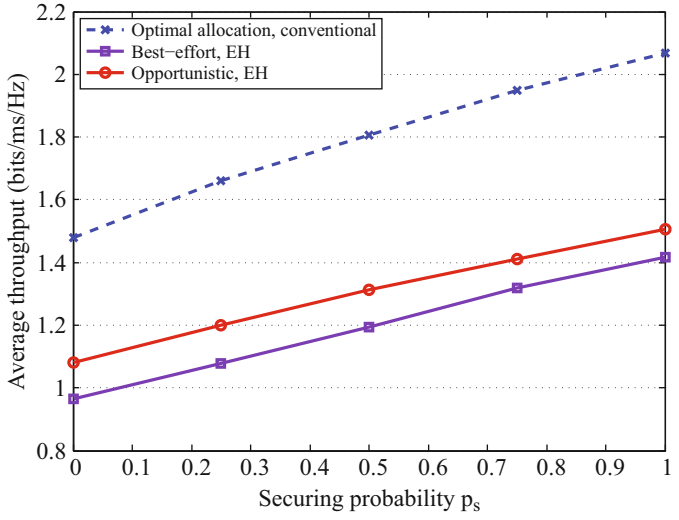


**Fig. 8** The throughput performance over different transmission policies

So far, we have discussed the multiple-channel scenario for both $N \geq 2$ and $N = 1$ cases. Again, we want to point out that accessing multiple channels at the same time may be too "luxury" for MUs. Some major issues include:

- Complexity at MU: The hardware at the MU should support multi-channel access, e.g., adopting multiple RF front ends. Meanwhile, power allocations over

multiple parallel channels may result in a higher power consumption on circuit overhead, which exerts a heavy burden on EH power supply.
- Interference: When the MU transmits via multiple channels, it means that other MUs cannot use these channels or may cause interference if they access. From the network operator point of view, allowing one MU to access multiple channels at the same time may decrease the whole benefits of the system.
- Needs of BS coordinations: The received signals from the same MU over multiple channels should be jointly processed. This requires the C-RAN to perfectly coordinate multiple BSs in a real-time fashion.

## 5 Summary and Discussion

EH-based power supply can drive wireless systems in a green fashion, while it also introduces both EH constraints and energy unreliability. These drawbacks should be mitigated as much as possible in different layers of the system design. In this chapter, we considered the EH-based heterogeneous wireless access, and proposed a throughput maximization framework across both physical and medium access layers. Transmission policies, including access and transmission power controls, were discussed in single and multi-channel scenarios.

EH-based heterogeneous wireless access is still under development, and many issues remain to be resolved. As the end of this chapter, some potential directions for future works are presented below.

**Learning Approach** Most of current studies adopt the dynamic programming technique to find the optimal transmission policies under the assumptions that the rewards and state transition probabilities of the system are known. However, given that the number of MUs keeps increasing, dynamic programming based solutions are no longer efficient when the state space of CSI/ESI of MUs is large and the statistics of CSI/ESI may not be fully accessible. Thus, it is worth to investigate learning based approaches, e.g., reinforcement learning, which are more feasible in the large-scale scenario since the learning approach does not require the statistics of the system and is extendable to deal with the large environments (see Chap. 3 in [5]).

**QoS vs. QoE** In general, the throughput-optimal transmission policies may not satisfy the quality of service (QoS) and the quality of experience (QoE) requirement at the same time. Sometimes there may exist a tradeoff between QoS and QoE. For example, too high QoS may decrease QoE since high QoS would consume more resources on both the user (e.g., energy stored in the battery) and operator (e.g., network resources, BS power) sides [1].

**Effect of EH-Based BSs** In this chapter, we only consider the heterogeneous access with EH-based MUs. Furthermore, BSs can also be driven or partially driven by EH-based power supplies to lower the energy cost on fuels and carbon-oxygen

emission. The employment of EH at BSs has drawn a great attention in recent years [10, 14, 42]. Particularly, the availability of EH-based BSs [8] has been analyzed. By accounting EH constraints at both MU and BS sides, the design of the channel access and power control becomes much more challenging.

**MU-to-MU Communications** The proposed heterogeneous wireless access model in this chapter captures the characteristic of the "MU-to-BS" communications, which belong to the medium-to-long range communications. As the short range communication being developed rapidly in recent years, 5G systems are also expected to incorporate device-to-device (D2D) communications where more and more EH-based devices could be involved. Thus, the wireless access control will mix multiple layers, and more tractable models need to be developed.

## References

1. P.K. Agyapong, M. Iwamura, D. Staehle et al., Design considerations for a 5G network architecture. IEEE. Commun. Mag. **52**(11), 65–75 (2014)
2. J.G. Andrews, S. Buzzi, W. Choi et al., What will 5G be? IEEE J. Sel. Areas Commun. **32**(6), 1065–1082 (2014)
3. P. Blasco, D. Gündüz, Multi-access communications with energy harvesting: a multi-armed bandit model and the optimality of the myopic policy. IEEE J. Sel. Areas Commun. **33**(3), 585–597 (2015)
4. P. Blasco, D. Gündüz, M. Dohler, A learning theoretic approach to energy harvesting communication system optimization. IEEE Trans. Wirel. Commun. **12**(4), 1872–1882 (2013)
5. L. Busoniu, R. Babuska, B. De Schutter et al., *Reinforcement Learning and Dynamic Programming Using Function Approximators*, vol. 39 (CRC press, Boca Raton, 2010)
6. Y. Chen, Q. Zhao, A. Swami, Distributed spectrum sensing and access in cognitive radio networks with energy constraint. IEEE Trans. Signal Process. **57**(2), 783–797 (2009)
7. China mobile, C-RAN the road towards green RAN. White paper, ver. 2.5 (2011)
8. H.S. Dhillon, Y. Li, P. Nuggehalli et al., Fundamentals of heterogeneous cellular networks with energy harvesting. IEEE Trans. Wirel. Commun. **13**(5), 2782–2797 (2014)
9. O.M. Gul, E. Uysal-Biyikoglu, A randomized scheduling algorithm for energy harvesting wireless sensor networks achieving nearly 100% throughput, in *IEEE WCNC 14: Topics in Mobile and Wireless Networks. IEEE Wireless Communications and Networking Conference*, Istanbul (2014)
10. H.A.H. Hassan, L. Nuaymi, A. Pelov, Renewable energy in cellular networks: a survey, in *IEEE OnlineGreenComm 2013: Topics in Green Wireless Networks. IEEE Online Conference on Green Communications* (2013)
11. A. He, A. Amanna, T. Tsou et al., Green communications: a call for power efficient wireless systems. J. Commun. **6**(4), 340–351 (2011)
12. X. Huang, T. Han, N. Ansari, On green energy powered cognitive radio networks. IEEE Commun. Surv. Tutorials **17**(2), 827–842 (2015)
13. International Telecommunication Union, The key 2005–2015 ICT data for the world. http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx. Accessed 26 May 2015
14. M. Ismail, W.H. Zhuang, E. Serpedin et al., A survey on green mobile networking: from the perspectives of network operators and mobile users. IEEE Commun. Surv. Tutorials. **17**(3), 1535–1556 (2015)
15. Z. Ji, Y. Yang, J. Zhou et al., Exploiting medium access diversity in rate adaptive wireless LANs, in *MobiCom 04. The 10th Annual International Conference on Mobile Computing and Networking*, Philadelphia (2004)

16. P.S. Khairnar, N.B. Mehta, Discrete-rate adaptation and selection in energy harvesting wireless systems. IEEE Trans. Wirel. Commun. **14**(1), 219–229 (2015)
17. M.B. Khuzani, P. Mitran, On online energy harvesting in multiple access communication systems. IEEE Trans. Inf. Theory **60**(3), 1883–1898 (2014)
18. K. Kim, Y. Han, S.L. Kim, Joint subcarrier and power allocation in uplink OFDMA systems. IEEE Commun. Lett. **9**(6), 526–528 (2005)
19. M.L. Ku, W. Li, Y. Chen et al., Advances in energy harvesting communications: past, present, and future challenges. IEEE Commun. Surv. Tutorials. **18**(2), 1384–1412 (2015)
20. H. Li, C. Huang, F.E. Alsaadi et al., Opportunistic multi-channel access in heterogeneous 5G network with renewable energy supplies (2015). arXiv:1412.4458. Accessed 21 May 2015
21. H. Li, C. Huang, P. Zhang et al., Distributed opportunistic scheduling for energy harvesting based wireless networks: a two-stage probing approach. IEEE/ACM Trans. Netw. **24**(3), 1618–1631 (2015)
22. J. Liu, H. Dai, W. Chen, On throughput maximization of time division multiple access with energy harvesting users. IEEE Trans. Veh. Technol. **65**(4), 2457–2470 (2015)
23. S. Luo, R. Zhang, T.J. Lim, Optimal save-then-transmit protocol for energy harvesting wireless transmitters. IEEE Trans. Wirel. Commun. **12**(3), 1196–1207 (2013)
24. N. Michelusi, M. Zorzi, Optimal adaptive random multiaccess in energy harvesting wireless sensor networks. IEEE Trans. Commun. **63**(4), 1355–1372 (2015)
25. R. Moghe, Y. Yang, F. Lambert et al., A scoping study of electric and magnetic field energy harvesting for wireless sensor networks in power system applications, in *ECCE. IEEE Energy Conversion Congress and Exposition*, San Jose (2009)
26. S. Park, D. Hong, Achievable throughput of energy harvesting cognitive radio networks. IEEE Trans. Wirel. Commun. **13**(2), 1010–1022 (2014)
27. A. Pasternack, The environmental costs (and benefits) of our cell phones (2009). http://www.treehugger.com/clean-technology/the-environmental-costs-and-benefits-of-our-cell-phones.html. Accessed 13 May 2015
28. K.J. Prabuchandran, S.K. Meena, S. Bhatnagar, Q-learning based energy management policies for a single sensor node with finite buffer. IEEE Wirel. Commun. Lett. **2**(1), 82–85 (2013)
29. R. Rajesh, V. Sharma, P. Viswanath, Capacity of Gaussian channels with energy harvesting and processing cost. IEEE Trans. Inf. Theory **60**(5), 2563–2575 (2014)
30. V. Sharma, U. Mukherji, V. Joseph, Optimal energy management policies for energy harvesting sensor nodes. IEEE Trans. Wirel. Commun. **9**(4), 1326–1336 (2010)
31. T. Shu, M. Krunz, Sequential opportunistic spectrum access with imperfect channel sensing. Ad Hoc Netw. **11**(3), 778–797 (2012)
32. S. Sudevalayam, P. Kulkarni, Energy harvesting sensor nodes: survey and implications. IEEE Commun. Surv. Tutorials **13**(3), 443–461 (2011)
33. D.N.C. Tse, S. Hanly, Multiple-access fading channels: Part I: polymatroid structure, optimal resource allocation, and throughput capacities. IEEE Trans. Inf. Theory **44**(7), 2796–2815 (1998)
34. S. Ulukus, A. Yener, E. Erkip et al., Energy harvesting wireless communications: a review of recent advances. IEEE J. Sel. Areas Commun. **33**(3), 360–381 (2015)
35. U.S. Energy Information Administration, How much coal, natural gas, or petroleum is used to generate a kilowatthour of electricity? (2015). http://www.eia.gov/tools/faqs/. Accessed 12 June 2015
36. L. Wang, F. Tian, T. Svensson et al., Exploiting full duplex for device-to-device communications in heterogeneous networks. IEEE Commun. Mag. **53**(5), 146–152 (2015)
37. Z. Wang, V. Aggarwal, X. Wang, Iterative dynamic water-filling for fading multiple-access channels with energy harvesting. IEEE J. Sel. Areas Commun. **33**(3), 382–395 (2015)
38. E. Yaacoub, Z. Dawy, A survey on uplink resource allocation in OFDMA wireless networks. IEEE Commun. Surv. Tutorials **14**(2), 322–337 (2012)
39. J. Yang, S. Ulukus, Optimal packet scheduling in multiple access channel with energy harvesting transmitters. J. Commun. Netw. **14**(2), 140–149 (2012)

40. S. Yin, E. Zhang, L. Yin et al., Optimal saving-sensing-transmitting structure in self-powered cognitive radio systems with wireless energy harvesting. In IEEE ICC 2013: topics in cognitive radio and networks symposium, in *IEEE International Conference on Communications*, Budapest (2013)
41. Q. Zhao, B.M. Sadler, A survey of dynamic spectrum access. IEEE Signal Process. Mag. **24**(3), 79–89 (2007)
42. S. Zhou, J. Gong, Z. Zhou et al., GreenDelivery: proactive content caching and push with energy-harvesting-based small cells. IEEE Commun. Mag. **53**(4), 142–149 (2015)

# Resource Management in Sustainable Green HetNets with Renewable Energy Sources

**Ran Zhang, Miao Wang, Lin X. Cai, Yu Cheng, Xuemin (Sherman) Shen, and Liang-Liang Xie**

**Abstract** Green energy has become a promising alternative energy source for powering wireless cellular networks by effectively reducing the network operational expenditure (OPEX) and carbon footprints. However, green energy sources such as solar and wind are harvested from the environment and their availability and capacity are by nature unstable, which poses great challenges to achieve sustainable network operation. In this chapter, we study the energy sustainable performance of a green HetNet where the small cell base stations (SBSs) are powered by green energy sources. Specifically, we first develop an analytical framework to study the energy sustainability of each SBS. The energy buffer at each SBS is modeled as a G/G/1 queue with arbitrary patterns of energy charging and discharging. We apply the diffusion approximation to analyze the transient evolution of the energy buffer, and derive the probability distribution of the queue length and the energy depletion time for a given initial energy level. Based on the energy sustainability analysis, we propose a distributed admission control strategy at SBSs striking a balance between high resource utilization and energy sustainability in the green HetNet. Extensive simulations are conducted to validate the analytical framework and evaluate the sustainability performance of the green HetNets using the proposed distributed admission control scheme. The simulation results demonstrate that relaxing the admission control criteria for the SBSs can improve the resource utilization (i.e, power and spectrum) of the system when the energy is abundant, but can significantly degrade the resource utilization instead when the energy comes short due to poor sustainability performance (i.e., frequent depletion of the SBSs).

R. Zhang (✉) • M. Wang • X. (Sherman) Shen • L.-L. Xie
Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, K2B8E3
e-mail: r62zhang@uwaterloo.ca

L.X. Cai • Y. Cheng
Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL, USA, 60616-3793



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_19

## 1 Introduction

The wireless communication industry is facing a great challenge with the expanding demand of mobile subscribers for broadband multimedia services. A recent study forecasted that the global mobile data traffic will increase tenfold between 2014 and 2019, and mobile video will account for more than 70 % of the total data by the end of 2019 [1]. As such, the main objective of the next generation mobile wireless networks is to further increase the network capacity and provision quality of service (QoS) guarantee for high densities of mobile users at the minimum cost. Dense deployment of heterogeneous networks (HetNets) has been recognized as a desirable and feasible solution for increasing the spatial network capacity and QoS provisioning in the fourth generation wireless networks [2]. However, a large number of small cell base stations (BSs) account for high capital expenditure (CAPEX) and operational expenditure (OPEX). In addition, a high volume of energy is required to power multiple small cell BSs (SBSs) in the cellular network, which also results in more greenhouse gases and carbon footprints, besides the increased energy costs. Therefore, one of the most critical issues in the next generation 5G wireless network is to design and develop new green technologies to reduce the energy consumption and costs of HetNets.

To achieve a green and sustainable wireless network, several research projects have been launched worldwide. OPERANET project aims to improve the power efficiency of mobile radio networks by monitoring and controlling the energy use and optimize the energy consumption at the BSs [3]. Mobile VCEs Green Radio programme examined the green techniques across the protocol stack for power reductio [4]. Most existing green communication networking solutions mainly focus on minimizing the energy consumption of the cellular network infrastructure, i.e., BSs, by improving hardware design, network planning [5, 6], energy efficient communications [7–9], and radio resource management [10–12]. With the recent advances in green energy techniques, it becomes a promising enabling technology to power SBSs with renewable energy sources such as solar and wind. By analyzing both the CAPEX and OPEX, it is shown in [13] that deploying green energy powered HetNets, especially the SBSs, is an economically convenient and sustainable solution for mobile operators, and it is possible to use harvested energy of BSs to serve plausible traffic loads in a dense HetNet [14].

Unlike traditional energy supply, green energy sources are by nature sustainable in the long term yet are variable and unstable in the short term, as the energy harvesting process is highly dependent on the charging environment, i.e., the weather, the geo-location, and the time. For example, a wind turbine provides intermittent power while a solar panel can supply relatively continuous power with varying output over a day and through the seasons. To fully utilize the harvested green energy, usually a rechargeable battery of large capacity is needed to store the charged energy and to provide a reliable and sustainable power supply for future use. In such a green energy powered wireless HetNet, the main performance metric has been shifted from energy efficiency to energy sustainability [15], i.e., the dynamically harvested energy at the BSs can sustain the traffic demands of

mobile users in a multi-tier HetNet. For instance, a green BS powered by renewable energy sources can be on and maintain normal operation when the charged energy is sufficient to provision services to its associated mobile users. However, a green BS will become off and out of service if the charged energy is used up and cannot sustain the service demands of mobile users. When this happens, all users associated with the out-of-power BS will experience excessive service delays or even service interruptions. Thus, it is indispensable to deliberately design the network resource management and admission control strategies to minimize the probability that an SBS depletes its energy and becomes out of service.

In this article, we first present a literature survey on existing green communication and networking techniques, and identify the relevant challenging issues in the design and development of a green HetNet powered by renewable energy sources. We then study resource management issues in a multi-tier HetNet where a macro cell BS (MBS) and multiple SBSs co-exist to provide services for mobile users in the network area. Specifically, we first develop an analytical model to analyze the energy buffer evolution at SBSs, characterizing the dynamics in the energy charging and discharging processes. Based on the analytical model, we propose an distributed admission control strategy to ensure the quality of service (QoS) requirements of admitted users, while maximizing the resource utilization and ensuring the sustainable performance of the green HetNets.

The main contributions of the chapter are four folds.

- Analytical model of energy buffer: A generic analytical model is developed to study the transient evolution of energy buffer at a HetNet BS. Basically, the rechargeable battery at each BS is used as an energy buffer to store the harvested energy, which can be modeled as a G/G/1 queue with general energy charging and discharging processes. Based on the first two statistical moments, i.e., the mean and variance, of the energy charging and discharging intervals, the diffusion approximation is applied to derive closed-form distributions of the transient queue length, i.e., the available energy level.
- Energy sustainability analysis: Using the developed analytical model, we then analyze the energy sustainability performance of the green HetNets. Specifically, two main metrics are derived to evaluate the sustainable performance of BSs: (1) the energy depletion probability: the probability that a BS depletes the energy and becomes unavailable for service; and (2) the duration for an BS to deplete its energy: given the current available energy level, the time duration that the BS can continue to serve the users until it depletes the energy and becomes unavailable.
- Distributed admission control: Based on the energy sustainability analysis, a distributed admission control strategy is deployed at BSs to provision QoS guarantee for admitted users and ensure desirable energy sustainability performance of the network while achieving high resource utilization.
- Performance evaluation: We conduct extensive simulations to validate the analytical model of energy buffer and examine the sustainable performance of the proposed distributed admission control scheme. We study the network stainability under different admission control parameters, and further investigate

the impacts of the energy depletion at one or multiple BSs on the overall performance of the green HetNets.

The remainder of the chapter is organized as follows. In Sect. 2, we give a broad overview of the existing works on green radio communication networks in the literature. In Sect. 3, the system model is presented including the network scenario and the energy model. An analytical framework is developed to study the energy sustainability performance of the BSs in Sect. 4, followed by an adaptive admission control strategy in Sect. 5. Numerical results are presented in Sect. 6. Finally, we conclude the chapter and introduce our future work in Sect. 7.

## 2 Understanding the Green HetNets

In this section, the role of the green HetNets in 5G wireless communications is first described. Then, the state-of-the-art of the research works related to green HetNets is introduced in details. Specifically, we classify the related research works into the following two categories: (1) energy efficient green HetNets; and (2) green energy powered wireless networks.

### *2.1 Green HetNets in 5G*

The wireless communications in 5G era are expected to fulfill a wide variety of requirements from both the operator and the user perspectives. On one hand, in order to meet the explosive growth of the mobile traffic, the wireless industry needs to significantly boost the system capacity by both excavating more spectrum resources and increasing the spectrum efficiency. On the other hand, the unprecedented expansion and densification of the wireless networks will cause huge energy consumption and cost, which has become a major challenge in commercializing the 5G systems. Green networking, as an energy-saving and eco-friendly approach to solve the energy-related challenges, has won ever-increasing attention in both academia and industry by its foreseeable potentials. The meaning of "green" has two folds: energy efficiency and renewable energy sources. Developing energy-efficient mechanisms for the radio access part of the system improves the system service quality without increasing much energy cost; and using renewable energy sources to power the wireless networks can effectively reduce the carbon footprint and relieve the greenhouse effects. Both folds have gained wide consensus and been incorporated as the main issues of the 5G development. Figure 1 illustrates the connections among different research issues covered by 5G, where rectangles with red rims show how the subject investigated in this chapter is connected to 5G.

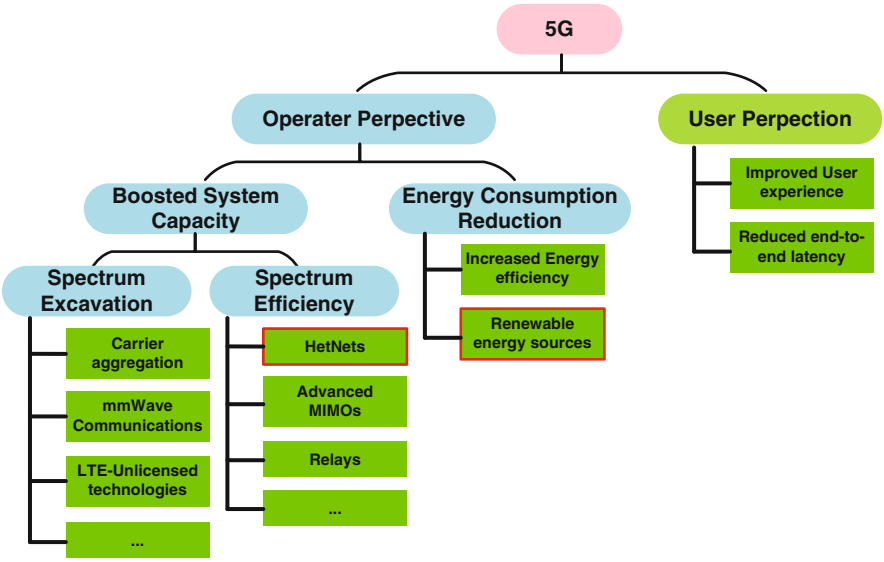


**Fig. 1** Connections among different research issues covered by 5G

## 2.2 *Energy Efficient Green HetNets*

It is reported that the major part of the energy consumption (as much as 70 %) for typical cellular networks comes from the radio access part [16]. Therefore, improving the energy efficiency of the cellular access network is of crucial importance in reducing the OPEX as well as cutting down the carbon footprints. To this end, the HetNet becomes a promising solution in significantly enhancing the network energy efficiency as densely deployed small cells can provide high-speed services at a low transmission power. There have been a massive number of existing works studying energy efficiency improvement in HetNets [7, 8], which can be generally divided into two categories according to different pointcuts, namely network topology control and radio resource management.

Network topology control is an effective methodology for improving energy efficiency of HetNets, by optimizing the energy efficiency through deliberate network planning and user association strategies subject to a set of system and user constraints, e.g., the maximum power constraints of the BSs and the user QoS requirements. As for the networking planning, the optimal network deployment in terms of the minimal number of BSs and the deployment locations are determined based on different metrics in different scenarios [17–21]. For instance, in [17], Zhang et al. studied the optimal energy efficiency in a HetNet consisting of a macrocell and multiple small cells. Exploiting the stochastic geometry theory [22], the optimal small cell density was derived to maximize the energy efficiency under coverage constraints. In [18], Klessig et al. investigated a HetNet composed of a

random number of microcells overlaid with multiple hexagonal macrocells. The optimal number of microcells for a reference cell area was determined to maximize the energy efficiency under different network loads. In [19], instead of dealing with the BSs that are always powered on, Saker et al. proposed to utilize macrocell sleeping mechanisms to improve the network energy efficiency. Specifically, the sleep and wake-up of the macrocells are carefully controlled to offload the users from macrocells switched to sleep mode to small cells while satisfying the users' QoS requirements. In [20], Zheng et al. put forward an energy-efficient scheme where femtocells cooperate in clusters to decide in an autonomous manner when and which femtocells should sleep and wake up. Particularly, femtocells with different user access policies are taken into account. In [21], Cao et al. examined how the energy efficiency of HetNets is impacted by the partial spectrum reuse (PSR) factor of the microcells, which is defined as the portion of spectrum reused by microcells. Optimal PSR factor was derived to determine whether to deploy more microcells or switch off less macro BSs. In the aspect of user association, energy efficiency is studied on both user side [23] and network side [24]. In [23], Zhang et al. studied the energy-saving performance of user equipment in a HetNet with carrier aggregation (CA) [25] and demonstrated that CA can be leveraged to save UE energy while satisfying the user demand. In [24], an user association algorithm was proposed to strike an energy-saving balance between the radio access and the backhaul links in HetNets.

Improving the energy efficiency of HetNets via resource management mainly aims at optimizing network resource allocation (e.g., bandwidth and power) to maximize energy efficiency, subject to the intra- and inter-tier interference in HetNets [10–12, 26, 27]. For instance, [10] studied how to optimize the transmission power of the femtocells to mitigate inter-cell interference while maximizing the energy efficiency of the system. The results revealed that a much higher energy efficiency could be achieved by sacrificing a relatively low loss in the spectral efficiency. In [11], the orthogonal frequency division multiple access (OFDMA) based HetNets were studied, where the joint subchannel allocation and power control problem was formulated to maximize the sum energy efficiency of all BSs subject to the BS power constraints and subchannel exclusivity constraints. In [12], the energy efficiency was investigated in a cognitive radio cellular network where the resource allocation problem was modelled as a Stackelberg game between cognitive macrocells and femtocells. In [26], energy-efficient pre-coding scheme for coordinated multi-point (CoMP) transmission in HetNets was studied, where the MBS and SBS can cooperatively transmit to a same user to provide better services. The optimal precoder structure was obtained to minimize the energy consumption with considering the inter-cell interference. In [27], the backhaul link power consumption of the CoMP HetNet was considered, and the resource allocation problem was modelled into an energy efficiency maximization problem subject to the spectral efficiency and user QoS constraints.

### *2.3 Green Energy Powered Wireless Networks*

Recent advances in green energy technologies make it a cost-effective solution to power wireless networks compared with the traditional energy supplies. The optimal rate adaptation in a single user energy harvesting system was studied in [28] to minimize the packet delivery time. In [29], the optimum transmission policies was presented for energy harvesting nodes with limited batteries. In [30], a directional water-filling algorithm was introduced to achieve optimal transmission of energy harvesting nodes over a wireless fading channel. Optimal power adaptation schemes were studied in [31] to maximize the achievable rate of a single-user channel in a device-to-device communication underlaying green cellular network, considering both total power constraint and the BS power constraint, and further extended in [15] to a single user relay channel for cooperative communications.

Besides the studies in a single user channel, the system performance of green energy powered network were examined in different contexts. Solar energy powered sensor network was first studied in [32], and experiments showed that near-perpetual operation of a sensor node was possible with the developed prototypes of sustainable sensors. The solar and wind powered wireless local area networks (WLANs) were studied in [33, 34] and a solar battery configuration methodology along with an outage control algorithm were proposed to minimize the energy outage at wireless nodes. In [35], an optimal green AP placement problem was formulated to find a minimal number of green access points (APs) powered by renewable energy to be deployed in constrained locations and to provision QoS to mobile users. The network throughput and energy sustainability in the green-energy-powered maritime wireless network was studied in [36]. Resource allocation and energy management in sustainable mesh networks have also been studied. In [37], the fundamental design criterion of a sustainable wireless network was elaborated. It explained that under the green network paradigm powered by renewable energy, communication and networking protocols across the protocol stack should be re-visited, as the design criteria has been shifted from energy efficiency to energy sustainability. Different routing schemes are examined and compared in the context of sustainable networks in [38], aiming to distribute the traffic loads evenly over the network to improve the network sustainability. It was shown in that traditional routing strategies which did not consider environmental energy supply achieve poor performance, and it was crucial to adapt the routing decisions to the time varying power supply conditions in a sustainable network environment. However, how to accurately capture and model the power conditions was not discussed. In our previous work [39], a generic energy model was developed to characterize the dynamic energy charging and discharging processes, and an adaptive resource management scheme was proposed for a green mesh network using renewable energy. The objective was to maximize the energy sustainability of the network, or equivalently, to minimize the failure probability that the mesh APs depleted their energy and went out of service by distributing the traffic loads appropriately across the mesh network.

With the dense deployment of small cells to improve the capacity of a multi-tiered cellular network, it is well recognized that small cells can benefit from green energy supplies while macro cells usually require fairly high power and less likely to use harvested energy to serve mobile users. Generally, green energy powered BSs have a higher CAPEX but a lower OPEX, and vice versa for traditional BSs. By careful analyzing the CAPEX and OPEX of the two types of BSs, it was shown in [13] that gradually deploying green energy powered small cell BSs was an economically convenient solution for mobile operators to provide sustainable network services. The spatial and temporal dynamics of energy harvesting in a cellular network was studied in [14, 40–42]. The availability region for a set of self-powered BSs was investigated in [14], considering the uncoordinated on and off operations of green BSs based on the available energy levels. A new network architecture that overlays an uplink cellular network was proposed in [40] to deploy power beacons which can charge mobile devices by microwave radiation. An energy field model was developed to analyze the coverage of self-powered green cellular network in [41], where the energy harvesters are distributively deployed in clusters and the generated energy from the same cluster was aggregated and then redistributed to power BSs. In [42], a joint network deployment and resource management problem was formulated for a two-tier cellular network such that the minimal number of small cell BSs were deployed with optimal subcarrier allocation to ensure full network connectivity and users' QoS requirements can be fulfilled with the harvested energy under the cost constraint.

The aforementioned works mainly studied network deployment or power management schemes to achieve a sustainable green cellular network. To the best of our technology, the resource management issues especially the admission control strategy for a green HetNet is not well investigated yet. In addition, due to the dynamic energy charging and limited battery size, the energy outage may occur stochastically. No existing work has extensively studied the impacts of energy depletion of one or multiple SBSs on the overall sustainable network performance.

## 3 System Model

### *3.1 Network Model*

We consider a heterogeneous network composed of different types of base stations serving different coverage areas. As shown in Fig. 2, there is one macrocell and multiple small cells. Due to the high power consumption and critical always-on operation requirements, the MBS is powered by both renewable energy and traditional energy from the electricity grid[1]; while the small cell base stations (SBSs) are powered by renewable energy supplies. In such a HetNet, the MBS uses the whole spectrum band for communications and each SBS uses only a

[1]For the MBS, renewable energy source is the first choice. The electricity grid provides supplemental energy source to guarantee the power supply of the MBS at all times.

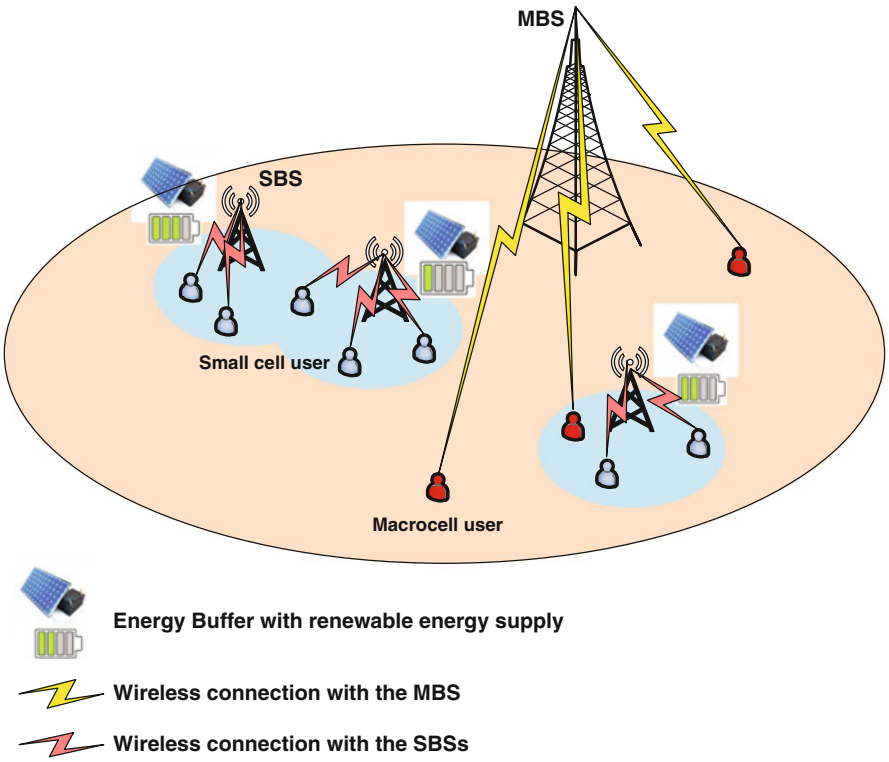


**Fig. 2** Green HetNets with renewable energy supplies

subset of the spectrum bands for communications. Generally, orthogonal frequency division multiple access (OFDMA) is adopted as the radio access technology for both macrocells and small cells, and small cells will experience co-channel interference from the macrocell and the neighboring small cells when they use the same frequency subcarriers for communication. Denote the MBS as $m$ and the set of SBSs as $\mathbb{S}$. Each small cell BS can provide high-rate multimedia services to the users within its coverage range. One typical network scenario is a household or an office building where a green SBSs with solar panels or wind turbines can be installed on the roof of the building to harvest energy from the environment, e.g., sunlight or wind. When one user arrives in the system, it checks for the available BSs in the neighborhood. In a high density HetNet, it is possible that a user is within the coverage areas of multiple SBSs, and the user will first choose the closest SBS to associate with, provided that it can be accommodated by the target SBS based on certain admission control criteria. If the association request is rejected, the user will choose the next closest SBS to associate with. The process is repeated until the user is successfully admitted by a BS, either a small cell BS in its neighborhood or the MBS if no nearby small cell has sufficient network resources to serve the user.

As the initial step, only downlink transmissions are considered for both macrocells and small cells. For small cells, bandwidth is usually sufficient and the renewable energy supply is the bottleneck. This is acceptable since bandwidth can be rich with advanced wireless technologies (e.g., carrier aggregation in both licensed and unlicensed band[2] and a small number of users due to the relatively small coverage of SBSs while energy is more limited resource when renewable energy is exploited to provide mobile multimedia services. A SBS is on operation when it has sufficient charged energy to serve the associated users, and will become off and out of service if the charged energy is used up. When this happens, all users associated with the out-of-power SMB need to re-associate with nearby SBSs or MBSs, which causes long delay or even interrupted services. A out-of-power SBS may resume to on state for services when the energy is charged to a certain threshold. For the always-on MBS, the energy is sufficient due to the complementary energy supply from electricity grid, yet the bandwidth may become the bottleneck when there are a large number of users associated with the MBS, due to the large communication coverage of MBS.

Each mobile user has a certain QoS requirement, i.e., the minimum transmission rate of user $i$ is denoted as $r_i,\ i \in \mathbb{U}_s\ (\mathbb{U}_m)$, where $\mathbb{U}_s\ (\mathbb{U}_m)$ is the set of users that are associated with SBS $s$ (MBS $m$). According to the Shannon theory, $r_i$ in small cells can be derived as

$$r_i = W_{si} \log(1 + \frac{P_t {d_{si}}^{-\alpha}}{n_0 W_{si} + \sum\limits_{j \in \{m\} \cup \mathbb{S} \setminus s} I_{ji}}),\quad \forall i \in \mathbb{U}_s,\ \forall s \in \mathbb{S} \tag{1}$$

where $W_{si}$ is the bandwidth assigned by SBS $s$ to its associated user $i$, $P_t$ is the transmission power of the SBS to the associated user $i$, $d_{si}$ is the distance between the SBS $s$ and the associated user $i$, $\alpha$ is the pass loss exponent, $n_0$ is the noise power spectrum density, and $I_{ji}$ is the perceived interference at user $i$ from base station (BS) $j$ (either MBS or SBS). The perceived interference $I_{ji}$ is a stochastic variable which is closely related to the distance between user $i$ and neighboring BS $j$, the transmission power of BS $j$ and the sub-carrier allocation at each BS. The perceived interference can be effectively estimated leveraging standard estimation methods [43]. Since bandwidth is rich for small cells in the considered scenario, $W_{si}$ is adaptively adjusted to combat the perceived interference to satisfy the transmission rate requirement (i.e., $r_i$) of each user. Similarly, if user $i$ is served by the MBS $m$, the transmission rate of user $i$ is

$$r_i = W_{mi} \log(1 + \frac{P'_t {d_{mi}}^{-\alpha}}{n_0 W_{mi} + \sum\limits_{j \in \{m\} \cup \mathbb{S} \setminus m} I_{ji}}),\quad \forall i \in \mathbb{U}_m. \tag{2}$$

where $P'_t$ denotes the transmission power of MBS to each macrocell user.

[2]FCC opened 100 MHz spectrum in 5 GHz band for unlicensed use in 2014, and plans to open an additional 195 MHz spectrum in the near future. The abundant unlicensed spectrum band provides great potential for cellular operators to deploy high density small cells that operates in unlicensed 5 GHz band.

### 3.2 Energy Model

Each green BS is equipped with an energy buffer that stores the harvested renewable energy for future use. Denote the buffer energy level at time $t$ by $E_s(t)$, $s \in \mathbb{S} \cup \{m\}$. Let $E_s(0)$ be the initial energy buffer level, and $C_s(t)$ be the amount of charged energy over the time interval $[t-1, t]$. Due to different environmental conditions (e.g., various sunlight intensities or wind strengths at different geographical positions and time periods), the charging rate of an SBS may vary uniquely with each other. Thus, in this paper, the buffer charging process is modelled as a continuous-time stochastic event arrival process with an arbitrary but stationary distribution; the charged energy units are considered as the arrival events in the buffer. Denote the mean and variance of the inter-arrivals of the energy charging as $\mu_c$ and $\sigma_c^2$, respectively. For SBS $s$, the energy stored in the buffer at time $t$ is subject to the boundary conditions, i.e., $0 < E_s^{min} \leq E_s(t) \leq E_s^{max}$. $E_s^{min}$ is the minimum required energy in the buffer to ensure normal BS operations such as signal processing and cooling system; $E_s^{max}$ is the maximum capacity of the energy buffer. Without loss of generality, this energy buffer model can be further simplified by letting $0 \leq E_s(t) \leq B_s$ where $B_s = E_s^{max} - E_s^{min}$. In the following section, we study the case that the buffer capacity is sufficiently large to store the harvested energy, i.e., $B_s \to \infty$. The analysis for finite buffer capacity will be included in the future work.

Denote the energy used for of a packet transmission as $e^u$. We normalized that $e^u = 1$ energy unit for presentation convenience. The packet arrival of each user is modelled as a stochastic process with an arbitrary but stationary distribution. The mean and variance of the inter-packet arrival are $\mu_0$ and ${\sigma_0}^2$, respectively. Given the traffic model for a single user, the mean and variance of the aggregated packet arrivals for one SBS $s$ can also be measured, which are equivalent to the mean (denoted as $\mu_{s,d}$) and variance (denoted as $\sigma_{s,d}^2$) of the discharging interval of the energy buffer in SBS $s$, respectively.

Given the above energy buffer model, the energy queue evolves as follows. The total amount of energy of SBS $s$ discharged during $[t-1, t]$, denoted as $D_s(t)$, is given by

$$D_s(t) = \sum_{i \in \mathbb{U}_s} N_i(t) e^u, \tag{3}$$

where $N_i(t)$ denotes the number of packets of user $i$ transmitted by SBS $s$ within $[t-1, t]$. Therefore, the energy level of SBS $s$ at time $t$ is calculated as

$$E_s(t) = \min\{\max\{E_s(t-1) + C_s(t) - D_s(t), 0\}, B_s\}.vs \tag{4}$$

The evolution of the energy buffer of an SBS is illustrated in Fig. 3. Generally, the energy charging and discharging are random processes, and the rates of which vary over time yet their first and second moments can be measured. The energy buffer increases when the energy charging rate is higher than the energy discharging rate, and decreases otherwise. An energy buffer may deplete temporally when the energy level reaches 0, i.e., $E_s(t) = 0$.

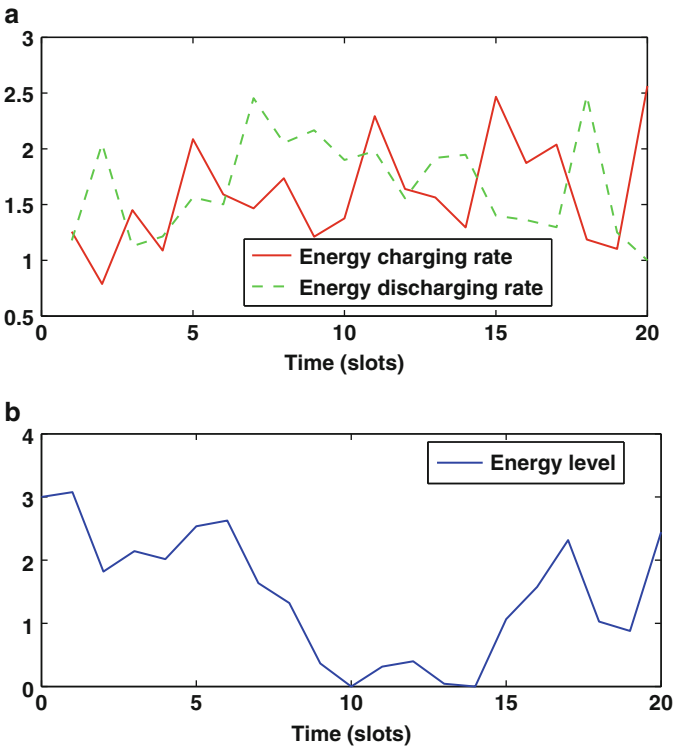


**Fig. 3** Energy buffer evolution under random charging and discharging processes. (**a**) Energy charging and discharging processes. (**b**) Energy buffer evolution

## 4 Energy Sustainability Analysis

In this section, a generic analytical framework is developed to analyze the energy buffer. The diffusion approximation approach is applied to study the energy buffer evolution in the infinite buffer size case. For notational simplicity, we omit the SBS index $s$ in the notations of the previous section (e.g., $E_s(t)$ to $E(t)$). Given the initial energy level $E(0)$, the buffer may be depleted when $E(t)$ reaches 0 due to energy discharging for downlink transmissions. When one SBS is depleted, it becomes temporally unaccessible and all the associated users need to find alternative BSs (other nearby SBSs or the MBS) to connect with based on the association policy described in Sect. 3.1. Therefore, it is critical to ensure energy sustainability of SBSs. To address this issue, we need to answer the following questions: (1) how to derive the depletion probability for a given initial energy level in the buffer and cell load conditions, (2) how to design effective admission control policy based on the depletion probability, and (3) to what extend the depletion of one or multiple SBSs will degrade the overall network performance.

With infinite buffer capacity, all the harvested energy can be stored in the buffer. According to the system model, both the charging and discharging processes are random processes with arbitrary but stationary distributions. Thus, the energy buffer can be modelled into a G/G/1 queue. The mean and variance of the inter-charging (inter-discharging) intervals are $\mu_c$ ($\mu_d$) and $\sigma_c^2$ ($\sigma_d^2$), respectively. To make the G/G/1 queue tractable for analysis, the diffusion approximation method is exploited by approximating the discrete-event queue via a continuous random process [44]. Specifically, the discrete energy queue length $E(t)$ is approximated to a continuous random process $Y(t)$ in the way that the increase of $Y(t)$ over a small period $dt$ follows a normal distribution, i.e.,

$$\begin{aligned} &dY(t) = Y(t+dt) - Y(t) = \mu dt + n(t)\sigma\sqrt{dt}, \\ &\text{where } \mu = \tfrac{1}{\mu_c} - \tfrac{1}{\mu_d},\ \ \sigma^2 = \tfrac{\sigma_c^2}{\mu_c^3} + \tfrac{\sigma_d^2}{\mu_d^3}, \end{aligned} \tag{5}$$

In Eq. (5) $\mu$ and $\sigma$ are the mean and standard variance of $dY(t)$, respectively, and $n(t)$ is the standard Gaussian process (i.e., with zero mean and unit variance). Given the initial energy level $E(0) = e_0$, the conditional probability density function (pdf) of $Y(t)$, which also refers to the energy buffer level at time $t$ $E(t)$, can be expressed as

$$f_Y(y, t|e_0) = \frac{\mathrm{P}(y < Y(t) \leq y + dy|e_0)}{dy}. \tag{6}$$

According to [44], $f_Y(y, t|e_0)$ satisfies the forward diffusion equation

$$\frac{\partial f_Y(y,t|e_0)}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 f_Y(y,t|e_0)}{\partial y^2} - \mu\frac{\partial f_Y(y,t|e_0)}{\partial y}, \qquad \text{for} \quad Y(t) \geq 0 \text{ and } t > 0. \tag{7}$$

To solve the above forward diffusion equations, the method of images [45] is utilized to obtain the expression of the conditional p.d.f. as

$$f_Y(y, t|e_0) = \tfrac{\partial}{\partial y}\left\{Z(\tfrac{y-e_0-\mu t}{\sigma\sqrt{t}}) - e^{2\mu y/\sigma^2} Z(-\tfrac{y+e_0+\mu t}{\sigma\sqrt{t}})\right\}, \tag{8}$$

where $Z(y)$ is the standard normal integration, $Z(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$.

After obtaining the conditional p.d.f. of the energy buffer size at time $t$, i.e., $f_Y(y, t|e_0)$, two main performance metrics can be derived: (1) the depletion probability given initial energy level $e_0$; and (2) the distribution of the energy depletion time, denoted by $T(e_0)$, i.e., from the initial time with the energy level of $e_0$ till the time when the energy buffer full depletes and reaches zero for the first time, $E[t] = 0$. Thus, we have

$$T(e_0) = \inf(t \geq 0 | Y(t) = 0, Y(0) = e_0). \tag{9}$$

The energy depletion time is an important sustainability metric for green HetNets which indicates the maximum time duration that the SBS can sustain the demands of the small cell operation before it becomes out of service, for a given initial energy level. In stochastic theory, $T(e_0)$ is also referred to as the first passage time of $Y(t)$ from $e_0$ to 0. The p.d.f. of $T(e_0)$, denoted as $f_T(t|e_0)$, can be derived similarly to $f_Y(y,t|e_0)$ based on the diffusion equations with the absorbing barrier at 0. According to [44], the diffusion equations are given by

$$f_T(t|e_0) = \lim_{x\to 0}\left\{\frac{\sigma^2}{2}\frac{\partial f_Y(y,t|e_0)}{\partial y} - \mu f_Y(y,t|e_0)\right\}. \tag{10}$$

Based on Eq. (10), the p.d.f. of $T(e_0)$ can be derived as

$$f_T(t|e_0) = \frac{e_0}{\sqrt{2\pi\sigma^2 t^3}}\exp\Big[-\frac{(e_0+\mu t)^2}{2\sigma^2 t}\Big]. \tag{11}$$

Then, the moment generation function of $T(e_0)$ can be expressed as

$$\Phi_T(s|e_0) = \int_0^{+\infty}\exp(st)f_T(t|e_0)dt = \exp\Big[-\frac{e_0}{\sigma^2}(\mu+\sqrt{\mu^2-2\sigma^2 s})\Big]. \tag{12}$$

The moment generation function can be utilized to derive the first and second moments of a random process by calculating the first-order and second-order differential equations of Eq. (12) with respective to $s$, respectively. Thus, the mean and variance of $T(e_0)$ when $\mu \neq 0$ are calculated as

$$E[T(e_0)] = \frac{\partial\Phi_T(s|e_0)}{\partial s}|_{s=0} = \begin{cases}\exp(-2e_0\mu/\sigma^2)e_0/\mu, & \text{if } \mu > 0,\\ -e_0/\mu, & \text{if } \mu < 0;\end{cases} \tag{13}$$

and

$$\begin{aligned} V[T(e_0)] &= E[T(e_0)^2] - E^2[T(e_0)] = \frac{\partial^2\Phi_T(s|e_0)}{\partial s^2}|_{s=0} - E^2[T(e_0)] \\ &= \begin{cases}\exp(\frac{-2\mu e_0}{\sigma^2})(\frac{e_0^2}{\mu^2}+\frac{e_0\sigma^2}{\mu^3}) - \exp(\frac{-4e_0\mu}{\sigma^2})\frac{e_0^2}{\mu^2}, & \text{if } \mu > 0,\\ -e_0\sigma^2/\mu^3, & \text{if } \mu < 0.\end{cases} \end{aligned} \tag{14}$$

If $s \to 0$, $\Phi_T(s|e_0)$ will result in the conditional probability of approaching to the absorbing barrier (i.e., $Y(t) = 0$), given the initial energy level of $e_0$. Thus, we have

$$\mathrm{P}(Y = 0|e_0) = \lim_{s\to 0}\Phi_T(s|e_0) = \begin{cases}\exp(-2e_0\mu/\sigma^2), & \text{if } \mu \geq 0\\ 1, & \text{if } \mu < 0.\end{cases} \tag{15}$$

It can be seen from Eq. (15) that when the energy charging rate is smaller than the energy discharging rate, i.e., $\mu < 0$ (or equivalently, $1/\mu_a < 1/\mu_d$), the energy

buffer will deplete eventually and $Y(t|e_0)$ will become 0 with probability 1. When $\mu \geq 0$, although the energy charging rate is no less than the energy discharging rate, it is still possible that the energy buffer may deplete with a certain probability due to the variances of the charging and discharging rates. In such case, the depletion probability is dependent on the initial energy level $e_0$, and the first and second moments of the energy charging and discharging processes.

## 5 Adaptive Admission Control Strategy for Small Cells

In this section, a distributed energy-aware user admission control strategy is proposed for small cells based on the queueing analysis of the energy buffer. Generally in a green HetNet, as a user will always attempt to associate with a nearby SBS first for achieving a higher transmission rate at a shorter distance and with a lower transmission energy cost, compared with the MBS; and a user will associate with the MBS only when no nearby SBS can accommodate it. Besides the bandwidth and energy efficiency of small cell communications, it is desirable because the MBS may need to use the energy from the electricity grid to serve wireless users, which involves extra cost. Thus, our admission control strategy at SBSs aims at protecting each SBS from being depleted its energy and becoming out of service due to admitting too many users while helping relieve the traffic burden from the MBS as much as possible, taking into consideration the transient energy level, cell load conditions, and charging capabilities. The strategy is implemented in each SBS and is conducted when the network conditions change (e.g., when a user arrives at or leaves a small cell, energy buffer of a small cell depletes, and a small cell recovers from depletion).

Basically, the admission control strategy tackles the tradeoff between the sustainability of the small cells and the traffic offloading from the macrocell. On one hand, according to Eq. (15), when the energy charging rate is higher than the energy discharging rate at one SBS (i.e., $\mu > 0$), the SBS depletion probability reduces as the discharging rate reduces. In other words, admitting a fewer number of users to one SBS can help improve the SBS sustainability. On the other hand, if a fewer number users are admitted to small cells, the MBS has to provision QoS to more users, probably resulting in bandwidth exhaustion after the network capacity is reached when the user density is high. When the bandwidth exhaustion occurs, no more users should be admitted, otherwise the QoS requirements of some macrocell users cannot be guaranteed, leading to significant user dissatisfaction. Therefore, the energy-aware admission control strategy should deliberately adjust the admission criteria to well balance the tradeoff.

In a small cell, each admitted user should be provided with guaranteed service; that is, once the user is admitted by a small cell, its transmission rate requirement is guaranteed by 100 %. When a new user sends an association request to a SBS at time $t$, the SBS will check the traffic demands of the user, admit it only if its depletion probability condition is under a desired threshold $\epsilon$ after admitting the user, based

on its current energy level (i.e., $E(t)$) and the energy charging capability, i.e.,

$$\begin{aligned} &\mathrm{P}(Y=0|E(t)) \leq \epsilon \ll 1. \\ &\Rightarrow \mathrm{P}(Y=0|E(t)) = \exp\left[-2E(t)\mu/\sigma^2\right] < \epsilon,\ \mu > 0. \end{aligned} \tag{16}$$

The threshold $\epsilon$ is an adjustable parameter but should be kept far less than 1, $\epsilon << 1$, to guarantee a sufficiently small risk of energy depletion, which leads to a necessary condition that the charging rate of the energy buffer is required to be larger than the discharging rate. Smaller $\epsilon$ indicates more strict criteria in admitting a user. Based on the admission criteria of small cells, the admission control strategy is outlined as follows.

- When a user arrives in the network, it first creates a BS set (denoted as $\mathbb{S}_c$) within whose coverage it is located, and sort the BSs (SBSs and the MBS) in the descending order of the signal strength. The user will first send a network association request to the BS with the highest signal strength, usually the closest SBS. Upon receiving the request, the SBS will examine whether the admission criteria is satisfied. If so, the user will be admitted and associated with the SBS; if the condition cannot be satisfied, the user will be rejected and have to attempt network association with the next BS in the list of $\mathbb{S}_c$. If no small cells can accommodate the user, the user will be associated to the MBS for communications, given the network capacity is not reached or the bandwidth is not exhausted.
- When a user leaves a small cell so that the SBS can accommodate more users, the SBS will broadcast this information periodically. If any user within the coverage receives the message, it will send an association request to the SBS if the SBS has a better channel condition compared with the currently associated BS (MBS or SBS). The SBS processes the requests based on the first-come-first-serve principle and accommodate the requesting user(s) if the admission criteria (16) is satisfied in order to offload the traffic burden from the MBS and provide more bandwidth and energy-efficient communications.
- In addition, if by any chance an SBS depletes its energy (with a probability smaller than $\epsilon$), the SBS will be temporally unaccessible, and all the associated users may need to find an available SBS or the MBS for network access.
- After an SBS depletes, no users can be served, and the energy discharging rate becomes zero. The SBS starts the recovery process where the energy is accumulated in the battery or energy buffer. When the energy queue accumulates to $e_0$, the SBS starts to broadcast the information periodically. Users receiving the messages update the list of available SBSs and sort them in the descending order of the signal strength. A user served by other BSs may choose to associate with the resumed SBS if it provides better services, i.e., a higher signal strength, for energy-efficient communications.

To summarize, flow charts are presented in the following to illustrate the admission control and user association procedures for all situations (Fig. 4).

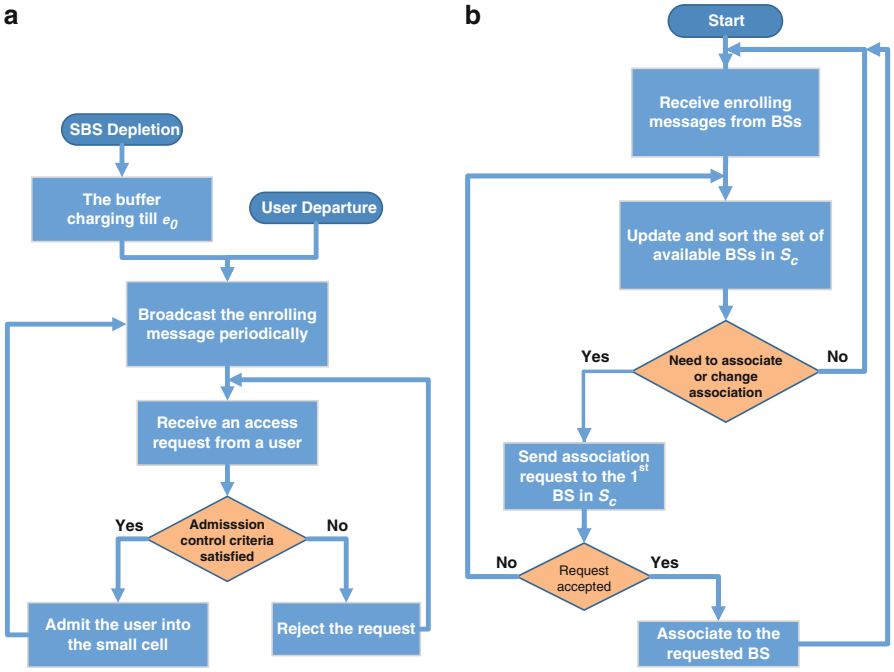


**Fig. 4** Flow chart for admission control and user association procedures for all situations. (**a**) Admission control procedure in SBSs. (**b**) User association procedure

## 6 Performance Evaluation

In this section, simulations are conducted to validate the analytical framework of energy buffer and evaluate the impact of admission control criteria on the resource (i.e., transmission power and bandwidth) utilization of the green HetNet powered by the renewable energy sources.

### *6.1 Simulation Setup*

In the simulation, we have a dense HetNet composed of one macrocell and $N_s$ ($N_s = 30$ if not otherwise specified) overlaid small cells, as shown in Fig. 2. The $N_s$ small cells (with a radius of $r$ meters) are uniformly distributed within the macrocell coverage (with a radius of $R$ meters) and may overlap with each other. The user arrivals and departures are modelled with Poisson processes with equal rates so that the number of users in the network can be kept at the same level in long term. For the user spatial distributions, one user may stay in a small cell with probability $5/6$ and outside of the small cells with probability $1/6$. Users are uniformly distributed

**Table 1** Simulation parameters

| Parameters | Values |
|---|---|
| Time slot duration | 1 ms |
| Coverage radius of the macrocell, $R$ | 500 m |
| Coverage radius of a small cell, $r$ | 30 m |
| MBS power spectrum density, $P_m$ | −23.5 dBm/Hz |
| SBS transmission power for a user packet, $P_t$ | −6.49 dBm |
| Noise power spectrum density, $n_0$ | −174 dBm/Hz |
| Available spectrum bandwidth for the MBS | 10 MHz |
| Subcarrier bandwidth | 15 kHz |
| Subcarriers allocation strategy | Random selection |
| User transmission rate requirement, $r_i$ | 300 kbps |
| User packet size | 150 bits |
| Path loss component, $\alpha$ | 2 |
| Initial energy level of SBS, $E_0$ | Uniform distributed on [5,25] energy units |
| Admission control criteria, $\epsilon$ | [0.001,0.5] |

both inside and outside the small cells. The simulation starts with 360 users with 10 users inside each small cell and 60 users outside the small cells.

For any user, the packet arrival process at the associated BS is modelled as Poisson process with an average inter-packet interval of $u_0 = 20$ time slots. According to the system model, the transmission energy of an SBS for one packet is $e^u = 1$ energy unit via adjusting the allocated bandwidth. For the energy charging process at a SBS, the inter-charging interval is randomly selected from $\{1, 2, 3, 4\}$ time slots with a certain probability mass function. For instance, if the probability for choosing $\{1, 2, 3, 4\}$ is $\{0.4, 0.4, 0.1, 0.1\}$, respectively, the mean and variance for the inter-charging interval will be $\mu_c = 1.9$ and $\sigma_c^2 = 0.89$, respectively. Other main parameters used in the simulations are summarized in Table 1. The simulation data is obtained via averaging over 1000 simulation runs.

## *6.2 Simulation Results*

We first examine the validity of the analytical model of energy buffer by investigating the evolution process of an energy buffer. The depletion probability of an energy buffer with different initial energy levels are shown in Fig. 5. We let $\mu > 0$ in the corresponding simulations since the depletion probability is 1 when $\mu < 0$. To reflect the relationship between the energy buffer and the cell load conditions, ten users are connected to an SBS without leaving the network. Two values of the average inter-packet interval are simulated (i.e., $\mu_0 = 20$ and $\mu_0 = 21$), respectively. It can be seen that the depletion probability descends with the increase of initial energy level $e_0$, and the exponential trends of the analytical

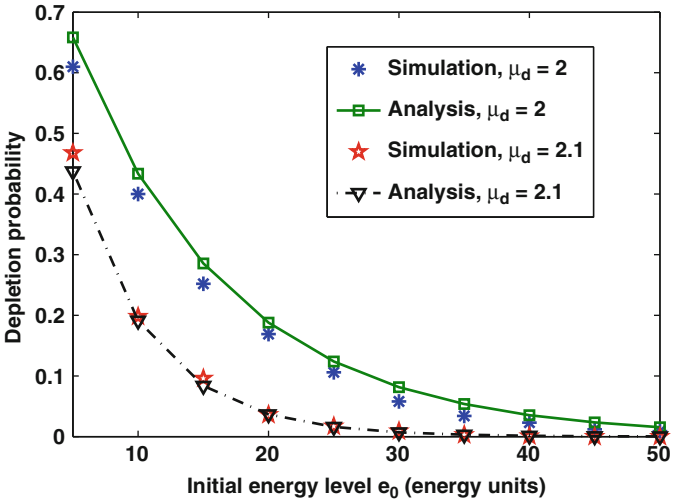


**Fig. 5** The depletion probability for an energy buffer $\mathrm{P}(Y(0) = 0|e_0)$ vs. initial energy level $e_0$. Default values: $\mu_c = 1.9$, $\sigma_c^2 = 0.89$, $\mu_d = 2$, and $\sigma_d^2 = 4$

results approach well with the simulation ones. With a higher initial energy level or an energy charging rate, it is less likely that the energy buffer will depletes, and thus the depletion probability is smaller. Besides that, the gap between the analytical and simulation results is slightly smaller when $\mu_d = 2.1$ than that when $\mu_d = 2$. This is as expected since the diffusion approximation approach is proven to be more accurate when the queue length is larger [44], and the average queue size when $\mu_d = 2.1$ is larger since the corresponding discharging rate is smaller than that when $\mu_d = 2$. This explanation also applies to the observation that the gap is slightly larger when the initial energy level $e_0$ is smaller.

The cumulative distribution function (CDF) of the first passage time from the initial energy level to 0 (i.e., $T(e_0)$) is presented in Fig. 6. Three $e_0$ values are simulated. It can be observed that the analytical results match well with the simulation results. In addition, for a given $T(e_0)$ value, the cumulative probability is larger for a smaller $e_0$. This is because when the energy charging rate is smaller than the energy discharging rate (i.e., $1/\mu_c < 1/\mu_d$), the queue depletes faster with a smaller initial energy $e_0$, leading to a smaller first passage time from $e_0$ to 0, i.e., $T(e_0)$. As a result, when $e_0$ is smaller, it is less likely that the SBS can sustain the traffic demands of its cell for a certain time, or equivalently the probability that $T(e_0)$ is smaller than a given value is higher.

After verifying the energy model, we then simulate the HetNet to investigate the impact of the admission control criteria $\epsilon$ on the resource utilization of the network, especially the MBS. For different $\epsilon$ values, two metrics are examined, i.e., the bandwidth exhaustion probability when the network capacity is reached, and the overall transmission power of the MBS. The results are shown in Fig. 7. It can

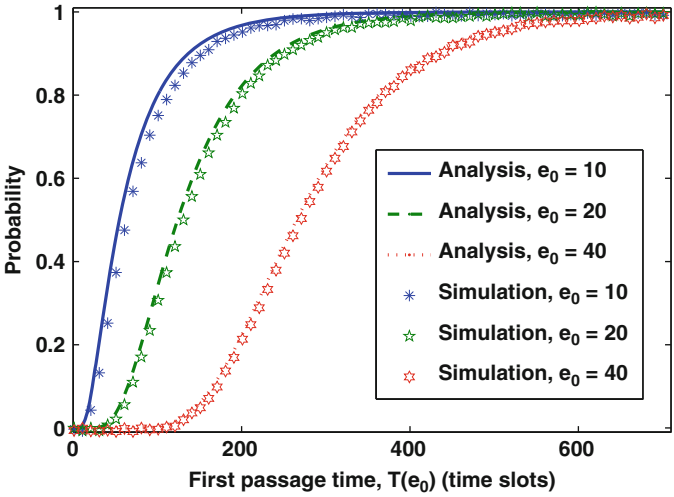


**Fig. 6** The CDF of the first passage time from $Y(0) = e_0$ to $Y(t) = 0$, i.e., $T(e_0)$. Default values: $\mu_c = 1.9$, $\sigma_c^2 = 0.89$, $\mu_d = 1.5$, and $\sigma_d^2 = 2.25$

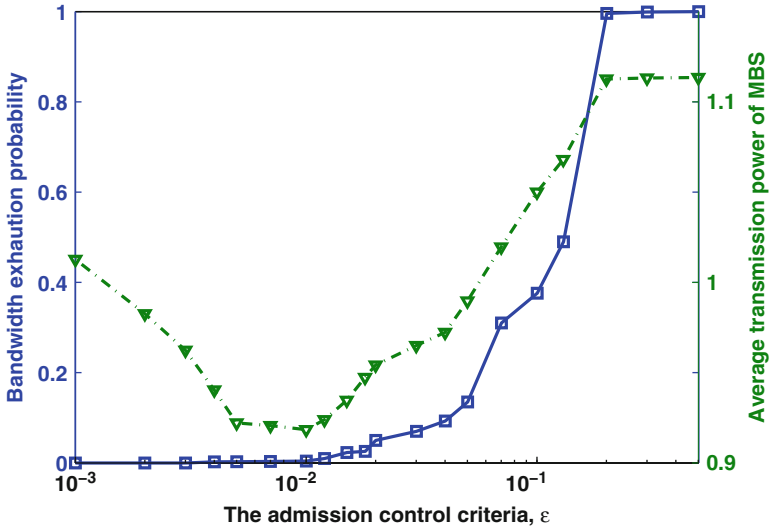


**Fig. 7** The impact of admission control criteria $\epsilon$ on bandwidth and power consumption of the MBS. Default values: $\mu_c = 1.9$, $\sigma_c^2 = 0.89$, $\mu_0 = 20$, $\sigma_0^2 = 400$, and $N_s = 30$

be seen that when $\epsilon$ increases from 0.001 to 0.5, the average overall transmission power of the MBS first decreases and then increases. This can be explained as follows. When $\epsilon$ is small, increasing $\epsilon$ will increase the number of admitted users in each small cell. As a result, fewer users are connected with the MBS, thus the overall transmission power of the MBS decreases. When $\epsilon$ is large, increasing $\epsilon$

will lead to a larger depletion probability for small cells. Consequently, although more users can be admitted with a larger $\epsilon$ for small cells, more small cells may deplete its energy and become unavailable for services; and the associated users with the unavailable SBSs will then connect to the MBS instead. In such a case, more transmission powers are used at the MBSs to provide services to the users whose associated SBSs are depleted.

The bandwidth exhaustion probability represents the probability that the MBS runs out of bandwidth or reaches its network capacity in providing QoS guaranteed service to the associated users. When bandwidth exhaustion occurs, the following user requests will be rejected, otherwise the QoS requirements of some associated users cannot be satisfied, which result in degraded service dissatisfaction of the users. It can be seen from Fig. 7 that when $\epsilon$ increases, the bandwidth exhaustion probability first stays 0, and then monotonically increases to 1 when the capacity is reached and no more user can be admitted. The monotonic trend indicates the dominative role that the small cell depletion probability plays in the bandwidth utilization of the MBS. When $\epsilon$ is very small, although a fewer number of users can be admitted to the small cells, the macrocell users do not used up the bandwidth. When $\epsilon$ becomes large, although more users can be admitted in the small cells, small cell energy depletion occurs more as well due to the loose admission control criteria. Depleted SBSs will results in more users associated with the MBS and thus more bandwidth is needed at the MBS. As a result, the bandwidth exhaustion probability increases. When bandwidth is exhausted, the average transmission power of the MBS remains unchanged when the network capacity is reached.

A more intuitive illustration for bandwidth exhaustion can be found in Fig. 8, which exhibits in one simulation run the total required bandwidth from all the macrocell users as time changes. It can be seen that when $\epsilon = 0.001$, the small cell depletion probability is very small, and the impulse caused by the depletion only occurs occasionally. The bandwidth exhaustion does not happen in the whole horizon. When $\epsilon$ increases, the depletion probability increases, resulting in that more than 1 small cells deplete their energy at the same time. Consequently, bandwidth

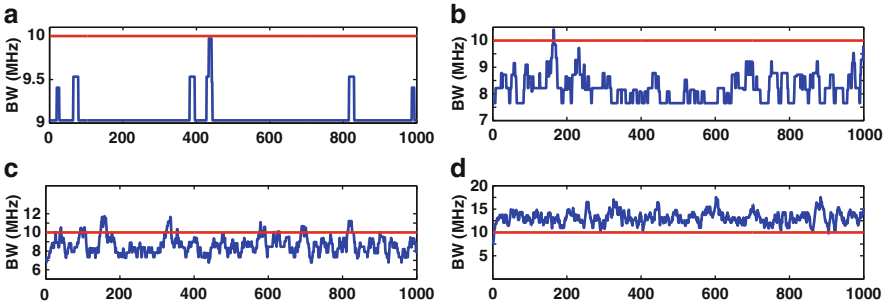


**Fig. 8** The bandwidth exhaustion conditions over time. The notation *BW* denotes the aggregate required bandwidth from all macrocell users in each time slot. (**a**) $\epsilon = 0.001$. (**b**) $\epsilon = 0.01$. (**c**) $\epsilon = 0.1$. (**d**) $\epsilon = 0.2$

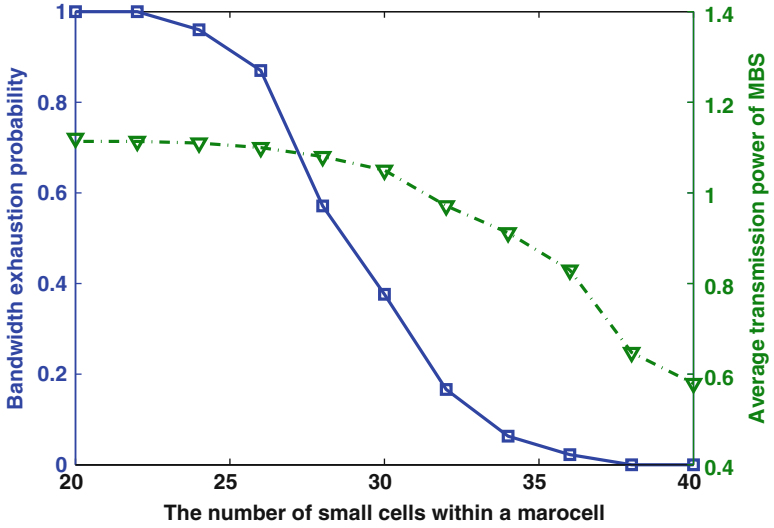


**Fig. 9** The bandwidth exhaustion probability and average transmission power of the MBS vs. the small cell density. Default values: $\mu_c = 1.9$, $\sigma_c^2 = 0.89$, $\mu_0 = 20$, $\sigma_0^2 = 400$, and $\epsilon = 0.1$

exhaustion may occur occasionally. When $\epsilon = 0.1$, bandwidth exhaustion becomes more frequent due to s higher depletion probability. When $\epsilon$ is large (e.g., 0.2), bandwidth is exhausted in most of the time slots except the very beginning. Therefore, it can be concluded from Figs. 7 and 8 that optimal admission control criteria $\epsilon$ exists for small cells in striking the tradeoff between small cell sustainability and macrocell resource utilization.

Figure 9 presents how the bandwidth exhaustion probability and average transmission power of the MBS are affected by the small cell density when the admission control criteria $\epsilon$ is fixed. It can be seen that with increased number of small cells in one macrocell, both the bandwidth exhaustion probability and the average MBS transmission power decrease. The results are intuitive since given $\epsilon$, more small cells can accommodate more users and reduce the number of macrocell users without increasing the depletion probability. Consequently, more MBS transmission power is saved and the bandwidth exhaustion probability is brought down.

Figure 10 shows how the bandwidth exhaustion probability and average transmission power of the MBS change with the average initial energy level of the SBSs. In each simulation run, the initial energy of each SBS is uniformly and randomly chosen from the interval $[e_0 - 3, e_0 + 3]$. The average value is obtained over 1000 runs. It can be observed that starting from the average value of 9, both examined metrics decrease as the average initial energy level of the SBSs increase. The reason is similar with that for Fig. 9, i.e., higher average initial energy level allows one SBS to accommodate more users and reduce the number of users connected to the MBS. In addition, one interesting observation may be noted that a trough occurs when the average initial energy level increases from 5 to 9 energy units. This phenomenon is

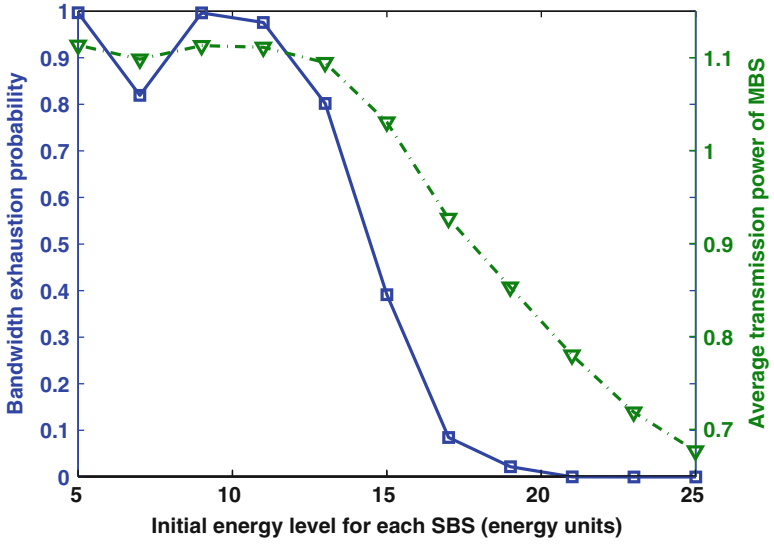


**Fig. 10** The bandwidth exhaustion probability and average transmission power of the MBS vs. the average initial energy level $e_0$ of the SBSs. Default values: $\mu_c = 1.9$, $\sigma_c^2 = 0.89$, $\mu_0 = 20$, $\sigma_0^2 = 400$, $\epsilon = 0.1$, and $N_s = 30$

counter-intuitive at first glance but reasonable after a second thought. When $e_0 = 5$ and $e_0 = 7$, the number of users that can be admitted into a small cell satisfying $\epsilon = 0.1$ is the same, but the number of admitted users gradually increases when $e_0 = 9$ and $e_0 = 11$, etc. This is because when the initial energy level is low, the energy depletion probability is relatively high, and users should be admitted conservatively. When the initial energy is abundant, more users can be admitted. This explains why the number of admitted users do not change when $e_0$ increases from 5 to 7, but changes afterwards when $e_0$ is large. We also validate the number of admitted users in the simulations by checking the intermediate results. Given the same number of admitted users, the SBS depletion probability when $e_0 = 7$ is considerably smaller than that when $e_0 = 5$ according to Eq. (15). As a result, the SBSs when $e_0 = 7$ are less likely to be depleted, leading to a smaller average number of macrocell users. When $e_0 = 9$, the number of users that can be admitted into a small cell increases (i.e., the discharging rate is increased), resulting into a larger SBS depletion probability and larger average number of macrocell users.

Figure 11 illustrates how the average small cell throughput changes with different $\epsilon$ values when $\mu_0 = 20$ and $\mu_0 = 21$, respectively. Different $\mu_0$ values indicate different energy discharging rates from the energy buffer. It can be found that for either curve, the average throughput of small cells first increases slightly and then drops dramatically as $\epsilon$ increases. The reason is as follows. There are two major factors that affect the average small cell throughput: SBS depletion probability and $\epsilon$. When $\epsilon$ plays a dominant role, i.e., the performance gain of more admitted users

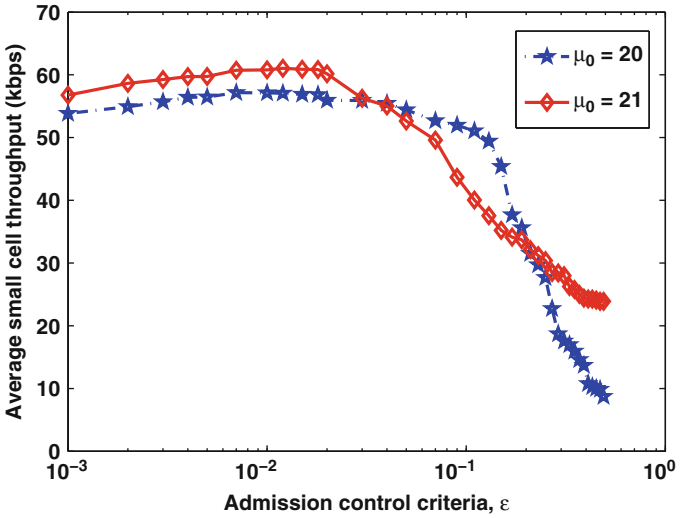


**Fig. 11** The impact of admission control criteria $\epsilon$ on the average small cell throughput. Default values: $\mu_c = 1.9$, $\sigma_c^2 = 0.89$, and $N_s = 30$

due to looser admission control criteria surpasses the performance loss due to higher depletion probability, the average small cell throughput increases. The situation is reverse when the negative impacts of energy depletion becomes the dominant factor that surpasses the performance gain of more admitted users. In addition, the dominant factors vary at different charging rates and admission control criteria, and we observe that there are two cross points between the two curves in Fig. 11.

## 7 Conclusion

In this chapter, we have developed a generic analytical framework to study the energy sustainability of a green HetNet powered by renewable energy sources, by characterizing the dynamics in the energy charging and discharging processes. We have derived the closed-form solutions of the energy buffer at SBSs, i.e., the energy depletion probability and average energy depletion time. Based on the analysis, we have further proposed a distributed admission control scheme to strike the balance between high resource utilization of the green HetNet and the energy sustainability of SBSs, by effectively controlling the volume of the offloaded traffic from the macrocell to small cells.

In our future work, we plan to jointly consider bandwidth allocation and power management in a green HetNet powered by sustainable energy sources, considering the inter- and intra-tier interference among different cells, and finite energy buffers

installed at BSs. How to optimally deploy the minimal number of green BSs to ensure network coverage and fully sustainable network operation of a green HetNet will also be under investigation.

## References


1. Cisco White Paper, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019 (2015)
2. R. Zhang, M. Wang, L.X. Cai, Z. Zheng, X. Shen, L.L. Xie, LTE-unlicensed: The future of spectrum aggregation for cellular networks. IEEE Wirel. Commun. Mag., **22**(3), 150–159 (2015).
3. OPERA-Net I/II, Optimising Power Efficiency in Mobile Radio Networks (2015). Available: http://projects.celticplus.eu/opera-net2/index.html
4. Mobile VCE Project, Green radio (2013). Available: http://www.mobilevce.com/green-radio
5. L.X. Cai, L. Cai, X. Shen, J.W. Mark, Resource management and QoS provisioning for IPTV over mmWave-based WPANs with directional antenna. Mob. Netw. Appl. **14**(2), 210–219 (2009)
6. P. Grant, S. Fletcher, Mobile basestations: reducing energy. Eng. Technol. Mag. **6**(2), (2011) x
7. S. Navaratnarajah, A. Saeed, M. Dianati, M.A. Imran, Energy efficiency in heterogeneous wireless access networks. IEEE Wirel. Commun. Mag. **20**(5), 37–43 (2013)
8. J.B. Rao, A.O. Fapojuwo, A survey of energy efficient resource management techniques for multicell cellular networks. IEEE Commun. Surv. Tutorials **16**(1), First Quarter (2014)
9. R. Zhang, M. Wang, Z. Zheng, X. Shen, L.L. Xie, Cross-layer carrier selection and power control for LTE-A uplink with carrier aggregation, in *Proceedings of IEEE Globecom'13* (2013)
10. S. Mclaughlin, P.M. Grant, J.S. Thompson, H. Haas, D.I. Laurenson, C. Khirallah, Y. Hou, R. Wang, Techniques for improving cellular radio base station energy efficiency. IEEE Wirel. Commun. Mag. **18**(5), 10–17 (2011)
11. X. Xiao, X. Tao, Y. Jia, J. Lu, An energy-efficient hybrid structure with resource allocation in OFDMA networks, in *Proceedings of IEEE WCNC'11* (2011)
12. R. Xie, F.R. Yu, H. Ji, Y. Li, Energy-efficient resource allocation for heterogeneous cognitive radio networks with femtocells. IEEE Trans. Wirel. Commun. **11**(11), 3910–3920 (2012)
13. G. Piro, M. Miozzo, G. Forte, N. Baldo, L.A. Grieco, G. Boggia, P. Dini, HetNets powered by renewable energy sources: sustainable next-generation cellular networks. IEEE Internet Comput. **17**(1), 32–39 (2013)
14. H.S. Dhillon, Y. Li, P. Nuggehalli, Z. Pi, J.G. Andrews, Fundamentals of heterogeneous cellular networks with energy harvesting. IEEE Trans. Wirel. Commun. **13**(5), 2782–2797 (2014)
15. Z. Zheng, X. Zhang, L.X. Cai, R. Zhang, X. Shen, Sustainable communication and networking in two-tier green cellular networks. IEEE Wirel. Commun. Mag. **21**(4), 47–53 (2014)
16. T. Edler, S. Lundberg, Energy efficiency enhancements in radio access networks, in *Ericsson Review* (2004). Available: http://www.ericsson.com/ericsson/corpinfo/publications/review/2004_01/files/2004015.pdf
17. X. Zhang, Z. Su, Z. Yan, W. Wang, Energy-efficiency study for two-tier heterogeneous networks (HetNets) under coverage performance constraints. Mob. Netw. Appl. **18**(4), 567–577 (2013)
18. H. Klessig, A.J. Fehske, G.P. Fettweis, Energy efficiency gains in interference-limited heterogeneous cellular mobile radio networks with random micro site deployment, in *IEEE 34th Sarnoff Symposium* (2011), pp. 1–6
19. L. Saker, S.E. Elayoubi, R. Combes, T. Chahed, Optimal control of wake up mechanisms of femtocells in heterogeneous networks. IEEE J. Sel. Areas Commun. **30**(3), 664–672 (2012)

20. W. Zheng, W. Li, Y. Xie, X. Wen, Hybrid BS-cooperative power management scheme with self-organized sleep mode in virtual cell-based femto networks. Int. J. Distrib. Sens. Netw. **2012**, 12 pp. (2012)
21. D. Cao, S. Zhou, Z. Niu, Improving the energy efficiency of two-tier heterogeneous cellular networks through partial spectrum reuse. IEEE Trans. Wirel. Commun. **12**(8), 4129–4141 (2013)
22. R. Zhang, M. Wang, Z. Zheng, X. Shen, L.L. Xie, Stochastic geometric performance analysis for carrier aggregation in LTE-A systems, in *Proceedings of IEEE ICC'14* (2014)
23. M. Lauridsen, H. Wang, P. Mogensen, LTE UE energy saving by applying carrier aggregation in a HetNet scenario, in *Proceedings of IEEE VTC Spring'13* (2013)
24. A. Mesodiakaki, F. Adelantado, L. Alonso, C. Verikoukis, Energy-efficient user association in cognitive heterogeneous networks. IEEE Commun. Mag. **52**(7), 22–29 (2014)
25. R. Zhang, Z. Zheng, M. Wang, X. Shen, L.L. Xie, Equivalent capacity in carrier aggregation-based LTE-A systems: a probabilistic analysis. IEEE Trans. Wirel. Commun. **13**(11), 6444–6460 (2014)
26. Z. Xu, C. Yang, G.Y. Li, Y. Liu, S. Xu, Energy-efficient CoMP precoding in heterogeneous networks. IEEE Trans. Signal Process. **62**(4), 1005–1017 (2014)
27. K.M.S. Huq, S. Mumtaz, J. Bachmatiuk, J. Rodriguez, X. Wang, R.L. Aguiar, Green HetNet CoMP: energy efficiency analysis and optimization. IEEE Trans. Veh. Technol. **64**, 4670–4683 (2014)
28. J. Yang, S. Ulukus, Optimal packet scheduling in an energy harvesting communication systems. IEEE Trans. Commun. **60**(1), 220–230 (2012)
29. K. Tutuncuoglu, A. Yener, Optimum transmission policies for battery limited energy harvesting nodes. IEEE Trans. Wirel. Commun. **11**(3), 1180–1189 (2012)
30. O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, A. Yener, Transmission with energy harvesting nodes in fading wireless channels: optimal policies. IEEE J. Sel. Areas Commun. **29**(8), 1732–1743 (2011)
31. X. Zhang, X. Shen, L.L. Xie, Joint subcarrier and power allocation for cooperative communications in LTE-advanced networks. IEEE Trans. Wirel. Commun. **13**(2), 658–668 (2014)
32. J. Taneja, J. Jeong, D. Culler, Design, modeling, and capacity planning for micro-solar power sensor networks, *Proceedings of IEEE ICPN'08* (2008)
33. A. Sayegh, T.D. Todd, M.N. Smadi, Resource allocation and cost in hybrid solar/wind powered WLAN mesh nodes, in *Wireless Mesh Networks: Architectures and Protocols* (Springer, New York, 2008), pp. 167–189
34. A. Farbod, T.D. Todd, Resource Allocation and Outage Control for Solar-Powered WLAN Mesh Networks. IEEE Trans. Mob. Comput. **6**(8), 960–970 (2007)
35. Z. Zheng, L.X. Cai, M. Dong, X. Shen, H.V. Poor, Constrained AP placement with rate adaptation in green WLAN mesh networks, in *Proceedings of IEEE GLOBECOM'11* (2011)
36. T. Yang, Z. Zheng, H. Liang, R. Deng, N. Cheng, X. Shen, Green energy and content aware data transmission in maritime wireless communication network. IEEE Trans. Intell. Transp. Syst. **16**(2), 751–762 (2015)
37. L.X. Cai, Y. Liu, T.H. Luan, X. Shen, J.W. Mark, H.V. Poor, Dimensioning network deployment and resource management in green mesh networks. IEEE Wirel. Commun. Mag. **18**(5), 58–65 (2011)
38. E. Lattanzi, E. Regini, A. Acquaviva, A. Bogliolo, Energetic sustainability of routing algorithms for energy-harvesting wireless sensor networks. Comput. Commun. **30**, 2976–2986 (2007). Available: http://dx.doi.org/10.1016/j.comcom.2007.05.035
39. L.X. Cai, Y. Liu, T.H. Luan, X. Shen, J.W. Mark, H.V. Poor, Sustainability analysis and resource management for wireless mesh networks with renewable energy supplies. IEEE J. Sel. Areas Commun. **32**(2), 345–355 (2014)
40. K. Huang, V.K.N. Lau, Enabling wireless power transfer in cellular networks: architecture, modelling and deployment. IEEE Trans. Wirel. Commun. **13**, 902–912 (2014)
41. K. Huang, V.O.K. Li, Renewables powered cellular networks: energy field and network coverage. IEEE Trans. Wirel. Commun., to appear

42. Z. Zheng, L.X. Cai, R. Zhang, X. Shen, H.V. Poor, RNP-SA: joint relay placement and subcarrier allocation in wireless communication networks with sustainable energy. IEEE Trans. Wirel. Commun. **11**(10), 3818–3828 (2012)
43. S. Wang, W. Guo, M.D. McDonnell, Downlink interference estimation without feedback for heterogeneous network interference avoidance, in *Proc. IEEE ICT'14* (2014)
44. D.R. Cox, H.D. Miller, *The Theory of Stochastic Processes* (Chapman and Hall, London, 1965)
45. H. Kobayashi, Application of the diffusion approximation to queueing networks I: equilibrium queue distributions. J. ACM **21**(2), 316–328 (1974)

# Resource Allocation for Cooperative D2D Communication Networks

**Shankhanaad Mallick, Roya Arab Loodaricheh, K.N.R. Surya Vara Prasad, and Vijay Bhargava**

**Abstract** Device-to-device (D2D) communications technology is currently being investigated as a potential enabler for the fifth generation (5G) communication networks. Significant performance gains are achievable in a cooperative D2D framework, wherein the user equipments (UEs) cooperate with each other to enable a variety of low-latency proximity-based services or to establish indirect communication links with the Base Station (BS) whenever direct service coverage is not possible. This chapter is focused on the throughput gains achievable in the latter scenario, i.e., when few UEs perform relaying operations to provide indirect service coverage to other UEs. In this direction, resource allocation problems are formulated for a variety of system models operating under the orthogonal frequency division multiple access (OFDMA) cellular or cognitive radio (CR) access architectures. The performance of mobile D2D relaying under different scenarios is evaluated. The system models are designed to study the benefits of incorporating additional capabilities at the devices, such as packet storage (using buffers), energy-harvesting, and cognitive spectrum access within the cooperative D2D framework. Depending on the system model, efficient algorithms are proposed to obtain optimal power allocation, subcarrier assignment, subcarrier pairing, and relay-UE selection policies which maximize the system throughput under a variety of system-dependant constraints. Simulation results demonstrate the effectiveness of our proposed algorithms and the performance improvement of mobile D2D-relaying networks over conventional networks.

## 1 Introduction

Device-to-device (D2D) communication is a radio technology which allows nearby devices (such as mobile phones and computers) to establish local communication links with each other. The local communication links, established either on the licensed or the unlicensed spectrum, allow traffic to flow directly between them without traversing the conventional path via the base stations (BSs) [1]. Consider,

S. Mallick (✉) • R. Arab Loodaricheh • K.N.R. Surya Vara Prasad • V. Bhargava
University of British Columbia, Vancouver, BC, Canada
e-mail: shankha@ece.ubc.ca; royaa@ece.ubc.ca; surya@ece.ubc.ca; vijayb@ece.ubc.ca



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_20

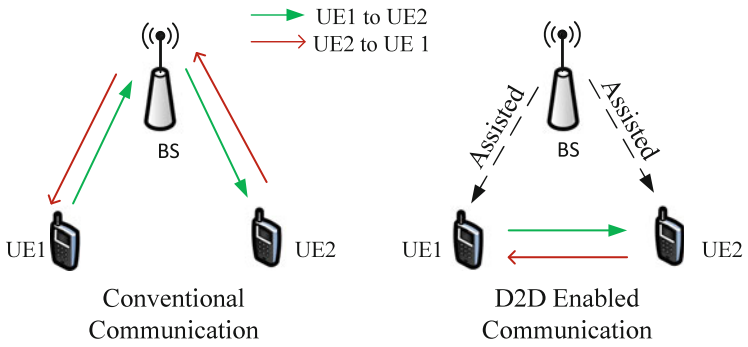


**Fig. 1** In conventional systems, UE1 and UE2 communicate via the BS, irrespective of their proximity to each other. However, in D2D enabled communication, network latencies and throughput rates can be significantly improved by establishing local communications with the help of signalling assistance from the BS

for example, a scenario where two user equipments (UEs), located within close proximity to each other, should share gigabytes of data on a regular basis through a wireless medium. If D2D communications are enabled, as shown in Fig. 1, the BS coordinating these devices is no longer a traffic bottleneck between the source and the destination.

D2D communications can potentially improve the end-user experience by reducing network latency, reducing power consumption, increasing peak data rates, and facilitating several proximity-based services [2]. Moreover, since D2D is a short-distance based communication technology, sufficiently separated D2D links can simultaneously operate on the same frequency band, thereby facilitating dense spectral reuse across the network. Due to these added benefits, D2D has been specified by 3GPP in LTE Release 12, as a technology with focus on the applications of proximity-based services and public safety communication [3, 4].

In proximity-based services, devices detect their proximity and subsequently trigger a variety of services including, for example, mobile social networking, mobile marketing, proximal multi-player gaming, advertising, local exchange of information, and smart vehicle-to-vehicle communications [5]. In public safety support, D2D can provide local connectivity to the UEs when the central radio infrastructure is damaged. Therefore, UEs can benefit significantly from a number of public safety services with the help of D2D technology. However, D2D brings a unique set of research challenges to the network designer. The most important research challenges include the design of smart devices, information security, interference management, and mobility management.

### *1.1 Cooperative D2D as a Technology Enabler for 5G*

D2D communication is also being considered as one of the state-of-the-art technologies for the 5G network architecture [6, 7]. This is because D2D-enabled 5G networks can support a plethora of services such as, cellular network offloading, low-latency tactile applications, vehicle-to-vehicle communication, national security and service advertisement. Recent research shows that D2D communications within the shared cellular spectrum controlled by cellular infrastructure achieve the following potential gains:

1. Spectral or network capacity gain: by sharing radio resources efficiently between conventional and D2D-enabled UEs.
2. User data rate gain: high peak rates may be achieved when UEs are close to each other with strong propagation links between them.
3. Latency gain: direct communication between UEs bypassing the network infrastructure reduces the end-to-end latency.

Additionally, D2D connectivity makes it possible for mobile devices in the multi-tier 5G network architecture to function as transmission relays for each other. Such cooperative mobile relaying methods have recently attracted significant attention because they can greatly improve the network performance without deploying new BSs [8]. By allowing relay-based communication, it is possible to improve network coverage and quality of service and to achieve similar performance benefits as in multiple-input-multiple-output (MIMO) systems. The idea of relaying is particularly attractive for the cell-edge users and for the users in coverage holes. In conventional relay systems, fixed relay stations are installed on a per-demand basis. Fixed relay stations not only result in significant infrastructure and maintenance costs for the operators, but also consume significant amounts of power. In this regard, mobile UEs with strong communication links to the BS can help as relays, via D2D communication, to other UEs with heavy blockage and shadowing at no additional cost to the mobile operators. In other words, mobile relaying via D2D communication can enhance the performance of cellular networks by taking the advantage of "shadowing diversity".

Consider the D2D relaying example illustrated in Fig. 2. The BS has to communicate some information with the destination UE. However, due to network blockage, direct communication with the destination UE cannot be established. The Relay UE, which is closer to the BS and has strong communication links with it, can act as a relay for the destination UE to enable indirect communication between the BS and the destination UE. In other words, if the cellular link of a mobile UE is weak, strong D2D links can establish a cellular connection via relay-UEs. Additionally, as opposed to the high power consumption in fixed relay stations, D2D relays operate at significantly lower power consumption levels and can therefore, potentially act as an energy-efficient replacement for the fixed relay stations.

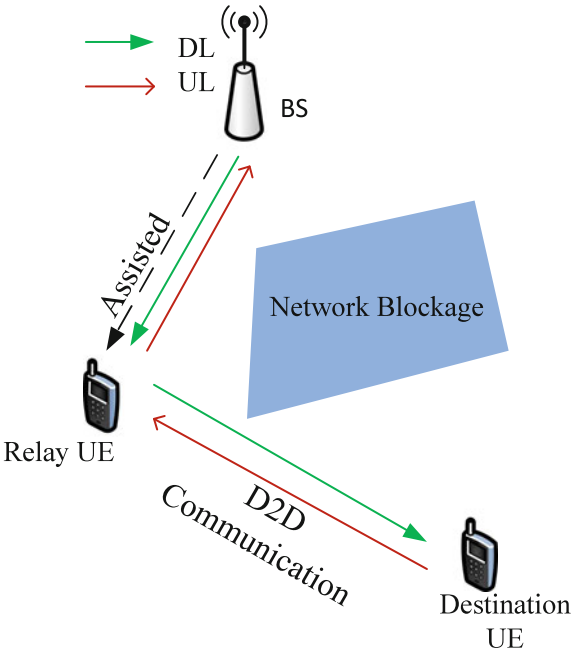


**Fig. 2** System model for cellular wireless systems with cooperative D2D communication. UEs with strong communication links from the BS can send/receive data directly from the BS. However, UEs with bad communications links from the BS (for example, due to network blockage) can benefit from D2D communication with relay UEs to send/receive the data from the BS

## *1.2 Chapter Organization*

The concept of D2D relaying has motivated researchers to investigate design methods which jointly exploit the advantages of both D2D and relaying technologies. Prominent research areas include opportunistic relay-UE selection, relay-UE power allocation, interference management, minimization of relay battery power, and maximizing network energy efficiency, improving coverage, and so on. In this chapter, different resource allocation problems associated with cooperative D2D networks operating under a variety of architectural settings are discussed, including OFDMA and cognitive radio (CR) networks. In Sect. 2, resource allocation problems for downlink transmission in OFDMA-based buffer-aided D2D relaying networks are investigated. Efficient algorithms are developed to solve the problem of power and subcarrier allocation such that the system capacity is maximized. In Sect. 3, energy-harvesting technology is considered in OFDMA systems where the UEs are capable of harvesting energy using the "wireless power transfer (WPT)" technology [9] and only the harvested energy is used for cooperative D2D-relaying. The joint optimization problem of power allocation, subcarrier allocation and relay-UE selection is studied. In Sect. 4, a cognitive radio network (CRN) is considered,

where the secondary users (SUs) from different small cell primary user (PU) networks attempt communication with each other. Cooperative D2D-relaying is utilized to enable communication among the SUs. Efficient relay-CR selection and power allocation schemes are developed for realistic networks under channel uncertainty. Finally, in Sect. 5, we provide few concluding remarks and directions for future work.

## 2 Resource Allocation for OFDMA-Based Buffer-Aided D2D Relaying Networks with Frequency Reuse

In this section, we consider the scenario illustrated earlier in Fig. 2, in which BS cannot communicate with the destination UE due to heavy obstructions or shadowing. Our goal is to establish indirect communication links between the BS and the UEs which are blocked from direct communication with the BS. In this regard, some other UEs with fair connection to the BS can assist as relays. As discussed earlier in Sect. 1.1, D2D relaying enhances the system performance by taking advantage of "shadowing diversity".

Specifically, we investigate D2D relaying for downlink transmission in OFDMA cellular networks. To analyze the performance gains achievable via D2D-relaying in this OFDMA system, we first classify the D2D pairs[1] based on the level of proximity with each other. Firstly, we consider a scenario where all the D2D pairs are very close to each other and their coverage areas overlap. In this case, since reusing the subcarriers among the D2D pairs will introduce significant interference to the D2D links, we do not attempt subcarrier reuse among the D2D pairs. Next, we consider a scenario where the D2D pairs are far enough from each other, and therefore their coverage areas are disjoint. In this case, full subcarrier reuse can be attempted with the help of buffer-aided relaying in all of the D2D pairs. We discuss relevant resource allocation problems and analyze the throughput gains achievable in both of the above scenarios. Later, we provide a general motivation on how frequency reuse can be attempted for a general scenario where D2D pairs are randomly distributed in a cell, i.e, the two aforementioned scenarios co-exist simultaneously within the cell.

### *2.1 System Model*

Let us assume that all the mobile UEs which can directly communicate with the BS are interested in relaying. However we opportunistically select few UEs for relaying based on the performance gains they can offer to the system. The important fact

[1]We refer to a relay-UE and destination-UE combination as a D2D pair if a D2D communication link can be established between them.

in the D2D communication is that the UEs can only communicate within short distances. Therefore, the relay-UEs (also referred to as relays interchangeably) and the UEs in coverage holes (also referred to as destination-UEs, or d-UEs, interchangeably) must be in the coverage area of each other. Also, to make the problem general, we assume that the d-UEs can receive assistance from multiple relays. Let $N$ orthogonal subcarriers, each with bandwidth $W$ be available for resource allocation in the system. Let $K$ be the number of d-UEs, i.e., UEs requiring relay assistance. Further, let us denote the BS, relay-UEs and d-UEs as B, R, and U, respectively. The set of the relays which can assist the $k$th UE is denoted by $R_k$. The channel gain and the transmission power from the BS to relay $m_k \in R_k$ over subcarrier $i$ at time slot $t$ are denoted by $g^{B,R}_{k,m_k,i,t}$ and $P^{B,R}_{k,m_k,i,t}$, respectively. Similarly, the channel gain from relay $m_k \in R_k$ to the $k$th UE over subcarrier $j$ at time slot $t$ is denoted by $g^{R,U}_{k,m_k,j,t}$. The transmission power of any relay in the candidate set, i.e, $R_k$, $\forall k$, over any subcarrier, i.e, $\forall i \in \{1, 2, \ldots N\}$, is the same fixed value denoted by $P_R$. We implement the popular amplify-and-forward (AF) protocol [10] for relaying.

## 2.2 D2D Relaying Without Frequency Reuse

We start with the first case in which all the D2D pairs are close to each other and their coverage areas overlap. As a result, frequency reuse may introduce excessive interference into the system. Note that, due to indirect communication via mobile relaying, downlink transmission from the BS to a given d-UE happens over two-time-slot intervals. During the first time slot, the BS transmits information to a relay-UE. This information is relayed by the relay-UE to the d-UE during the second time-slot. We also assume that each relay-UE has a unit buffer size, i.e, if all the BS-to-relay transmissions occur in the $t$th time slot, all the relay-to-d-UE transmissions occur in the $(t + 1)$th time sot. Under the AF relaying protocol, the achievable throughput $R_{k,m_k,i}$ for downlink transmission from the BS to the $k$th UE via relay $m_k$ over subcarrier $i$ is given by

$$R_{k,m_k,i} = \frac{1}{2}\log_2\left(1 + \frac{P^{B,R}_{k,m_k,i} P_R g^{B,R}_{k,m_k,i} g^{R,U}_{k,m_k,i}}{P^{B,R}_{k,m_k,i} g^{B,R}_{k,m_k,i} + P_R g^{R,U}_{k,m_k,i}} \cdot \frac{1}{N_0 W}\right), \tag{1}$$

where $N_0$ is the noise variance assuming additive white Gaussian noise (AWGN) channels. In this orthogonal transmission case, the time index is omitted, i.e, $g^{B,R}_{k,m_k,i}$ is used in place of $g^{B,R}_{k,m_k,i,t}$, because we assume the unit buffer size at all the relays. Using (1), we formulate the following resource allocation problem to maximize the sum throughput in the system

$$
\begin{aligned}
&\underset{P,S}{\text{maximize}} \quad \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} s_{k,m_k,i} R_{k,m_k,i} \\
&\text{subject to :} \\
&C_1 : \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} s_{k,m_k,i} P^{B,R}_{k,m_k,i} \leq P_{max}, \\
&C_2 : s_{k,m_k,i} \in \{0,1\}, \quad \forall i,k,m_k, \\
&C_3 : \sum_{k=1}^{K} \sum_{m_k \in R_k} s_{k,m_k,i} \leq 1, \quad \forall i, \\
&C_4 : P^{B,R}_{k,m_k,i} \geq 0, \quad \forall i,k,m_k,
\end{aligned}
\tag{2}
$$

where $P = \{P^{B,R}_{k,m_k,i}\}$ and $S = \{s_{k,m_k,i}\}$ denote the power allocation and subcarrier allocation policies, respectively.

$C_1$ is the power budget constraint with $P_{max}$ being the power budget for the BS. $C_2$ shows that the subcarrier assignment indicators $s_{k,m_k,i}$ are binary integer variables. $s_{k,m_k,i} = 1$ if BS is transmitting to the $k$th UE via relay $m_k$ over subcarrier $i$ and $s_{k,m_k,i} = 0$ otherwise. $C_3$ ensures that the subcarriers are not reused among the D2D pairs to avoid interference. $C_4$ ensures that the power allocation variables are non-negative.

In simple terms, our goal in (2) is to allocate the BS transmission power and the $N$ subcarriers such that the sum-throughput in the system is maximized. From an optimization perspective, the problem in (2) is a mixed integer non-linear program (MINLP) which, in its original form, is difficult to solve in polynomial time due to the coupled integer and continuous variables [11–13]. Therefore, to obtain a real-time solution, we attempt a series of mathematical modifications and simplifications as shown in Fig. 3.

Firstly, we relax the integer binary variables, i.e., the subcarrier assignment indicators $s_{k,m_k,i}$ into continuous variables in the interval [0,1]. Using this relaxation, we introduce auxiliary power variables as

$$
\tilde{P}^{B,R}_{k,m_k,i} = s_{k,m_k,i} P^{B,R}_{k,m_k,i}. \tag{3}
$$

When the original power variables in (2) are replaced by the auxiliary power variables defined in (3), the optimization problem (2) is relaxed into the following convex optimization problem

$$
\min_{\lambda \geq 0} \max_{\tilde{P},S} L(\lambda, \tilde{P}, S), \tag{4}
$$

where $L$ is a Lagrangian function and $\lambda$ is the associated Lagrangian multiplier. Therefore, efficient convex optimization algorithms can be used to solve the relaxed problem. Consider, for example, the popular dual decomposition technique wherein a master problem and few sub-problems are solved iteratively until convergence is achieved. The dual decomposition technique can be adopted to solve (4) by allowing the master problem to update $\lambda$s and the sub-problems to find optimal power and subcarrier assignment for a given $\lambda$.

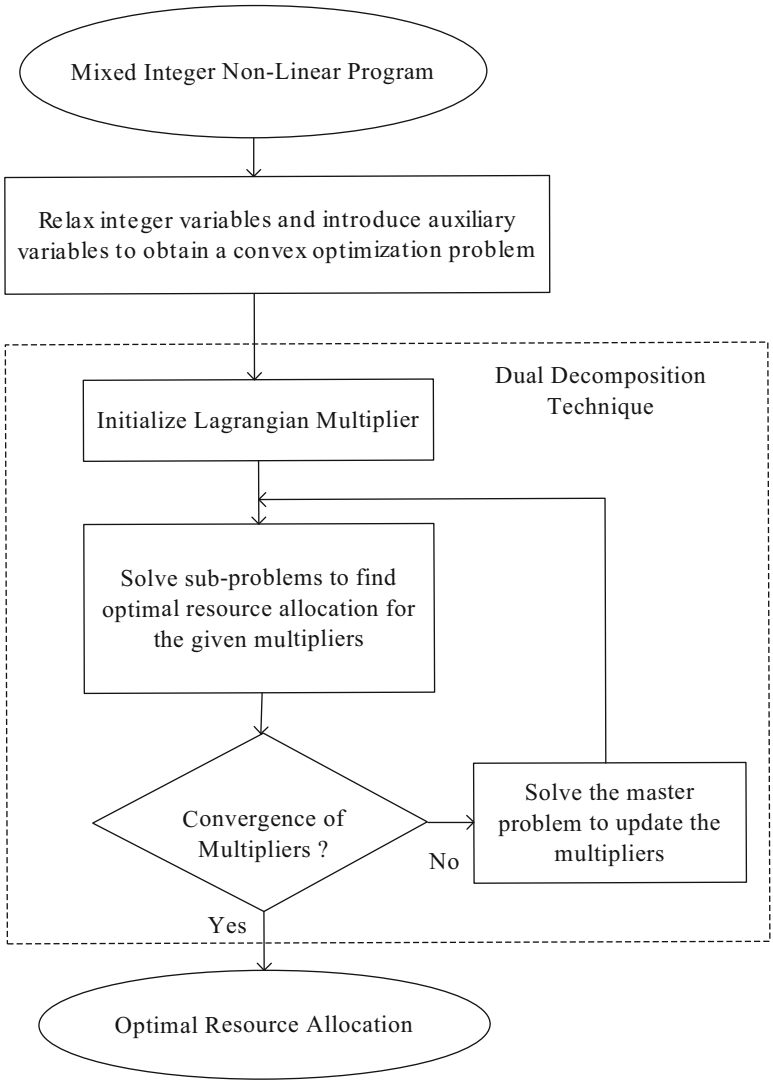


**Fig. 3** Optimization framework to solve (2)

We begin with the following definition of the Lagrangian function

$$L(\lambda, \tilde{P}, S) = \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} s_{k,m_k,i} \frac{1}{2} \log_2 \left( 1 + \frac{\frac{\tilde{P}^{B,R}_{k,m_k,i} g^{B,R}_{k,m_k,i}}{s_{k,m_k,i} i N_0 W} P_R g^{R,U}_{k,m_k,i}}{\frac{\tilde{P}^{B,R}_{k,m_k,i} g^{B,R}_{k,m_k,i}}{s_{k,m_k,i}} + P_R g^{R,U}_{k,m_k,i}} \right) + \lambda \left( P_{max} - \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} \tilde{P}^{B,R}_{k,m_k,i} \right), \tag{5}$$

where we consider the power constraint $C_1$ only and initialize the Lagrangian multiplier $\lambda$ with a non-negative value. Constraints $C_2, C_3, C_4$ are not considered in (5) but will be satisfied in subsequent steps. By applying KKT conditions to (5), we can derive optimal auxiliary power variables $\tilde{P^*}^{B,R}_{k,m_k,i}$ and the corresponding throughput $\tilde{R^*}_{k,m_k,i}$ in terms of the subcarrier assignment indicators $s_{k,m_k,i}$ and the Lagrangian multiplier $\lambda$. In the derivation process, constraint $C_4$ can be satisfied by offsetting negative $\tilde{P^*}^{B,R}_{k,m_k,i}$ values to zero. Plugging back $\tilde{P^*}^{B,R}_{k,m_k,i}$ and $\tilde{R^*}_{k,m_k,i}$ into the sub-problem (5) and re-arranging the equations, we obtain the simplified problem

$$\begin{aligned} &\underset{S}{\text{maximize}} \quad \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} s_{k,m_k,i} H_{k,m_k,i} + \lambda P_{max} \\ &\text{subject to} \ : \ C_3, \end{aligned} \tag{6}$$

where we have

$$H_{k,m_k,i} = \tilde{R^*}_{k,m_k,i} - \lambda \tilde{P^*}^{B,R}_{k,m_k,i}. \tag{7}$$

Upon inspection, one may observe that (6) is a linear programming problem [14] with respect to $s_{k,m_k,i}$, i.e., it is a linear subcarrier assignment problem. Applying totally unimodularity theorem [15, 16] shows that the problem (6) has integer optimal values even if the optimization variables $S = \{s_{k,m_k,i}\}$ span over the continuous interval [0,1]. Optimal solution to the linear subcarrier assignment problem (6) can be obtained using simple algorithms such as Algorithm 1.

**Algorithm 1** Subcarrier allocation algorithm I

1: Find the best user-relay pair for each subcarrier as $(k^*, m_k^*) = \underset{k,m_k}{\text{maximize}} \ H_{k,m_k,i}, \forall i.$
2: **return** $S^*$

Note that, constraint $C_3$ can be satisfied while implementing Algorithm 1. Also, since Algorithm 1 yields optimal solutions, constraint $C_2$ is satisfied inherently due to totally unimodularity theorem. Since the optimal power allocation and subcarrier assignment values are now available, the master problem of updating the Lagrange multiplier $\lambda$ can be solved using standard sub-gradient methods [17] as follows

$$\lambda(l+1) = \left[\lambda(l) - \eta\left(P_{max} - \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} \tilde{P}^{B,R}_{k,m_k,i}\right)\right]^+, \tag{8}$$

where $\eta$ is a constant step size parameter, $l$ is the iteration index, and $[x]^+ = \max(0, x)$.

The master problem and the sub-problems are solved iteratively until we observe a convergence in the Lagrangian multiplier, i.e, $\lambda$, values. The complexity of the proposed algorithm is polynomial in time and is significantly lower than the brute-force exhaustive search method.

### *2.3 Buffer-Aided D2D Relaying with Frequency Reuse*

We now consider the second scenario wherein the D2D pairs are far enough from each other such that their coverage areas do not overlap. Therefore, frequency reuse can now be attempted to fully utilize the available spectrum. The resource allocation scheme proposed in the previous section, i.e, for orthogonal transmission, may not be spectrally efficient for this system because frequency reuse was not considered. Note that, frequency reuse is not possible in the first time slot, as the BS is broadcasting the messages. However, in the second time slot, frequency reuse can be attempted to establish D2D communication between relay-UEs and d-UEs. To exploit the benefits of frequency reuse, we propose a communication protocol using buffer-aided relaying.

As a motivational example, let us consider a system with two D2D pairs and two subcarriers. During the first time-slot, the BS broadcasts two packets, i.e, one to each relay-UE, over the two subcarriers. During the second time-slot, the BS broadcasts another two packets over the two subcarriers. Therefore, at the end of the second time-slot, we have four packets at the relay-UE buffers, i.e, two packets at each relay-UE. Since the coverage areas of the two D2D pairs do not overlap, both the relay-UEs can use the two subcarriers simultaneously for D2D communications. That is, during the third time slot, each relay-UE can forward its two packets over the two subcarriers without causing any interference to other relay-UEs. Therefore, using buffer-aided relaying with full frequency reuse, four packets are transmitted to two d-UEs in three time slots. On the other hand, the orthogonal transmission, i.e, without frequency reuse, scheme discussed previously in Sect. 2.2 requires four time slots to transfer the same four packets from the BS to the two d-UEs.

To understand the benefits of frequency reuse, we now formulate a resource allocation problem similar to (2) wherein the objective is to maximize the sum throughput in the system. Similar to the previous scenario, we have $K$ destination UEs and $N$ subcarriers. We assume that the BS broadcasts messages to the relay-UEs over $T$ time slots and that the relay-UEs forward these messages to d-UEs during time slot $T + 1$. Thereby, the following resource allocation problem can be formulated as

$$
\begin{aligned}
&\underset{P,S}{\text{maximize}} \quad \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} s_{k,m_k,i,j,t} R_{k,m_k,i,j,t} \\
&\text{subject to :} \\
&C_1 : \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} \sum_{j=1}^{N} s_{k,m_k,i,j,t} P^{B,R}_{k,m_k,i,t} \leq P_{max}, \forall t, \\
&C_2 : s_{k,m_k,i,j,t} \in \{0,1\}, \quad \forall i,j,k,m_k,t, \\
&C_3 : \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{j=1}^{N} s_{k,m_k,i,j,t} \leq 1, \quad \forall i,t, \\
&C_4 : \sum_{t=1}^{T} \sum_{m_k \in R_k} \sum_{i=1}^{N} s_{k,m_k,i,j,t} \leq 1, \quad \forall k,j, \\
&C_5 : \sum_{t=1}^{T} \sum_{m_k \in R_k} \sum_{i=1}^{N} \sum_{j=1}^{N} s_{k,m_k,i,j,t} \leq N, \quad \forall k, \\
&C_6 : P^{B,R}_{k,m_k,i,t} \geq 0, \quad \forall i,k,m_k,t,
\end{aligned}
\tag{9}
$$

where $P = \{P^{B,R}_{k,m_k,i,t}\}$ and $S = \{s_{k,m_k,i,j,t}\}$ denote the power allocation and subcarrier allocation policies, respectively. Here, $C_1$ is the power budget constraint for the BS and $P_{max}$ is the maximum power budget for the BS. $C_2$ shows that the subcarrier assignment indicators $s_{k,m_k,i,j,t}$ are binary integer variables. $s_{k,m_k,i,j,t} = 1$ if the BS is transmitting messages to a relay $m_k$ over subcarrier $i$ during time slot $t$, and the relay $m_k$ forwards these messages to the $k$th UE over subcarrier $j$ during time slot $(T+1)$. $s_{k,m_k,i,j,t} = 0$ otherwise. $C_3$ implies that the BS cannot reuse any subcarrier within a given time-slot for broadcasting, i.e, to avoid interference. $C_4$ implies that the subcarriers cannot be reused during a given time-slot for communication within a D2D pair, i.e, to avoid interference. $C_5$ implies that each relay-UE can have a buffer length upto $N$ because a maximum of $N$ messages, i.e, using all the $N$ subcarriers, can be forwarded by a given relay-UE in a given time slot. $C_6$ ensures that the power variables are non-negative.

In simple terms, our goal in (10) is to allocate the BS power and assign the $N$ subcarriers such that the total system throughput is maximized. The achievable throughput $R_{k,m_k,i,j,t}$ for downlink transmission from BS to UE $k$, via the relay $m_k$, over subcarrier $i$ and $j$ in the $t$th and $(T+1)$th time slots , respectively, is given by

$$
R_{k,m_k,i,j,t} = \frac{1}{2} \log_2 \left( 1 + \frac{P^{B,R}_{k,m_k,i,t} P_R g^{B,R}_{k,m_k,i,t} g^{R,U}_{k,m_k,j,T+1}}{P^{B,R}_{k,m_k,i,t} g^{B,R}_{k,m_k,i,t} + P_R g^{R,U}_{k,m_k,j,T+1}} \cdot \frac{1}{N_0 W} \right). \tag{10}
$$

Similar to the scenario in the previous section, the original problem is MINLP. Therefore, we follow a similar solution approach to solve (10). We begin by relaxing the integer variables into continuous intervals and defining auxiliary power variables similar to (3). Plugging these changes into the original problem makes it convex and the dual decomposition technique described in Fig. 3 can be applied. The corresponding dual Lagrangian function is given by

$$L(\boldsymbol{\lambda}, \tilde{P}, S) = \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} s_{k,m_k,i,j,t}$$
$$\frac{1}{2}\log_2\left(1 + \frac{\frac{\tilde{P}^{B,R}_{k,m_k,i,t} g^{B,R}_{k,m_k,i,t}}{s_{k,m_k,i,j,t} N_0 W} P_R g^{R,U}_{k,m_k,j,T+1}}{\frac{\tilde{P}^{B,R}_{k,m_k,i,t} g^{B,R}_{k,m_k,i,t}}{s_{k,m_k,i,j,t}} + P_R g^{R,U}_{k,m_k,j,T+1}}\right) + \tag{11}$$
$$\sum_{t=1}^{T} \lambda_t \left(P_{max} - \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} \sum_{j=1}^{N} \tilde{P}^{B,R}_{k,m_k,i,t}\right),$$

where $\boldsymbol{\lambda} = \{\lambda_t\}$ is the vector of non-negative Lagrangian multipliers associated with the power constraints in $C_1$. The constraints $C_2, C_3, C_4, C_5, C_6$ are not considered in (11) but will be satisfied in subsequent steps. As in the previous solution approach, for a given Lagrangian multiplier $\lambda$, optimal auxiliary power variables $\tilde{P^*}^{B,R}_{k,m_k,i,t}$ and corresponding throughput rates $\tilde{R^*}_{k,m_k,i,j,t}$ can be obtained by applying KKT conditions to the sub-problems. While deriving $\tilde{P^*}^{B,R}_{k,m_k,i,t}$, constraint $C_6$ can be satisfied by offsetting negative values to zero. Plugging back $\tilde{P^*}^{B,R}_{k,m_k,i,t}$ and $\tilde{R^*}_{k,m_k,i,j,t}$ into the sub-problem and re-arranging the equations results in the following simplified problem

$$\begin{aligned} &\underset{S}{\text{maximize}} \quad \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} s_{k,m_k,i,j,t} H_{k,m_k,i,j,t} + \sum_{t=1}^{T} \lambda_t P_{max} \\ &\text{subject to} : C_3, C_4, C_5, \end{aligned} \tag{12}$$

where

$$H_{k,m_k,i,j,t} = \tilde{R^*}_{k,m_k,i,j,t} - \lambda_t \tilde{P^*}^{B,R}_{k,m_k,i,t}. \tag{13}$$

That is, as encountered in the previous solution approach, we are required to solve a linear integer programming problem (12) so as to obtain optimal subcarrier assignment for a given Lagrangian multiplier $\lambda$. Therefore, we propose Algorithm 2 to solve the linear integer programming problem (12) in polynomial time.

Algorithm 2 begins with finding the best relay-UE for all the subcarriers $i$ used by the BS for broadcasting messages and all the subcarriers $j$ used by the relay-UEs for forwarding those messages. Here, the subcarriers are guaranteed to be used at most once in each broadcasting time-slot, i.e, constraint $C_3$ is satisfied. However, subcarrier reuse is permitted in the second time slot because the D2D pairs are far apart and their coverage areas do not overlap, i.e, constraint $C_4$ is satisfied. We also ensure that each relay receives and transmits over the same number of subcarriers. In the process, constraint $C_5$ is also satisfied. Towards the end, we use ordered subcarrier pairing [18] such that the subcarrier with the greatest channel gain among the BS-relay communication links is assigned to the subcarrier with the greatest channel gain among the relay to d-UE communication links. Similarly, subcarriers with the second largest gains are paired, and so on. Such a subcarrier pairing is known to maximize the system throughput [18]. Note, however, that an

**Algorithm 2** Subcarrier allocation algorithm II

1: For each subcarrier, find the relay-UE with the best channel gain to each destination-UE: $m_k^* = \underset{m_k}{\text{maximize}}\, g^{R,U}_{k,m_k,j,T+1}, \forall k, j.$
2: Count the number of subcarriers assigned to each relay-UE and denote it as $a(m_k)$.
3: **for** i=1:N
4: **for** t=1:T
5: For each subcarrier $i$ used by the BS to broadcast messages to the relay-UEs, find the D2D pair with the best channel gain at time slot $t$: $\left(k^*, m_k^*\right) = \underset{k,m_k}{\text{maximize}}\, g^{B,R}_{k,m_k,i,t}.$
6: $a(m_k^*) = a(m_k^*) - 1;$
7: **if** $a(m_k^*) \geq 0$, assignment $\left(i, t, k^*, m_k^*\right)$ is valid.
**else** assignment $\left(i, t, k^*, m_k^*\right)$ is not valid and relay-UE $m_k^*$ is omitted from the relay selection pool.
8: **end**
9: **end**
10: For each relay $m_k$, implement an ordered subcarrier pairing [18].
11: **return** $S^*$

optimal subcarrier assignment can not be found in polynomial time and that the discussed subcarrier assignment algorithm, although sub-optimal, provides a real-time solution. With subcarrier assignment, the solution process for sub-problems is complete.

The Lagrangian multiplier set $\{\lambda_t\}$ is updated in the master problem using sub-gradient methods [14] such as follows

$$\lambda_t(l+1) = \left[\lambda_t(l) - \eta\left(P_{max} - \sum_{k=1}^{K} \sum_{m_k \in R_k} \sum_{i=1}^{N} \sum_{j=1}^{N} \tilde{P}^{B,R}_{k,m_k,i,t}\right)\right]^+, \forall t, \tag{14}$$

where $\eta$ is a constant step size parameter. The master problem and the sub-problems are iteratively solved until a convergence in the Lagrangian multiplier values, i.e, $\lambda_t$, is observed. This completes our discussion on resource allocation for buffer-aided D2D relaying systems.

## *2.4 Simulation Results*

In this section, we present two numerical results to illustrate and compare the performance gains achieved by the proposed resource allocation methods. The simulation setup comprises three candidate relays for each UE. The transmission power for each relay $P_R$ is set to 0.1 W. Figure 4 shows average system capacity versus $P_{max}$ for different number of UEs $K$ considering $N = 16$ subcarriers. We compare the achievable performance gains of a scenario where frequency reuse is possible, i.e, when D2D pairs are far enough from each other with disjoint coverage area, against a scenario where frequency reuse is not possible, i.e, D2D pairs have

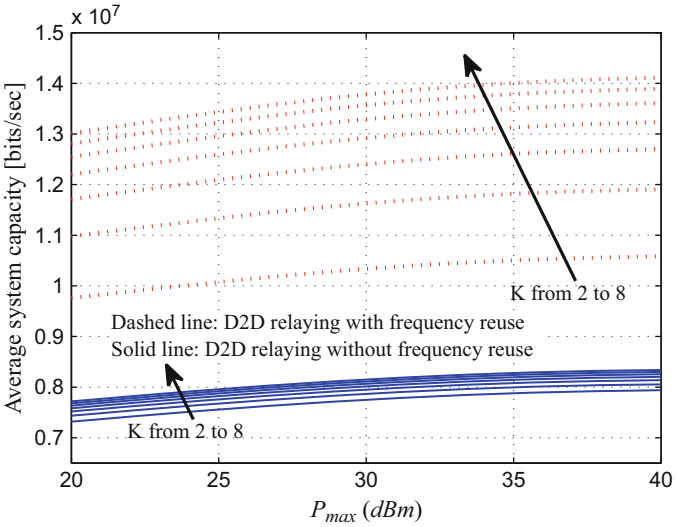


**Fig. 4** Average system capacity versus BS power budget $P_{max}$ for different number of UEs $K$. Result shows significant improvement of system capacity when frequency is reused in D2D links

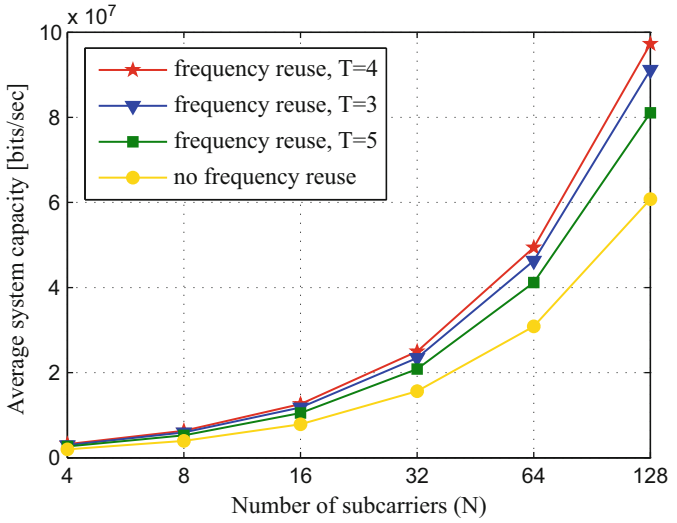


**Fig. 5** Average system capacity versus number of subcarriers $N$ for $P_{max} = 35$ dBm and $K = 4$

overlapping coverage area. Significant system capacity improvement is observed for resource allocation schemes with frequency reuse. With the increase of number of UEs, i.e, $K$, the frequency reusing factor improves and as a result, the capacity gain increases.

In Fig. 5, we show average system capacity versus number of subcarriers $N$ for $P_{max} = 35\,\text{dBm}$ and $K = 4$. As expected the system capacity increases with the increase of the number of subcarriers. Here, we show the performance of the resource allocation scheme with frequency reuse for different values of $T$, where $T$ is the number of time slots for broadcasting. We notice that by increasing $T$ from 3 to 4, the system capacity increases. However, the system capacity decreases when we change $T$ from 4 to 5. The reason is when $T$ is greater than the number of D2D pairs, the number of messages are greater than the size of the buffer at the relays. Therefore, some of the messages are discarded which will result in having lower system capacity than expected.

## *2.5 D2D Relaying: Motivation for a General Scenario*

In the discussions so far, we have considered two theoretically extreme scenarios in the context of D2D relaying. In the first scenario, the D2D pairs are close to each other and their coverage areas overlap. Therefore, since frequency reuse cannot be attempted, we proposed resource allocation algorithms considering the orthogonality of the subcarriers. In the second scenario, the D2D pairs are far enough from each other and their coverage areas are disjoint. Therefore, we proposed resource allocation which fully reuse the subcarriers for D2D communication.

However, in a practical setup, D2D pairs would be distributed randomly in the cell. As a result, both the scenarios discussed in the previous sections can co-exist simultaneously. That is, few D2D pairs would experience coverage overlap with each other, while few other D2D pairs would be disjoint from each other. Resource allocation for such a general scenario can be attempted by developing an analogy based on the conventional frequency reuse patterns used in cellular networks.

In conventional cellular systems, the available frequency spectrum is generally reused in sufficiently separated cells. While adjacent cells use different frequencies to reduce inter-cell interference, the available set of frequencies are usually exhausted within a cluster of cells due to limited spectrum. However, the same frequencies can be reused in adjacent cell-clusters. A more aggressive and effective approach from an interference perspective is the fractional frequency reuse (FFR) technique, where the cells are in-turn partitioned into smaller spatial regions and the available frequency spectrum is reused over these regions.

Similar frequency reuse ideas can be applied in the context of D2D relaying networks. The cellular operator can reserve a dedicated in-band spectrum for D2D communications. Since D2D communications occur over short distances, ultra-dense frequency reuse can be attempted. In other words, several clusters of D2D pairs, with each cluster fully exhausting the available frequency resources, can co-exist within each cell. The D2D pairs within in the same cluster should use orthogonal frequencies, i.e, resource allocation as per the first scenario (c.f. Sect. 2.2) should be attempted. On the other hand, D2D pairs in different clusters can reuse the same set of available frequencies, i.e, resource allocation as per the second

scenario (c.f. Sect. 2.3) should be attempted. On a slightly advanced note, intelligent FFR techniques can also be adopted for D2D communications (for example, see [19]) to minimize interference in the system.

### *2.6 Summary*

In this section, we proposed resource allocation algorithms for OFDMA cellular systems, in which BS communicates with destination-UEs via mobile relay-UEs using D2D links. We considered two scenarios based on the proximity of the D2D pairs. Firstly, we considered that D2D pairs are very close to each other and their coverage areas overlap. In this case, we proposed subcarrier and power allocation without frequency reuse in order to avoid the interferences among the D2D links. Next, we considered that the D2D pairs are far enough and their coverage areas are disjoint. In this case, we proposed subcarrier allocation policy with full frequency reuse with the help of buffer-aided relaying. We also provided motivation for resource allocation in a general scenario where D2D pairs are randomly distributed in a cell, i.e., the two aforementioned scenarios co-exist within the cell. Numerical results show that buffer-aided D2D relaying with frequency reuse can significantly improve the performance of the system compared to the conventional D2D relaying scheme with no frequency reuse.

## 3 Resource Allocation for Energy Harvesting-Based OFDMA Systems with Selective D2D Relaying

In the previous section, we assumed all the relay-UEs are interested in relaying information. However, this assumption is too optimistic for UEs due to their capacity-limited batteries. Without any incentive, the UEs may not be interested in relaying information using their own battery power. In this regard, energy harvesting technology can offer energy incentives for the relay-UEs.

D2D relaying improves the performance of the system if energy harvesting (EH) capabilities are incorporated at the relay-UEs. This is because the relay-UEs can utilize the energy harvested from external resources for relaying operations. Typically, the relays can harvest energy from environmental resources such as wind, solar, and thermal energy or through wireless power transfer from the BS. Environmental energy resources are often not reliable for wireless communication because they are subject to environmental fluctuations and therefore, may not be available anytime and anywhere. On the other hand, in the wireless power transfer (WPT) technology [20], the BS transfers both information and power to the relays using radio frequency (RF) signals. The relays can harvest the energy transmitted by the BS and utilize the harvested energy to establish D2D links with the UEs in

the coverage holes. Thereby, relays facilitate indirect communication between the BS and the UEs in coverage holes. Note that, WPT also offers reliability because the downlink RF transmissions to the relays can be controlled at the BS.

In this section, we study the performance gains achieved by incorporating energy harvesting capabilities into a D2D-relaying OFDMA system. Specifically, the relays harvest energy transferred via WPT by the BS and utilize the harvested energy for relaying operations. Among the various kinds of energy-harvesting receivers proposed in the literature, we focus on relays with energy splitting receivers [21]. Basically, using an energy splitting receiver, the relays split a received signal into two streams: one of them is used for decoding information and the other is used for energy harvesting. Thereby, we develop resource allocation algorithms for maximizing the system throughput. Specifically, we discuss the joint optimization problem of power allocation, subcarrier assignment, and relay-UE selection considering selective cooperation of UE devices. The optimization problem is again an MINLP, which is in general very difficult to solve in its original form due to the coupled integer and continuous variables [11–13]. Similar to the solution approach discussed in Sect. 2.2, we propose a computationally efficient solution approach based on the dual decomposition method. The proposed scheme is more effective and energy-efficient in comparison to the other existing schemes in the literature.

## *3.1 System Model*

Consider a downlink single-cell network where the BS is placed in the center of the cell. We consider an OFDMA system with $N$ orthogonal subcarriers, each of them with a bandwidth of $W$. There are two groups of mobile UEs and the number of UEs in the first and second group is $M$ and $K$, respectively. The first group UEs are close to the BS and have strong direct link channels. The second group UEs are far apart from the BS and have weak direct link channels due to heavy blockage. When cooperation is beneficial for the second group, UEs in the first group act as relays for them via D2D cooperation. In this section, we consider the downlink transmission from the BS to the second group UEs only via D2D-relay assistance from the first group of UEs. We also assume that the first group UEs use separate bands for their own communication with the BS. In the remainder of the section, we will consider second group as users and the first group as relays. The popular DF relaying protocol [22] is used whenever D2D cooperation is beneficial. The relay UEs decide whether to cooperate or not based on the channel conditions.

We denote the relays, i.e., UEs in the first group, by $m$, and the users, i.e, UEs in the second group, by $k$. The transmission modes corresponding to cooperative and non-cooperative are superscripted with $(C)$ and $(NC)$, respectively. The channel gains from BS $B$ to relay $m$, and from relay $m$ to user $k$ are defined as $g_{B,m}^{i,(C,1)}$ and $g_{m,k}^{i,(C,2)}$, respectively. The superscript $i,(C,1)$ and $i,(C,2)$ indicate the cooperative transmission mode over subcarrier $i$ in time slot one and two, respectively. Similarly,

the channel gain for the non-cooperative, i.e, direct, link is defined as $g_{B,k}^{i,(NC)}$. In addition, we define $\alpha_{B,m}^{i,(C,1)} = \frac{g_{B,m}^{i,(C,1)}}{N_0 W}$, $\alpha_{m,k}^{i,(C,2)} = \frac{g_{m,k}^{i,(C,2)}}{N_0 W}$ and $\alpha_{B,k}^{i,(NC)} = \frac{g_{B,k}^{i,(NC)}}{N_0 W}$, where $N_0$ is the variance of the AWGN received by each receiver.

We assume that each UE has two sources of power, one from the battery and the other from the harvested energy. The harvested energy is received by the power splitting operation. Particularly, in the case of cooperative transmission, relays split the received signal from the BS into two streams: one for harvesting the energy ($E$), and the other for decoding the information ($I$) with the ratios of $\rho_{m,k}^{i,(E)}$ and $\rho_{m,k}^{i,(I)}$, respectively. The harvested energy is used by the relays for D2D transmission, i.e., only to forward the data for the users. Therefore, UEs do not consume their own batteries for the purpose of relaying. Since we consider downlink transmissions only, the role of energy harvesting at the UEs in the second group is ignored in our discussions.

### 3.2 Resource Optimization Problem Formulation

We now formulate a resource allocation problem, where the objective is to maximize the total capacity of the system. The capacity of the cooperative link can be calculated as

$$R_{m,k}^{i,(C)} = \frac{1}{2}\min\Big\{\log_2(1+\alpha_{B,m}^{i,(C,1)}\rho_{m,k}^{i,(I)}P_{B,m}^{i,(C)}), \log_2(1+\alpha_{m,k}^{i,(C,2)}\alpha_{B,m}^{i,(C,1)}\rho_{m,k}^{i,(E)}P_{B,m}^{i,(C)})\Big\}, \tag{15}$$

where $P_{B,m}^{i,(C)}$ is the transmission power of BS $B$ on subcarrier $i$ over the cooperative link to send the data for relay $m$. $\alpha_{B,m}^{i,(C,1)}P_{B,m}^{i,(C)}$ is the power received by relay $m$ after attenuation by the fading channel in the first hop. The relay splits the received signal into two power streams with the ratios of $\rho_{m,k}^{i,(I)}$ and $\rho_{m,k}^{i,(E)}$, where $\rho_{m,k}^{i,(I)}$ and $\rho_{m,k}^{i,(E)}$ are used for decoding the information and for harvesting, respectively. In the second time slot, the relay will use the harvested power given by $\alpha_{B,m}^{i,(C,1)}\rho_{m,k}^{i,(E)}P_{B,m}^{i,(C)}$ to forward the data for user $k$. It is assumed that all the energy harvested in first time slot in the relay will be completely utilized for forwarding the information in the second time slot. The remaining harvested energy, if any, at the end of each transmission frame is assumed to be discarded. This assumption simplifies the problem as resources can be optimized over each transmission frame. On the other hand, if we assume that the UEs are capable of storing the harvested energy for future use, dynamic programming methods [23] can be implemented to efficiently utilize the harvested energy over multiple transmission frames. For simplicity, we do not consider the second scenario where dynamic programming is required.

If the direct link is better than the cooperative link, the BS directly transmits data to the user. The capacity of the direct link is calculated as

$$R_k^{i,(NC)} = \log_2(1 + \alpha_{B,k}^{i,(NC)} P_{B,k}^{i,(NC)}), \tag{16}$$

where $P_{B,k}^{i,(NC)}$ is the transmission power of BS $B$ on subcarrier $i$ over the direct link to user $k$.

Our objective is to maximize the total capacity of the system subject to the power constraint of the BS. The joint optimization problem of subcarrier assignment, power allocation, transmission mode, i.e, cooperative or non-cooperative, and relay selection is formulated as

$$\begin{aligned}
&\underset{P,S,\rho}{\text{maximize}} \quad \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{N} s_{m,k}^{i,(C)} R_{m,k}^{i,(C)} + \sum_{k=1}^{K}\sum_{i=1}^{N} s_k^{i,(NC)} R_k^{i,(NC)} \\
&\text{subject to :} \\
&C_1 : \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{N} s_{m,k}^{i,(C)} P_{B,m}^{i,(C)} + \\
&\qquad \sum_{k=1}^{K}\sum_{i=1}^{N} s_k^{i,(NC)} P_{B,k}^{i,(NC)} \le P_{max}, \\
&C_2 : s_{m,k}^{i,(C)}, s_k^{i,(NC)} \in \{0,1\}, \quad \forall i,k,m, \\
&C_3 : \sum_{k=1}^{K}\sum_{m=1}^{M} s_{m,k}^{i,(C)} + \sum_{k=1}^{K} s_k^{i,(NC)} \le 1, \quad \forall i, \\
&C_4 : \rho_{m,k}^{i,(I)} + \rho_{m,k}^{i,(E)} = s_{m,k}^{i,(C)}, \quad \forall i,k,m, \\
&C_5 : P_{B,m}^{i,(C)}, P_{B,k}^{i,(NC)}, \rho_{m,k}^{i,(I)}, \rho_{m,k}^{i,(E)} \ge 0, \quad \forall i,k,m, \\
&C_6 : \sum_{k=1}^{K}\sum_{i=1}^{N} s_{m,k}^{i,(C)} \le n_m, \quad \forall m, \\
&C_7 : \alpha_{B,m}^{i,(C,1)} \rho_{m,k}^{i,(I)} = \alpha_{B,m}^{i,(C,1)} \alpha_{m,k}^{i,(C,2)} \rho_{m,k}^{i,(E)}, \quad \forall i,k,m,
\end{aligned} \tag{17}$$

where $P$, $\rho$ and $S$ denote the power allocation policy, power splitting factors and subcarrier allocation policy, respectively. $C_1$ is the power budget constraint, where $P_{max}$ is the power budget of the BS. $C_2$ shows that $s_{m,k}^{i,(C)}$ and $s_k^{i,(NC)}$ are binary integer variables, i.e, indicators. If BS is transmitting to user $k$ with the assistance of relay $m$ over subcarrier $i$, $s_{m,k}^{i,(C)}$ is one, otherwise it is zero. Similarly $s_k^{i,(NC)}$ is defined for non-cooperative transmission. $C_3$ implies that each subcarrier can be used only once to avoid interference. It also states that over each subcarrier either cooperative or non-cooperative mode can be used. $C_4$ indicates that the sum of the power splitting ratio in each subcarrier should be equal to one if that subcarrier is selected for cooperation, i.e., the splitter is not producing any energy and power wastage is ignored. $C_5$ states that the power and splitting variables are non-negative. $C_6$ indicates that each relay can only assist over $n_m$ subcarriers due to the limitation of the power splitter and capacity of the energy harvester, where $n_m$ is known a priori. The constraint $C_7$ is incorporated to avoid information loss and wastage of power at the relays. The problem stated above is a MINLP which is very difficult to solve in general. An efficient approach which uses the dual decomposition method to obtain an optimal solution for this problem is given in [24].

### 3.3 Solution Approach

In order to solve the problem (17), first we obtain the power splitting ratios for each cooperative link from $C_4$ and $C_7$ as

$$\begin{aligned} \rho_{m,k}^{i,(I)*} &= \frac{h_{m,k}^{i,(C,2)}}{1 + h_{m,k}^{i,(C,2)}}, \\ \rho_{m,k}^{i,(E)*} &= \frac{1}{1 + h_{m,k}^{i,(C,2)}}. \end{aligned} \tag{18}$$

Since we have obtained the optimal power-splitting factors in (18), the variables of the optimization problem in (17) are reduced to $(P, S)$, i.e., transmission power and subcarrier allocation policy.

We follow a solution strategy similar to our previous discussions on MINLP problems (c.f. Sect. 2.2). An overview of the solution approach is provided here. The major concern for solving problem (17) is the presence of integer subcarrier assignment variables and constraints. Therefore, we relax the indicators $S$ to the continuous [0,1] interval and define auxiliary power variables as $\tilde{P}_{A,m}^{i,(C)} = s_{m,k}^{i,(C)} P_{A,m}^{i,(C)}$ and $\tilde{P}_{A,k}^{i,(NC)} = s_{k}^{i,(NC)} P_{A,k}^{i,(NC)}$ to make the optimization problem tractable. The resulting optimization problem becomes convex with respect to the auxiliary and relaxed variables given by $\tilde{P}$ and $S$, respectively. Moreover, strong duality holds, i.e., the duality gap is zero. Therefore, we can solve the Lagrangian dual problem and still obtain the optimal solution of the relaxed problem [24]. An optimal subcarrier allocation policy can be obtained using the popular Hungarian algorithm [25] (c.f. [24] for details) . It can be proved that the solution of the relaxed problem is optimal and has integer values and hence, we can obtain the optimal solution of the original problem (17). The proof is based on the concept of totally unimodularity [15]. The proof of convexity and optimality of the solution for integer variables can be derived similar to [16].

### 3.4 Simulation Results

Let us evaluate the performance of the proposed D2D relaying system for two scenarios: (1) with energy harvesting via WPT and (2) without energy harvesting capabilities. We assume that in both cases $P_{max}$ is the limit on the total power budget of the system. The noise power spectral density $N_0$ and bandwidth of each subcarrier $W$ are assumed to be $5 \times 10^5$ W/Hz and 20 kHz, respectively. The number of users in each group is 8, i.e., $M = K = 8$ and number of subcarriers for each relay $n_m$ is 4.

Figure 6 depicts the average system capacity in terms of the power budget of the BS $P_{max}$ for different number of subcarriers. As expected, the total capacity of the system grows both with the increase of $P_{max}$ and the number of subcarriers. Figure 7

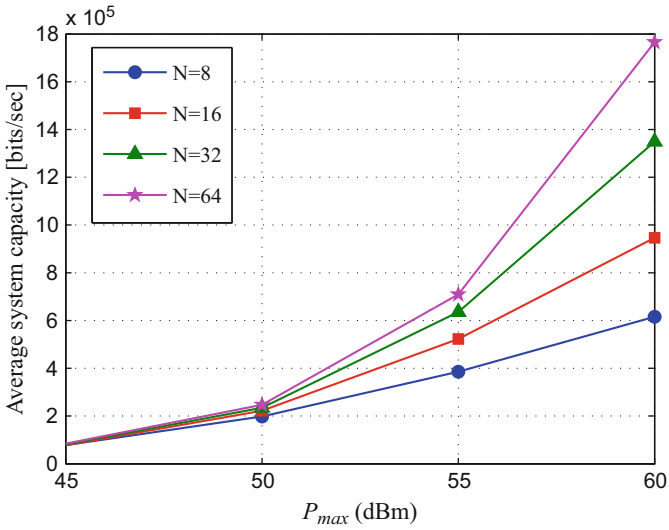


**Fig. 6** Average System Throughput vs. $P_{max}$ for different number of subcarriers

shows the energy efficiency in terms of $P_{max}$ for different number of subcarriers. The energy efficiency is defined as $R_T/P_T$, where $R_T$ is the total system capacity and $P_T$ is the total power consumed in the system. $P_T$ contains both the transmission and the constant power of all the nodes. We compare our proposed scheme with a scenario where there is no energy harvesting. As shown in the figure, our proposed scheme performs better than "No harvesting" scenarios in terms of energy efficiency. It is worth mentioning that in our proposed energy harvesting scheme, the relay UEs do not consume energy from their batteries for the purpose of relaying. Therefore, in terms of battery fairness, our scheme outperforms the "No harvesting" scenario.

## *3.5 Summary*

In this section, we proposed resource allocation algorithm for energy harvesting based OFDMA systems with selective D2D relaying We assumed that the UEs are capable of energy harvesting and do not use their own power for relaying purpose. The harvested energy works as an incentive for the UEs to participate in relaying. A novel method based on dual decomposition technique is proposed to solve the joint optimization problem of power and subcarrier allocation, and relay-UE and transmit mode selection, optimally. Numerical results show the performance improvement of our proposed system over a system without energy harvesting.

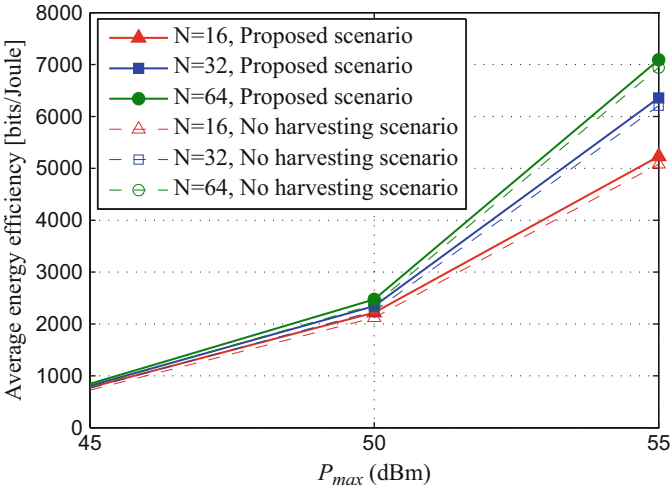


**Fig. 7** Average energy efficiency vs. $P_{max}$ for different number of subcarriers

## 4 Robust Resource Allocation for Cognitive Small Cell Networks with D2D (CR to CR) Cooperation

D2D cooperation can also be utilized to hold in-band communications between devices in the network. Consider, for example, a CR network where an unlicensed UE, i.e., secondary UE, can transmit over the licensed, i.e., primary UE (PU), band, as long as the interference received in the PU band is within the limit set by the regulatory bodies [26]. In a dense CR network with several closely-spaced small cells, the secondary UEs (SUs) operating in different small cells might be interested in communicating with each other for reasons such as data offloading and low-latency application sharing. However, due to interference constraints in the PU bands or due to poor direct communication links, these SUs may not be able to hold direct D2D communications. In such scenarios, D2D cooperation techniques can establish indirect D2D communication links [26–28]. An example scenario is illustrated in Fig. 8.

Figure 8 describes a D2D communication scenario between a source SU $s$ and a destination SU $d$ located in two different small cell networks operating in PU sub-bands $B_s$ and $B_d$ within the system bandwidth, respectively. The two SUs $s$ and $d$ are reasonably far apart but are within the D2D communication range of each other. In such a scenario, any transmission between the SUs in their respective PU bands can add significant interference to the system because the PU receivers can be much closer than the SU receivers. For example, if the source $s$ transmits information directly to the destination $d$, the PU receivers close to the source $s$ experience significant interference. This may not be permissible because the SUs, being unlicensed UEs, should only opportunistically utilize the available spectrum

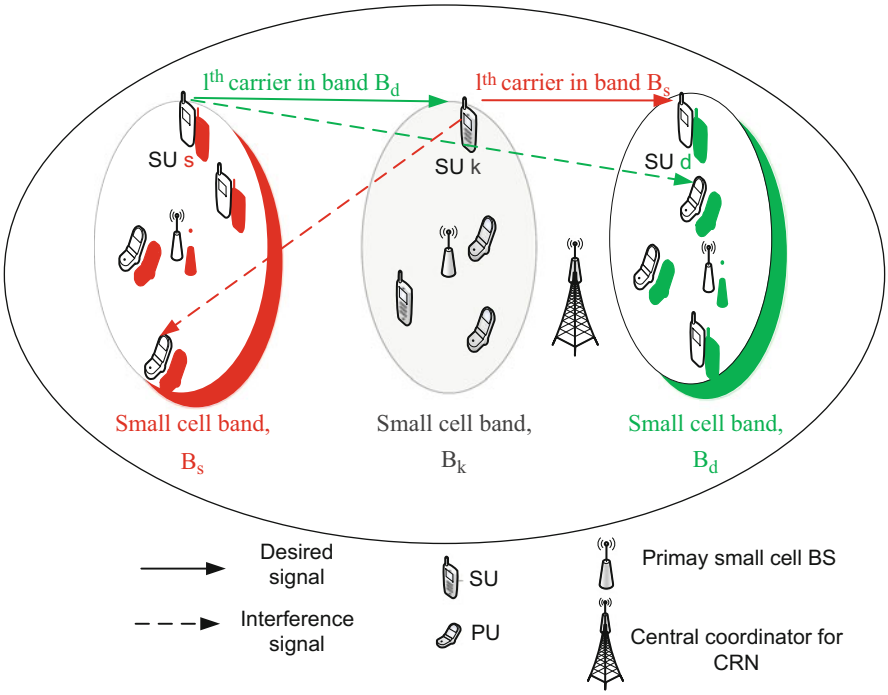


**Fig. 8** SUs *s* and *d* operating in two different small cells cannot establish direct D2D communication links with each other due to poor communication links or due to PU band interference constraints. Another SU *k* operating in a different small cell can relay information between *s* and *d* by establishing separate D2D links with *s* and *d*

while maintaining the interference levels to PUs within an acceptable limit. In such a scenario with PU band interference constraints, another SU *k* operating in a different PU small cell can relay information between SU *s* and SU *d*. For example, the source can transmit a message to the relay *k* over the PU band of the destination, i.e., $B_d$, without exceeding the interference limit. The relay *k* can then forward that message to the destination *d* over the PU band of the source or the destination depending on the received interference level. With the help of such D2D cooperation, low-powered transmissions are enabled over the small cell PU bands while ensuring that the PU interference levels are acceptable. A similar argument is also valid for scenarios where the SUs *s* and *d* are not within the D2D communication range of each other. The relay SU *k* can facilitate a low-powered indirect communication link between SUs *s* and *d*.

In this section, we focus on such CRN scenarios wherein D2D relays establish indirect communication links between SUs located in different small cell networks. We study optimal resource allocation in such D2D relaying systems to improve the CRN performance while keeping the PU band interference below acceptable limits. Specifically, we focus on solving the joint relay-CR selection and power allocation

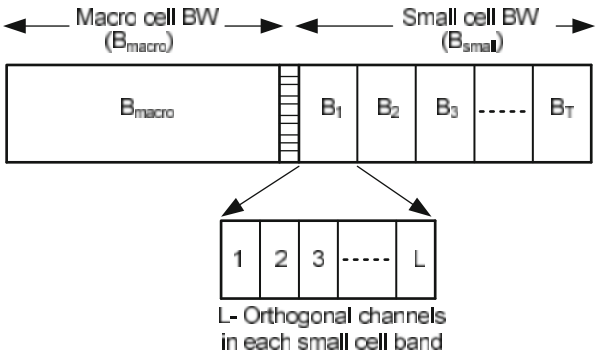


**Fig. 9** Dedicated spectrum allocation to avoid macro cell interferences

problem. Note that, unlike with the systems discussed in previous sections, the performance of CR networks depends heavily on channel uncertainties for the SU-PU channels and channel estimation errors for the SU-SU channels. Therefore, our goal is to propose robust solution to the joint optimization problem by taking channel uncertainties and channel estimation errors into account.

## *4.1 System Model*

Consider a CR network where the SUs are operating from small cell (for example, micro, pico, or femto-cell) PU networks. We also assume that the small cell networks operate in conjunction with a macro cell and that a dedicated spectrum allocation technique (see [29, 30] for example techniques) is employed to avoid interference in the small cells due to macro cell. Basically, as shown in Fig. 9, the total licensed spectrum bandwidth BW is divided into two non-overlapping parts $B_{macro}$ and $B_{small}$, where $B_{macro}$ is used by the macrocell layer and $B_{small}$ is used by the small cell layer. The total small cell BW $B_{small}$ is further divided into $T$ non-overlapping parts and each small cell BW contains $L$ orthogonal channels to avoid the interference among the licensed small cell UEs.

Under such a deployment scenario, as shown in Fig. 8, a set of SU sources $\mathscr{S} = \{1, 2, \ldots, s, \ldots, S\}$ needs to communicate with a set of SU destinations $\mathscr{D} = \{1, 2, \ldots, d, \ldots, D\}$, where the source and destination nodes are far apart and operating from two different small cell PU networks with frequency bands $B_s$ and $B_d$, respectively. We consider that SUs can use the spectrum bands of PUs such that the maximum amount of interference generated at the PU receivers is below a specified limit. As PU receivers are much closer than the SU receivers to the SU transmitter, any transmission of the SUs in their respective PU bands creates significant interference to the PU receivers. For example, a SU source $s$ may not be able to communicate in band $B_s$ and keep the interference to the PU receivers within the limits. Since destination node $d$ is far apart, a direct long distance transmission in band $B_d$ requires high transmission power, which may

not be available at the SUs. High power SU transmission may cause excessive interference to the other intermediate small cell PU receivers operating in the same frequency band. As a result, SU $s$ may not be able to transmit directly to SU $d$ in band $B_d$ and therefore, we look for low-powered CR transmission, which does not create significant interference to the PU networks.

Consider another set of SUs $\mathscr{R} = \{1, 2, \ldots, k, \ldots, K\}$, which is located in between the source and destination nodes and operates from a different PU network with frequency band $B_k$. These intermediate group of users can assist in establishing a communication between the source and the destination nodes via D2D relaying. An SU source group $\mathscr{S}$ can transmit their messages in band $B_d$, which can be decoded by the intermediate relay group $\mathscr{R}$. This transmission over band $B_d$ keeps the interference received at the PUs operating in band $B_d$ within limits. The relay group $\mathscr{R}$ decodes the messages sent by $\mathscr{S}$ and forwards them to the destination over band $B_s$ or $B_d$ depending on their location and the interference received over the PU band. However, the frequency band switching from $B_d$ to $B_s$ is beneficial as it facilitates full-duplex communication at the relay SUs [26]. Similarly, the destination group $\mathscr{D}$ can transmit their messages in band $B_s$ and the relay group can forward them over band $B_d$, and a dual hop D2D communication is possible with the help of the relays. We assume that the SUs in different small cell networks can operate in different frequency bands [26].

The CR network is assumed to have dedicated control channels. During the signaling period, which is a small portion of a transmission frame, SU-SU channels are estimated from a training sequence and resource allocation is performed. Each transmission frame consists of a block of messages and the channel impulse response is assumed to be time invariant, i.e, slow fading, within a frame. A central coordinator for the CR network performs the resource allocation and scheduling. We further assume, for simplicity, that each relay SU can help only one SU during one transmission frame of the CR network. This is a reasonable assumption as relays are small SU devices and supporting multiple users require higher power, which is only possible for the fixed dedicated relay stations mainly used for cell extension purposes in a cooperative or heterogeneous network. Since each relay SU can support a single SU at a time, the problem of choosing a suitable relay from the candidate relay set for each SU arises. In this direction, a simple relay selection strategy, where the relay with the *best* channel conditions is selected, is known to be close to optimal. Moreover, such a simple relay selection strategy has low computational complexity and introduces low overhead into the system [31].

The dual hop communication between a user $s$ and destination $d$ via relay $k$ consists of two phases. In the first phase, user $s$ transmits the message $x_s$ to the relays in carrier $l$ of band $B_d$ with power $p_s$. Without loss of generality, we assume that the $l$-th carrier of the PU band $B_d$ and $B_s$ is assigned to user $s$ and relay $k$, respectively, before the data transmission phase. The received message at relay $k$ is given by

$$y_k = \sqrt{p_s} h_{s,k} x_s + z_k, \tag{19}$$

where $h_{s,k}$ is the channel coefficient for the user $s$ to relay $k$ link. The channel coefficients are statistically modeled as independent and identically distributed (i.i.d) circularly symmetric complex Gaussian (CSCG) random variables (RVs) with zero-mean. $z_k$ denotes the i.i.d. AWGN at the receiver with zero-mean and unit variance.

Relay $k$ decodes the received message and forwards it in the second phase to destination $d$ in carrier $l$ of band $B_s$ with power $p_k$. The received message at destination $d$ forwarded by relay $k$ is given by

$$y_d = \sqrt{p_k} h_{k,d} x_k + z_d, \tag{20}$$

where $h_{k,d}$ is the channel coefficient between relay $k$ and destination $d$. $z_d$ denotes AWGN at the receiver with zero-mean and unit variance.

PUs $pu(d)$ and $pu(s)$ that use the carrier $l$ in band $B_d$ and $B_s$, respectively, will also receive the transmission from the SUs as interference during the first and second phase of the transmission. The interference received at the PUs can be expressed as

$$\begin{aligned} y_{pu(d)} &= \sqrt{p_s} h_{s,pu(d)} x_s, \quad \forall pu(d) \in \mathscr{B}_d^{(l)}, \\ y_{pu(s)} &= \sqrt{p_k} h_{k,pu(s)} x_k, \quad \forall pu(s) \in \mathscr{B}_s^{(l)}, \end{aligned} \tag{21}$$

respectively. $h_{s,pu(d)}$ and $h_{k,pu(s)}$ are the channel coefficients between SU $s$ and PU receiver $pu(d)$ and SU $k$ and PU receiver $pu(s)$, respectively. We denote $\mathscr{B}_d^{(l)}$ and $\mathscr{B}_s^{(l)}$ as the set of PUs using the $l$-th carrier of band $B_d$ and $B_s$ in different small cells, respectively.

The received signal from the CR network, $y_{pu(d)}$ and $y_{pu(s)}$ will result in added interference to the PU receivers in band $B_d$ and $B_s$. To protect the rights of the PUs as licensed owners of the bandwidth, interference generated at the PU receivers should be kept below a specified limit. There are two types of interference constraints (1) per-channel or carrier interference constraint and (2) total interference constraint. Since per-carrier interference constraint is more restrictive than that of the total interference constraint [32], we consider per-carrier interference constraint for resource allocation in our research, which is given by

$$\begin{aligned} g_{s,pu(d)} p_s &\leq I_{th}^{B_d(l)}, \quad \forall pu(d) \in \mathscr{B}_d^{(l)}, \\ g_{k,pu(s)} p_k &\leq I_{th}^{B_s(l)}, \quad \forall pu(s) \in \mathscr{B}_s^{(l)}, \end{aligned} \tag{22}$$

in which $I_{th}^{B_d(l)}$ and $I_{th}^{B_s(l)}$ are the maximum allowed per-carrier interference limit for each CR transmission in carrier $l$ of band $B_d$ and $B_s$, respectively. $g_{s,pu(d)} = |h_{s,pu(d)}|^2$ and $g_{k,pu(s)} = |h_{k,pu(s)}|^2$ are the channel power gains from SU transmitters to PU receivers.

Considering half-duplex decode-and-forward (DF) transmission [33], the mutual information of the links between user $s$ and relay $k$ in the first time slot and between relay $k$ and destination $d$ in the second time slot are given by

$$C_{s,k}^{(1)} = \frac{1}{2}\log_2(1 + |h_{s,k}|^2 p_s),$$
$$C_{k,d}^{(2)} = \frac{1}{2}\log_2(1 + |h_{k,d}|^2 p_k), \tag{23}$$

respectively, where the factor $\frac{1}{2}$ comes from two time slot transmission. $|h_{s,k}|^2$ and $|h_{k,d}|^2$ are the channel power gains of the links between user $s$ and relay $k$, and relay $k$ and destination $d$, respectively. We assume that there is no data buffer at the UEs of the CRN and the maximum achievable data rate of the dual hop link between user $s$ and destination $d$ is limited by the minimum data rate of the two hops.

For our considered cooperative CRN, estimation errors and outdated CSI make the SU-SU channels imperfect [12, 34]. To satisfy the interference constraints in the PU bands, the channels between SU-PU need to be estimated. However, it is extremely difficult to estimate these channels perfectly. Keeping this in mind, we formulate a resource allocation problem by considering two types of channel estimation error models.

## *4.2 SU to SU Channel Model with Imperfectness*

For cooperative D2D-based CR transmission, we consider that training sequences are used to obtain the CSI of SU-SU channels and the resulting channel uncertainty is unbounded Gaussian [35]. SU to SU transmission channel coefficients are modeled as $h = \hat{h} + e$, where $\hat{h}$ is the estimated channel and $e$ denotes the error in estimation. The error $e$ is modeled as i.i.d and distributed as complex Gaussian with zero mean and variance given by $\sigma_e^2$. We also assume that the estimated channel $\hat{h}$ and the error $e$ are statistically independent.

Using $h = \hat{h} + e$, the mutual information expressions in (23) can be re-written as

$$C_{s,k}^{(1)} = \tfrac{1}{2}\log_2(1 + \left|\hat{h}_{s,k} + e_{s,k}\right|^2 p_s),$$
$$C_{k,d}^{(2)} = \tfrac{1}{2}\log_2(1 + \left|\hat{h}_{k,d} + e_{k,d}\right|^2 p_k). \tag{24}$$

Depending on the signal detection methods employed, lower bounds on the ergodic capacities can be derived. Note that, with CSI error, a channel outage occurs whenever the actual instantaneous mutual information drops below the target data rate $r_s$. Such a condition is mathematically given by

$$\min(C_{s,k}^{(1)}, C_{k,d}^{(2)}) < r_s, \tag{25}$$

even when channel capacity achieving coding is applied for error protection [12]. This is because the instantaneous mutual information in (24) is a random variable

due to the error in estimation. Note that the lower bounds on ergodic capacities do not capture such outage events. Therefore, we should define a new system performance measure which captures the effect of outage events.

In this direction, we adopt the system goodput [12, 36] as the performance measure and weighted system goodput as the objective function of the problem. In the following, we first define the instantaneous goodputs $U_{s,k}^{inst}$ and $U_{k,d}^{inst}$, i.e., bits per second per Hz successfully delivered to relay $k$ and destination $d$, respectively as

$$\begin{aligned} U_{s,k}^{inst} &= r_s \times 1(r_s \leq C_{s,k}^{(1)}), \\ U_{k,d}^{inst} &= r_s \times 1(r_s \leq C_{k,d}^{(2)}), \end{aligned} \tag{26}$$

where $1(\cdot)$ is an indicator function, i.e., the function value is 1 if $(\cdot)$ is true and zero otherwise.

The average goodput $\mathrm{U}_{s,k}^{goodput}$ for each $(s,k)$ pair can be defined as minimum of the total average bits per second per Hz successfully delivered to relay $k$ and destination $d$, and is given by

$$\begin{aligned} &\mathrm{U}_{s,k}^{goodput} = E\left\{\min(U_{s,k}^{inst}, U_{k,d}^{inst})\right\} \\ &= E_{\hat{h}_{s,k},\hat{h}_{k,d}} \times \left\{\min\left(r_s E_{\hat{h}_{s,k}} \times \left[1\left(r_s \leq C_{s,k}^{(1)}\right)\right],\right.\right. \\ &\left.\left. r_s E_{\hat{h}_{k,d}} \times \left[1\left(r_s \leq C_{k,d}^{(2)}\right)\right]\right)\right\} \\ &= E_{\hat{h}_{s,k},\hat{h}_{k,d}} \times \left\{\min\left[r_s \times \Pr.\left(r_s \leq C_{s,k}^{(1)} | \hat{h}_{s,k}\right),\right.\right. \\ &\left.\left. r_s \times \Pr.\left(r_s \leq C_{k,d}^{(2)} | \hat{h}_{k,d}\right)\right]\right\}. \end{aligned} \tag{27}$$

$E_{\hat{h}}(\cdot)$ denotes the statistical expectation with respect to the RV $\hat{h}$. Finally, by incorporating the channel outage events, we define a new objective function of the problem as the average weighted system goodput,

$$\mathrm{U}_{goodput} = \sum_{s=1}^{S} \sum_{k=1}^{K} w_s \, \mathrm{U}_{s,k}^{goodput}. \tag{28}$$

## *4.3 SU to PU Channel Models with Uncertainty*

### (a) Model-1: Probabilistic

We assume that CR transmitters cannot estimate the instantaneous channels and have information only about the distribution type and the corresponding distribution parameters of the channel gains among the CR transmitter to the PU receivers. We consider a probabilistic model, which requires only the knowledge of channel gain statistics instead of instantaneous CSI. We assume that the CR transmitters have some knowledge of the transmit power of the PUs and hence can obtain the

mean channel gains from PUs to themselves by exploiting the pilot signals from the PUs [37]. Using the reciprocal property of the wireless channels, some statistics of the SU-PU channels can be obtained [37, 38].

For example, suppose the SU to PU channels $h_{s,pu(d)}$ and $h_{k,pu(s)}$ are statistically modeled as i.i.d circularly symmetric complex Gaussian RVs with zero mean and variances $\sigma^2_{s,pu(d)}$, and $\sigma^2_{k,pu(s)}$, respectively. We assume that only the mean and the variance of the channels are known to the CRN. Since the channel coefficients are distributed as CSCG, their channel power gains $g_{s,pu(d)}$ and $g_{k,pu(s)}$ are exponentially distributed with parameters $\lambda_{s,pu(d)} = 1/(\sigma^2_{s,pu(d)})$ and $\lambda_{k,pu(s)} = 1/(\sigma^2_{k,pu(s)})$, respectively.

**(b) Model-2: Bounded Uncertainty**

In this case, we assume that the uncertainty among the SU-PU channels is bounded and no statistical knowledge is available about it. Without restricting ourselves to any specific estimation schemes, we model the channel as $h = \hat{h} + \epsilon$, where $\epsilon$ is the estimation error which determines how far the actual channel $h$ can extend in both real and imaginary parts from the estimated value $\hat{h}$. The value of $\epsilon$ depends on the type of the error and the accuracy of the estimation method used. We use the bounded uncertainty model presented in [39]. This model can be viewed as deterministic by considering the worst case error, i.e., $|\epsilon| = \epsilon_{max}$ [40]. The interference constraints $C_3$ and $C_4$ are satisfied for the worst case and hence guaranteed for all channels.

The channel gain with uncertainty can be written as

$$g = |h|^2 = (\hat{h} + \epsilon)(\hat{h} + \epsilon)^* = \hat{g} + \delta, \tag{29}$$

where $\hat{g} = \hat{h}\hat{h}^*$ is the estimated channel gain, $\delta = \epsilon\epsilon^* + 2\Re(\hat{h}\epsilon^*)$ is the uncertainty in estimation and $\Re(\cdot)$ is the real part of $(\cdot)$. For each $h$ that lies in some bounded region, the channel gain $g$, lies on a set of line segment [39], given by

$$g \in \mathscr{L} = \{\hat{g} + \delta v \mid |v| \leq 1\}. \tag{30}$$

To determine the worst possible case, we consider the maximum error given by $|\epsilon| = \epsilon_{max}$. Note that the line segment, $\mathscr{L}$, is generally asymmetric around the estimated gain $\hat{g}$ for $\epsilon \neq 0$. For the worst case scenario, we take the maximum absolute uncertainty, $\delta^{max} = \epsilon_{max}\epsilon^*_{max} + 2|\Re(\hat{h}\epsilon^*_{max})|$, as the uncertainty on both sides of $\hat{g}$, and re-write the channel gain as

$$g \in \mathscr{L} = \{\hat{g} + \delta^{max} v \mid |v| \leq 1\}, \tag{31}$$

where the estimated channel gain $\hat{g}$ is the mid point of the line segment $\mathscr{L}$. Therefore, the channel gains to the PU receiver in the $l$-th carrier of band $B_d$ and $B_s$ from the CR transmitter $s$ and $k$ can be written as

$$
\begin{aligned}
g_{s,pu(d)} \in \mathscr{L}_{s,pu(d)} &= \{\hat{g}_{s,pu(d)} + \delta^{max}_{s,pu(d)} v \mid |v| \leq 1\}, \\
g_{k,pu(s)} \in \mathscr{L}_{k,pu(s)} &= \{\hat{g}_{k,pu(s)} + \delta^{max}_{k,pu(s)} v \mid |v| \leq 1\}.
\end{aligned} \tag{32}
$$

### 4.4 Problem Formulation with Imperfect CSI

In this section, we formulate the robust resource allocation problem taking into account the estimation errors in the SU-SU channels and the uncertainties among the SU-PU channels. The optimization objective is to maximize the weighted system goodput of the CRN. To achieve that, we need to optimally select the relays and allocate power for CR transmission. In addition, we need to satisfy the QoS of the SUs such that the interference received at the PU receivers remain within limits. In order to solve the optimization problem, we first obtain the power allocation solution for each feasible $(s, k)$ pair and then find the optimal assignment pairs $(s^*, k^*)$ [41].

The robust power allocation problem is probabilistic and can be written as

$$
\begin{aligned}
\forall (s,k) \in S \times K: \quad & \underset{\{p_s, p_k\}}{\text{argmax}} \quad \mathrm{U}^{goodput}_{s,k}, \\
& \text{subject to :}, \\
& C_1 : \Pr.\left(r_s > C^{(1)}_{s,k} | \hat{h}_{s,k}\right) \leq p_{out}, \\
& C_2 : \Pr.\left(r_s > C^{(2)}_{k,d} | \hat{h}_{k,d}\right) \leq p_{out}, \\
& C_3 : \Pr.\left(g_{s,pu(d)} p_s \leq I^{B_d(l)}_{th}\right) \geq c, \quad \forall pu(d) \in \mathscr{B}^{(l)}_d, \\
& C_4 : \Pr.\left(g_{k,pu(s)} p_k \leq I^{B_s(l)}_{th}\right) \geq c, \quad \forall pu(s) \in \mathscr{B}^{(l)}_s, \\
& C_5 : 0 \leq p_s,\, p_k \leq P_{max},
\end{aligned} \tag{33}
$$

where $C_1$ and $C_2$ are the QoS constraints, which guarantee the target data rate $r_s$ of user $s$ with a probability of $1 - p_{out}$. Constraints $C_3$ and $C_4$ incorporate *Model-1* of SU-PU channel uncertainty. These constraints satisfy the interference to PU bands with a probability $c$, which is a predefined term known to the CR coordinator. $C_5$ is the non-negative individual power budget constraint for the UEs and limits the power to a maximum of $P_{max}$.

For *Model-2* that considers bounded uncertainty, the optimization problem given in (33) would remain the same, except for the constraints $C_3$ and $C_4$ which depend on the channel uncertainty model. New constraints corresponding to $C_3$ and $C_4$ can be written from (32) considering the worst case scenario as

$$
\begin{aligned}
C_3 &: \sup\{g_{s,pu(d)} p_s \mid g_{s,pu(d)} \in \mathscr{L}_{s,pu(d)}\} \leq I^{B_d(l)}_{th}, \\
&\qquad\qquad \forall pu(d) \in \mathscr{B}^{(l)}_d \\
C_4 &: \sup\{g_{k,pu(s)} p_k \mid g_{k,pu(s)} \in \mathscr{L}_{k,pu(s)}\} \leq I^{B_s(l)}_{th}, \\
&\qquad\qquad \forall pu(s) \in \mathscr{B}^{(l)}_s.
\end{aligned} \tag{34}
$$

Note that supremum of a set, sup{·}, captures the essence of worst case optimization.

### 4.5 Optimization Framework for the Proposed Scheme

In this section, we briefly describe the optimization framework for the formulated probabilistic optimization problem (33). We first transform the probabilistic problem to deterministic and develop a power allocation algorithm. Once the power allocation solution is obtained, we can determine the optimal relay assignment for the transformed problem based on the Hungarian algorithm [25].

In order to solve the problem efficiently, we incorporate the outage-probability (or, QoS) constraints $C_1$ and $C_2$ of (33) into the objective function. This is possible if the constraints are fulfilled with equality at the optimal point, which is generally the case for the low outage probabilities required in practical applications [12]. Thus, the resulting optimization problem contains stronger constraints than the original problem. Accordingly, the first QoS constraint becomes

$$C_1 : \Pr.\left(r_s > C_{s,k}^{(1)} | \hat{h}_{s,k}\right) = p_{out}. \tag{35}$$

Using (24), $C_1$ can be expressed as

$$\Pr.\left(\left|\hat{h}_{s,k} + e_{s,k}\right|^2 \leq \frac{(2^{2r_s} - 1)}{p_s} | \hat{h}_{s,k}\right) = p_{out}, \tag{36}$$

where $\chi^2_{2_{s,k}} = \left|\hat{h}_{s,k} + e_{s,k}\right|^2$ is a non-central chi-square distributed random variable with non-centrality parameter, $\lambda^2_{s,k} = \dfrac{|\hat{h}_{s,k}|^2}{\sigma^2_{e_{s,k}}/2}$ and degrees of freedom $n = 2$.

To obtain a closed-form result, we approximate the non-central chi-square distribution by the central chi-square $\chi^2_{2_{s,k}}(0)$ [42, 43] and evaluate as

$$\Pr.\left(\chi^2_{2_{s,k}} \leq \frac{2^{2r_s} - 1}{p_s}\right) \approx \Pr.\left(\chi^2_{2_{s,k}}(0) \leq \frac{\left(2^{2r_s} - 1\right)/p_s}{1 + \lambda^2_{s,k}/2}\right) = 1 - e^{-\frac{\left(2^{2r_s} - 1\right)/p_s}{2\left(1 + \lambda^2_{s,k}/2\right)}}. \tag{37}$$

Using the expression of (37) in (36), we can obtain the guaranteed QoS considering channel estimation error in SU source to relay links as

$$r_s = \frac{1}{2}\log_2(1 + p_s\alpha_{s,k}), \tag{38}$$

where $\alpha_{s,k} = [-\ln(1-p_{out})] \times 2\left(1+\lambda_{s,k}^2/2\right)$.

Similarly, $C_2$ can be transformed into an equality constraint and a similar expression to (38) can be derived.

Since the power and interference constraints are instantaneous in (33), as far as the optimization problem is concerned, the average goodput $\mathrm{U}_{s,k}^{goodput}$ maximization is equivalent to the maximization of the instantaneous goodput for each set of channel gains, although both criteria are different in general. Therefore, we can write

$$\begin{aligned}
&\underset{\{p_s,p_k\}}{\text{argmax}} \quad \mathrm{U}_{s,k}^{goodput} \\
&= E_{\hat{h}_{s,k},\hat{h}_{k,d}} \times \left\{\min\left(r_s \times \Pr.\left(r_s \le C_{s,k}^{(1)}|\hat{h}_{s,k}\right), r_s \times \Pr.\left(r_s \le C_{k,d}^{(2)}|\hat{h}_{k,d}\right)\right)\right\} \\
&\equiv \underset{\{p_s,p_k\}}{\text{argmax}} \quad \left\{\min\left(r_s \times \Pr.\left(r_s \le C_{s,k}^{(1)}|\hat{h}_{s,k}\right), r_s \times \Pr.\left(r_s \le C_{k,d}^{(2)}|\hat{h}_{k,d}\right)\right)\right\} \\
&= \underset{\{p_s,p_k\}}{\text{argmax}} \quad \min\left\{(1-p_{out})\tfrac{1}{2}\log_2(1+p_s\alpha_{s,k}), (1-p_{out})\tfrac{1}{2}\log_2(1+p_k\alpha_{k,d})\right\},
\end{aligned} \tag{39}$$

where "≡" denotes the equivalence relation.

## 4.6 Optimization Framework for Model-1: Probabilistic Constraints

Since $g_{s,pu(d)}$ is exponentially distributed with parameter $1/(\sigma_{s,pu(d)}^2)$, the left hand side of the first constraint $C_1$ in (33) can be evaluated as

$$\Pr.\left(g_{s,pu(d)}p_s \le I_{th}^{B_d(l)}\right) = 1 - e^{-\left(\frac{I_{th}^{B_d(l)}}{\sigma_{s,pu(d)}^2 p_s}\right)}. \tag{40}$$

Therefore, the problem (33) can be written as a deterministic optimization problem given by

$$\begin{aligned}
\forall(s,k) \in S \times K: \quad &\underset{\{p_s,p_k\}}{\text{argmax}} \quad \min\left\{\frac{(1-p_{out})}{2}\log_2(1+p_s\alpha_{s,k}), \frac{(1-p_{out})}{2}\log_2(1+p_k\alpha_{k,d})\right\} \\
&\text{subject to :} \\
&C_1 : p_s \le \frac{I_{th}^{B_d(l)}}{\sigma_{s,pu(d)}^2\{-\ln(1-c)\}}, \quad \forall pu(d) \in \mathscr{B}_d^{(l)}, \\
&C_2 : p_k \le \frac{I_{th}^{B_s(l)}}{\sigma_{k,pu(s)}^2\{-\ln(1-c)\}}, \quad \forall pu(s) \in \mathscr{B}_s^{(l)}, \\
&C_3 : 0 \le p_s, p_k \le P_{max}.
\end{aligned} \tag{41}$$

### 4.7 Optimization Framework for Model-2: Bounded Uncertainty Constraints

For model-2 with bounded uncertainty region given in (34), we propose worst case optimization approach in which the interference constraints are satisfied for all possible values of the channel gains that lie in some bounded region. Therefore, the constraint $C_3$ in (34) can now be expressed as

$$\begin{aligned} & \sup\{g_{s,pu(d)}p_s \mid g_{s,pu(d)} \in \mathscr{L}_{s,pu(d)}\} \leq I_{th}^{B_d(l)}, \quad \forall pu(d) \in \mathscr{B}_d^{(l)} \\ \Rightarrow \; & \sup\{(\hat{g}_{s,pu(d)} + \delta_{s,pu(d)}^{max} v)p_s \mid |v| \leq 1\} \leq I_{th}^{B_d(l)}, \quad \forall pu(d) \in \mathscr{B}_d^{(l)} \\ \Rightarrow \; & \hat{g}_{s,pu(d)}p_s + \delta_{s,pu(d)}^{max} p_s \leq I_{th}^{B_d(l)}, \quad \forall pu(d) \in \mathscr{B}_d^{(l)}. \end{aligned} \tag{42}$$

Following a similar procedure, a similar expression to (42) can be derived for the second interference constraint $C_4$ of (34). Finally, replacing $C_1$ and $C_2$ in (41), respectively with the constraints derived here, we obtain an equivalent deterministic optimization problem for Model-2.

### 4.8 Solution Approach

It can be proved that the robust power optimization problem in (41) (for both the channel uncertainty models) is convex and can be solved efficiently. The algorithm to obtain the solution for problem (41) is given in Algorithm 3.

In Algorithm 3, we obtain the closed form analytical expression for optimal $p_s^*$ and $p_k^*$ for user-relay $(s,k)$ pair for problem (41).

The optimal relay assignment can be obtained by forming a $S \times K$ matrix $\mathbf{B} = [b(s,k)]$, such that

$$b(s,k) = w_s \left[ (1 - p_{out}) \frac{1}{2} \log_2(1 + p_s^* \alpha_{s,k}) \right] = w_s \left[ (1 - p_{out}) \frac{1}{2} \log_2(1 + p_k^* \alpha_{k,d}) \right].$$

By applying Hungarian algorithm [25] on matrix $\mathbf{B}$, we obtain the optimal $(s^*, k^*)$ pairs that maximize the weighted goodput of the CRN.

### 4.9 Simulation Results

We conduct simulations to show the effectiveness of our proposed solution and compare the performance with the baseline scheme that ignores the imperfectness in CSI. Different system parameters are considered to analyze the influence of different channel estimation errors on the system performance. Frequency-flat Rayleigh fading channels are assumed in all simulations. Each of the channel coefficients is

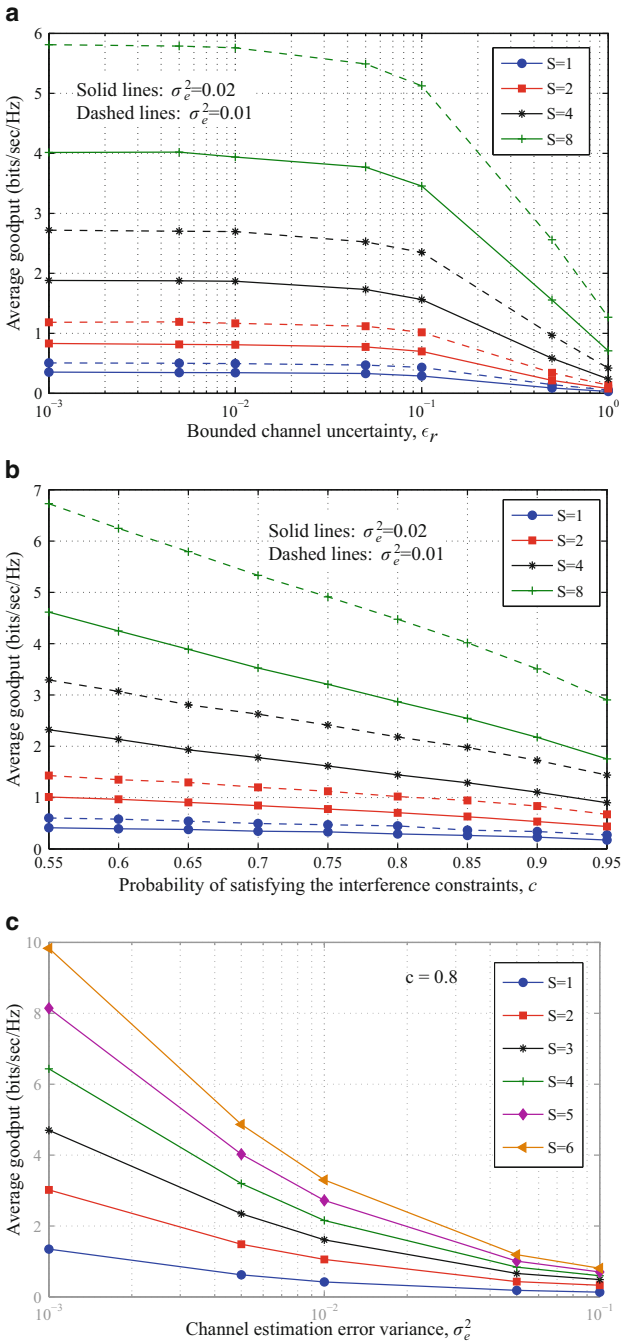


**Fig. 10** Average goodput versus different uncertainty models among the SU-PU channels and SU-SU channels. (**a**) Average goodput versus uncertainty distance, $\epsilon_r$ among the channels of SU transmitters to PU receivers. (**b**) Average goodput versus probability of satisfying the interference constraints $c$. (**c**) Average goodput versus channel estimation error variance $\sigma_e^2$ of SU-SU channels

**Algorithm 3** Power allocation for problem (41)

**Require:** $\alpha_{s,k},\ \alpha_{k,d},\ I_{th}^{B_d(l)},\ I_{th}^{B_s(l)},\ c,\ P_{max},\ \sigma_{s,pu(d)}^2,\ \sigma_{k,pu(s)}^2,\ \hat{g}_{s,pu(d)},\ \hat{g}_{k,pu(s)},\ \delta_{s,pu(d)}^{max},\ \delta_{k,pu(s)}^{max}$.

1: Calculate $\text{Pow}_s$ and $\text{Pow}_k$ given by

$$\text{Pow}_s = \min\left(\frac{I_{th}^{B_d(l)}}{\sigma_{s,pu(d)}^2\{-\ln(1-c)\}}, P_{max}\right); \text{for model-1}$$

$$\text{Pow}_s = \min\left(\frac{I_{th}^{B_d(l)}}{\hat{g}_{s,pu(d)} + \delta_{s,pu(d)}^{max}}, P_{max}\right); \text{for model-2}$$

$$\text{Pow}_k = \min\left(\frac{I_{th}^{B_s(l)}}{\sigma_{k,pu(s)}^2\{-\ln(1-c)\}}, P_{max}\right); \text{for model-1}$$

$$\text{Pow}_k = \min\left(\frac{I_{th}^{B_s(l)}}{\hat{g}_{k,pu(s)} + \delta_{k,pu(s)}^{max}}, P_{max}\right); \text{for model-2}$$

2: **if** $\left(\frac{\text{Pow}_s\alpha_{s,k}}{\alpha_{k,d}} \le \text{Pow}_k\right)$ **then**

3: $\quad p_s^* = \text{Pow}_s\ , p_k^* = \frac{\text{Pow}_s\alpha_{s,k}}{\alpha_{k,d}}$

4: **else**

5: $\quad p_k^* = \text{Pow}_k\ , p_s^* = \frac{\text{Pow}_k\alpha_{k,d}}{\alpha_{s,k}}$

6: **end if**

an i.i.d. circularly symmetric complex Gaussian random variable and distributed as $\mathscr{CN}\left(0, \frac{1}{(1+d)^\alpha}\right)$ [44], where $d$ is the normalized distance among various nodes in the network and $\alpha = 4$ is the path loss coefficient.

The bounded uncertainty region for model-2, as discussed in Sect. 4.7, is considered to be rectangular with four equal sides. The uncertainty distance, i.e., the distance from the center of the rectangle to the center of each side (or, equivalently half length of a side) is given by $\epsilon_r$. Same channel estimation error variance $\sigma_e^2$ is assumed for all SU to SU channels. We also consider that the maximum per-carrier interference limit in both band $B_d$ and $B_s$ is the same, i.e., $I_{th}^{B_d(l)} = I_{th}^{B_s(l)} = I_{th}$. Since both the source and the relays are UEs, the maximum power available are assumed to be the same. Unless stated otherwise, probability of outage, $p_{out}$, is assumed to be 0.1, equal priority is given to all the users by setting $w_s = 1$, maximum power, $P_{max}$ and interference threshold limit, $I_{th}$, are set at 1 and 0.1 W, respectively. The results are averaged over 10,000 independent channel realizations.

Figure 10a–c show the average system goodput versus the uncertainty distance $\epsilon_r$, probability of satisfying the interference constraints $c$, and the channel estimation error variance $\sigma_e^2$ of the SU channels, respectively, for different number of users $S$. We see that average goodput decreases with the increase of $\epsilon_r$, $c$, and $\sigma_e^2$ and increases with the increase of the number of users $S$. The effect of error variance $\sigma_e^2$ of the SU channels is more significant at the lower values of $\epsilon_r$ and $c$ and at higher number of users. This is because low $\epsilon_r$ or $c$ values indicate better estimation of the channels between SU transmitters and PU receivers. Low $\epsilon_r$ or $c$ values allow higher transmission capability of the CRN as the interference threshold

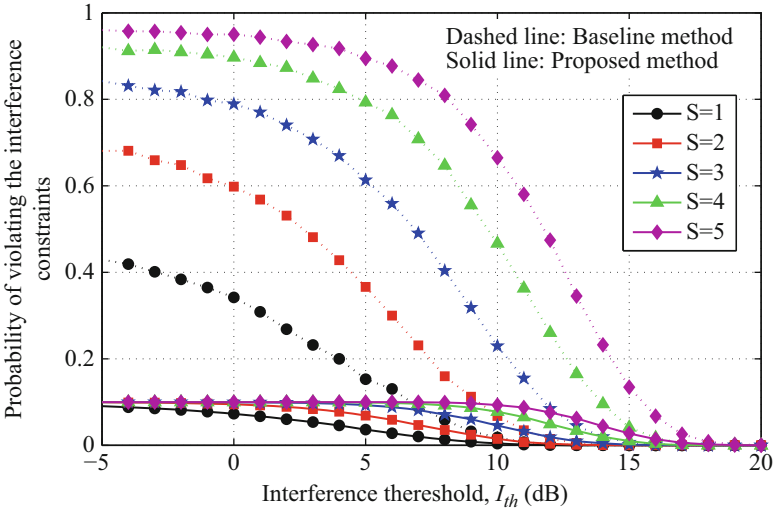


**Fig. 11** Probability of violating the interference constraints versus interference threshold $I_{th}$

margin becomes more flexible. Similarly, with increasing number of users, more power is transmitted by the CRN. Since the average goodput is proportional to the transmission power and inversely proportional to $\sigma_e^2$, the effect of estimation error in the SU-SU channels becomes more detrimental at high power regimes.

Figure 11 shows the probability of violating the interference constraints given in (33) versus interference threshold $I_{th}$ for different number of users. In this simulation, we consider that some uncertainty is associated with the channels of SU transmitters to PU receivers and compare the performance of our proposed scheme with the baseline scheme which assumes the estimated channel to be perfect. We observe frequent violations of the constraints for the baseline scheme in the active region of $I_{th}$ for ignoring the channel uncertainty in the design process. Since increasing the number of users use more power to improve the system goodput, the probability of violation increases with the increase of the number of users. Similar trends of interference violation can be observed for the resource allocation schemes proposed in [45, 46]. For our proposed probabilistic interference scheme, i.e., model-1, we set $c = 0.9$. That is, we allow the interference constraints to be violated with a probability of 0.1 when the constraints are active. As a result, we see that the maximum value of interference violation is $1 - c = 0.1$, which is much lower than the baseline scheme. Model-2, i.e., the bounded uncertainty model is designed to satisfy the interference constraints for the worst case scenario. As a result, no violation is observed for our proposed scheme when the uncertainty lies in some bounded region.

### *4.10 Summary*

We have developed robust power allocation and relay-CR selection schemes for a CR network where the SUs from different small cell PU networks communicate with each other. In such a network with tight interference constraints in the PU band, close proximity of the PUs and limited power budgets of SUs, the communication among the SUs is established via D2D cooperation. We have developed efficient resource allocation schemes with the provision of QoS to each SU considering different channel uncertainty models in the network. Channel outage events, which have resulted from the imperfect CSI under slow fading channels, have been taken into account in our algorithm. System goodput of the CR network has been maximized while satisfying the interference constraints for the PU bands both probabilistically and for the worst case scenario. The joint optimization problem is probabilistic and non-convex, which is very difficult to solve in its original form. The problem has been approximated and transformed into a convex deterministic form and a closed form analytical solution for power allocation has been derived. The closed form power allocation solution helped us to develop an efficient relay-CR selection scheme based on Hungarian algorithm. Our results reveal that the implications of ignoring the imperfectness among different channels are violations of interference constraints and failure to provide the QoS, which ultimately result in retransmissions and wastage of power.

## 5 Conclusion and Future Research Directions

In this chapter, we discussed several resource allocation problems which attempt throughput maximization in cooperative D2D networks operating under OFDMA or cognitive radio architectures. A variety of system models were considered to understand how cooperative D2D communications can benefit from incorporating packet storage using buffers, energy-harvesting, and cognitive radio capabilities at UEs in the network. For each system, we provided efficient solutions to joint resource optimization problems including power allocation, subcarrier allocation, subcarrier pairing, and relay-UE selection. Significant throughput gains are observed in each system, thus proving that the system performance can be significantly improved whenever the UEs are enabled with cooperative D2D capabilities.

Several interesting research works can be pursued to further investigate the potential performance gains offered by cooperative D2D communications. Energy Efficiency is becoming an important performance metric for 5G networks because there are growing concerns about the global carbon footprint of telecommunication systems. Energy Efficiency, which is the ratio of the total achieved throughput to the total power consumption in the system, is a fractional and nonlinear function and is different from the conventional performance measures. Therefore, potential future research works can be to develop energy efficient resource allocation schemes

for the cooperative D2D systems discussed in this chapter. Another extension of this research can be to develop an intermix of the presented system models to develop a comprehensive understanding of the benefits offered by cooperative D2D. For example, resource allocation problems can be formulated to understand the benefits of incorporating both packet-storage and energy-harvesting capabilities at the UEs in cooperative D2D networks. Also, investigating the role of strict and soft fractional frequency reuse patterns in enabling dense cooperative D2D networks with spectrum availability constraints can be a possible future research work.

Another research direction in the field of cooperative D2D and mobile relays can be investigating radio resource management problems such as opportunistic D2D mode selection, i.e, switch opportunistically between overlay D2D and underlay D2D, user association, link adaptation, multi-casting, and delay-tolerant data offloading. Appropriate interference mitigation, load-balancing, and scheduling algorithms can be derived when cooperative D2D capabilities are incorporated in ultra-dense multi-tier heterogenous networks (HetNets). Also, the benefits of cooperative D2D communications can be evaluated when operated in conjunction with other 5G enabling technologies, such as massive MIMO, millimeter wave spectrum, and full-duplex radio technologies. Wherever applicable, distributed resource allocation algorithms can be proposed based on suitable trade-offs between computational complexities and expected delays.

## References

1. S. Mumtaz, J. Rodriguez, *Smart Device to Smart Device Communication* (Springer, Cham, 2014)
2. L. Song, D. Niyato, Z. Han, E. Hossain, *Wireless Device-to-Device Communications and Networks* (Cambridge University Press, New York, NY, 2015)
3. K.W. Choi, Z. Han, Device-to-device discovery for proximity-based service in LTE-advanced system IEEE J. Sel. Areas Commun. **33**(1), 55–66 (2015)
4. L. Lei, Z. Zhong, C. Lin, X. Shen, Operator controlled device-to-device communications in LTE-advanced networks. IEEE Wirel. Commun. **19**(3), 96–104 (2012)
5. L. Wei, R. Q. Hu, Y. Qian, G. Wu, Enable device-to-device communications underlaying cellular networks: challenges and research aspects. IEEE Commun. Mag. **52**(6) 90–96 (2014)
6. L. Song, D. Niyato, Z. Han, E. Hossain, Game-theoretic resource allocation methods for device-to-device (D2D) communication. IEEE Wirel. Commun. **21**(3), 136–144 (2014)
7. M.N. Tehrani, M. Uysal, H. Yanikomeroglu, Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions. IEEE Commun. Mag. **52**(5), 86–92 (2014)
8. A. Asadi, Q. Wang, V. Mancuso, A survey on device-to-device communication in cellular networks. IEEE Commun. Surv. Tutorials **16**(4), 1801–1819 (2014)
9. S. Bi, C. Ho, R. Zhang, Wireless powered communication: opportunities and challenges. IEEE Commun. Mag. **53**(4), 117–125 (2015)
10. S. Berger, M. Kuhn, A. Wittneben et al., Recent advances in amplify-and-forward two-hop relaying. IEEE Commun. Mag. **47**(7), 50–56 (2009)
11. K.T.K Cheung, S. Yang, L. Hanzo, Achieving maximum energy-efficiency in multi-relay OFDMA cellular networks: a fractional programming approach. IEEE Trans. Commun. **61**(7), 2746–2757 (2013)

12. D.W.K. Ng, R. Schober, Cross-layer scheduling for OFDMA amplify-and-forward relay networks. IEEE Trans. Veh. Technol. **59**(3), 1443–1458 (2010)
13. M.S. Alam, J.W. Mark, X. Shen, Relay selection and resource allocation for multi-user cooperative OFDMA networks. IEEE Trans. Wirel. Commun. **12**(5), 2193–2205 (2013)
14. S. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, New York, NY, 2004)
15. S.O. Krumke, *Integer Programming. Polyhedra and Algorithms*. Draft, January 4, 2006
16. R. Arab Loodaricheh, S. Mallick, V.K. Bhargava, Energy efficient resource allocation for OFDMA cellular networks with user cooperation and QoS provisioning. IEEE Trans. Wirel. Commun. **13**(11), 6132–6146 (2014)
17. S. Boyd, L. Xiao, A. Mutapcic, Subgradient methods, in *Notes for EE392o Stanford University Autumn*, 2003–2004
18. W. Dang, M. Tao, H. Mu, J. Huang, Subcarrier-pair based resource allocation for cooperative multi-relay OFDM systems. IEEE Trans. Wirel. Commun. **9**(5), 1640–1649 (2010)
19. H. Zhu, J. Wang, Device-to-device communication in cellular networks with fractional frequency reuse, in *Proceedings of 2014 IEEE ICC*, pp. 5503–5507
20. L. Liu, R. Zhang, K.C. Chua, Wireless information transfer with opportunistic energy harvesting. IEEE Trans. Wirel. Commun. **12**(1), 288–300 (2013)
21. D.W.K. Ng, E.S. Lo, R. Schober, Wireless information and power transfer: energy efficiency optimization in OFDMA systems. IEEE Trans. Wirel. Commun. **12**, 6352–6370 (2013)
22. T. Wang, A. Cano, B. Giannakis et al., High-performance cooperative demodulation with decode-and-forward relays. IEEE Trans. Commun. **55**(7), 1427–1438 (2007)
23. D.P. Bertsekas, *Dynamic Programming and Optimal Control*. vol. 1, no. 2 (Athena Scientific, Belmont, MA, 1995)
24. R. Arab Loodaricheh, S. Mallick, V.K. Bhargava, Resource allocation for OFDMA systems with selective relaying and energy harvesting, in *Proceedings of 2014 IEEE VTC (Fall)*, pp. 1–5
25. H.W. Kuhn, The Hungarian method for the assignment problem. Naval Res. Logistic Q. **2**, 83–97 (1955)
26. G. Zhao, C. Yang, G.Y. Li, D. Li, A. Soong, Power and channel allocation for cooperative relay in cognitive radio networks. IEEE J. Sel. Top. Sign. Proces. **5**(1), 151–159 (2011)
27. X. Gong, W. Yuan, W. Liu, W. Cheng, S. Wang, A cooperative relay scheme for secondary communication in cognitive radio networks, in *Proceedings of 2008 IEEE Globecom*, pp. 1–6
28. C. Sun, K.B. Letaief, User cooperation in heterogeneous cognitive radio networks with interference reduction, in *Proceedings of 2008 IEEE ICC*, pp. 3193–3197
29. D. López-Pérez, A. Valcarce, G. Roche, J. Zhang, OFDMA femtocells: a roadmap on interference avoidance. IEEE Commun. Mag. **47**(9), 41–48 (2009)
30. I. Demirdogen, I. Guvenc, H. Arslan, A simulation study of performance trade-offs in open access femtocell networks, in *Proceedings of 2010 IEEE PIMRC*, pp. 151–156
31. S. Kadloor, R. Adve, Relay selection and power allocation in cooperative cellular networks. IEEE Trans. Wirel. Commun. **9**(5), 1676–1685 (2010)
32. T. Al-Khasib, M. Shenouda, L. Lampe, Dynamic spectrum management for multiple-antenna cognitive radio systems: design with imperfect CSI. IEEE Trans. Wirel. Commun. **10**(9), 2850–2859 (2011)
33. J.N. Laneman, D.N.C. Tse, G.W. Wornell, Cooperative diversity in wireless networks: efficient protocols and outage behavior. IEEE Trans. Inf. Theory **50**(12), 3062–3080 (2004)
34. M. Medard, The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel. IEEE Trans. Inf. Theory **46**(3), 933–946 (2000)
35. G. Zheng, S. Ma, K.-K. Wong, T.-S. Ng, Robust beamforming in cognitive radio. IEEE Trans. Wirel. Commun. **9**(2), 570–576 (2010)
36. Z.K.M. Ho, V.K.N. Lau, R.S.K. Cheng, Closed loop cross layer scheduling for goodput maximization in frequency selective environment with no CSIT, in *Proceedings of 2007 IEEE WCNC*, 299–303
37. D. I. Kim, L. B. Le, E. Hossain, Joint rate and power allocation for cognitive radios in dynamic spectrum access environment. IEEE Trans. Wirel. Commun. **7**(12), 5517–5527 (2008)

38. L. Zhang, Y.-C. Liang, Y. Xin, H. Poor, Robust cognitive beamforming with partial channel state information. IEEE Trans. Wirel. Commun. **8**(8), 4143–4153 (2009)
39. S. Mallick, M.M. Rashid, V.K. Bhargava, Joint relay selection and power allocation for decode-and-forward cellular relay network with channel uncertainty. IEEE Trans. Wirel. Commun. **11**(10), 3496–3508 (2012)
40. G. Zheng, K.-K. Wong, B. Otterston, Robust cognitive beamforming with bounded channel uncertainties. IEEE Trans. Signal Process. **57**(12), 4871–4881 (2009)
41. S. Mallick, R. Devarajan, R. Arab Loodaricheh, V.K. Bhargava, Robust resource optimization for cooperative cognitive radio networks with imperfect CSI. IEEE Trans. Wirel. Commun. **14**(2), 907–920 (2015)
42. P.-J. Chung, H. Du, J. Gondzio, A probabilistic constraint approach for robust transmit beamforming with imperfect channel information. IEEE Trans. Signal Process. **59**(6), 2773–2782 (2011)
43. D.R. Cox, N. Reid, Approximations to non-central distributions. Can. J. Stat. **15**(2), 105–114 (1987)
44. I. Hammerstrom, A. Wittneben, On the optimal power allocation for nonregenerative OFDM relay links, in *Proc. 2006 IEEE ICC*, pp. 4463–4468
45. M. Choi, J. Park, S. Choi, Simplified power allocation scheme for cognitive multi-node relay networks. IEEE Trans. Wirel. Commun. **11**(6), 2008–2012 (2012)
46. M. Shaat, F. Bader, Asymptotically optimal resource allocation in OFDM-based cognitive networks with multiple relays. IEEE Trans. Wirel. Commun. **11**(3), 892–897 (2012)

# Fog Computing and Its Applications in 5G

**Longxiang Gao, Tom H. Luan, Bo Liu, Wanlei Zhou, and Shui Yu**

**Abstract** With smartphones becoming our everyday companions, high-quality mobile applications have become an important integral of people's lives. The intensive and ubiquitous use of mobile applications have led to explosive growth of mobile data traffics. To accommodate the surge mobile traffic yet providing the guaranteed service quality to mobile users represent a key issue of 5G mobile networks. This motivates the emergence of Fog computing as a promising, practical and efficient solution tailored to serving mobile traffics. Fog computing deploys highly virtualized computing and communication facilities at the proximity of mobile users. Dedicated to serving the mobile users, Fog computing explores the predictable service demand patterns of mobile users and typically provides desirable localized services accordingly. Stitching above features, Fog computing can provide mobile users with the demanded services through low-latency and short-distance local connections. In this chapter, we introduce the main features of Fog computing and describe its concept, architecture and design goals. Lastly, we discuss on the potential research issues from the perspective of 5G networking.

## 1 Introduction

Smartphones have already become our everyday companions. In 2011, the smartphone shipment worldwide overtook that of PCs for the first time in history, and now the smartphone penetration has reached 75 % in US. It is envisioned by Cisco that the average number of connected mobile devices per person will hit 6.56 in 2020, due to the proliferating use of "Internet of Things" applications, e.g., smart home, smart community, and emerging mobile electronics, e.g., wearable devices.

Smart devices have brought rich computing and communication capability to the palm of our hand. As a result, rich mobile applications are enabled to enhance our day-to-day experiences by enabling productivity, connectivity and achievement

L. Gao • T.H. Luan (✉) • B. Liu • W. Zhou • S. Yu
School of Information Technology, Deakin University, Burwood, Victoria, Australia
e-mail: longx@deakin.edu.au; tom.luan@deakin.edu.au; bo.liu@deakin.edu.au; wanlei.zhou@deakin.edu.au; shui.yu@deakin.edu.au



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_21

of our goals. For example, we may be already addicted to mobile applications everyday for social connectivity and to fulfill our multimedia messaging needs. On each passing day, mobile applications connect with wearable devices to read our heartbeat and track our health conditions, adjust the temperature and light of our room by communicating with smart home facility, coerce us to take that morning walk, offer a different route that will help us avoid the rush hour traffic, and become increasingly intelligent by understanding our mobilities, gestures and social activities. Apparently, from changing the way we communicate to revolutionizing the way we work and live, mobile electronics and applications pervade our daily lives everywhere.

The proliferation and pervasive use of mobile applications inevitably leads to the explosive growth of the mobile data traffic. To accommodate the surge mobile traffic and in the meantime provide guaranteed service quality to mobile users represent the key issue of next generation mobile networks. This motivates the emergence of Fog computing as a promising, practical and efficient solution that extends cloud computing to better serving mobile traffics. The term "Fog computing" was first proposed by Cisco in 2012 [1]. Similar systems typically known as edge computing, such as Cyber Foraging [2], Cloudlets [3] can date back to early 2000.

In this chapter, we introduce why the Fog computing is promising, the main features of Fog computing and describe its concept, architecture and design goals. Then we demonstrate a case study on how the Fog computing can improve the network performance in 5G environment, followed by a discussion on the potential research issues from the perspective of 5G networking.

## 2 Fog Computing Architecture

Fog computing extends the cloud-based Internet by introducing an intermediate layer between mobile devices and cloud, aiming at the smooth, low-latency service delivery from the cloud to mobile. This accordingly leads to a three hierarchy Mobile-Fog-Cloud architecture as depicted in Fig. 1.

The intermediate Fog layer is composed of geo-distributed Fog servers which are deployed at the edge of networks, e.g., parks, bus terminals, shopping centers, etc. Each Fog server is a highly virtualized computing system, similar to a light-weight cloud server, and is equipped with the on-board large-volume data storage, compute and wireless communication facility. The role of Fog servers is to bridge the mobile users and cloud. On one hand, Fog servers directly communicate with the mobile users through single-hop wireless connections using the off-the-shelf wireless interfaces, such as WiFi and Bluetooth. With the on-board compute facility and pre-cached contents, they can independently provide pre-defined service applications to mobile users without assistances from cloud or Internet. On the other hand, the Fog servers can be connected to the cloud so as to leverage the rich functions and application tools of the cloud. The next section describes some typical examples of Fog computing in details.

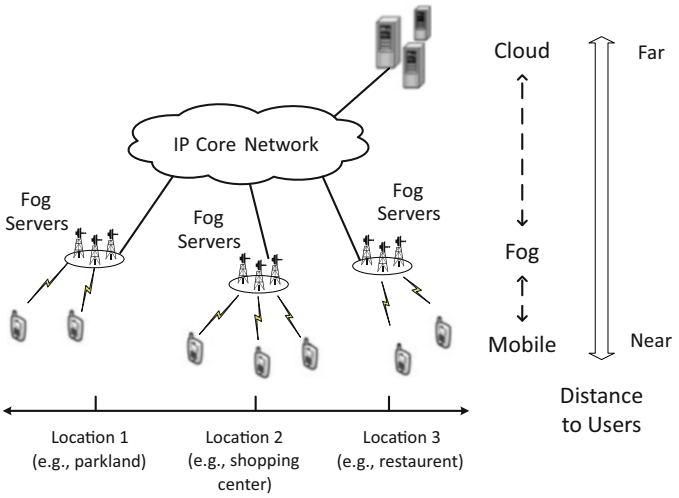


**Fig. 1** Fog computing architecture

To summarize, the purpose of Fog computing is to place a handful of compute, storage and communication resources in the proximity of mobile users, and therefore to serve mobile users with the local short-distance high-rate connections. This overcomes the drawback of cloud which is far to mobile users with elongated service delays. Therefore, the fog is interpreted as "the cloud close to the ground" [1].

## 3 Why Fog Computing?

Nowadays, the evolving of Internet has shown two obvious trends. First, the could-based architecture is adopted to host major applications and storage. As predicted by Cisco Global Cloud Index, the global cloud traffic will account for more than three-fourths of total data center traffic by 2018. Second, the Internet users have shifted predominantly from using desktop computers to smartphones and tablets. With cloud becoming the overarching approach for service delivery and information retrieval, and mobile users becoming the major service consumers, the seamless interconnection of cloud computing and mobile applications therefore represents a key issue in the 5G mobile networks and motivates the emergence of Fog computing.

To show the rationale of Fog computing, in what follows we take a retrospect study by revisiting the design of cloud-based Internet and service requirements of mobile users.

### 3.1 Global and Local Information

The cloud computing represents an efficient and scalable centralized solution for information management and distribution. It is efficient to serve the information requests from the traditional desktop users. To be specific, the desktop users, typically accessing Internet at homes and offices are often interested in the information which is irrelevant to their locations, such as the world news, stock market at different cities or countries, to name a few. We refer to such information as the *global information*. As a contrast, we refer to the location-based service information related to the location of users as the *local information*. Cloud computing favors desktop users with an optimized approach for serving the global information services. With a scalable and efficient approach to store and manage the information originated at different locations of the world and using a static public IP, cloud computing conveniently distribute the cached global information from a remote central server to desktops worldwide.

The mobile users, however, have distinguished service requirements from the desktop users. This requires the current cloud-based Internet to be modified accordingly to cater to the specific service requirements of mobile users. In specific, different from desktop users, mobile users, particularly smartphones, are typically in the outdoor environments. This makes their service requirements closely related to their current locations. In other words, mobile users are more interested in the *local information* around them. For example, a mobile user in a shopping center tends to be interested in the sales, open hour, restaurants and events inside the attended shopping center; such information become useless once he/she leaves the shopping center. In another example, a traveller to a city would seek for information on the places of interest, local news and weather conditions of the specific city, while such information of other places is useless. The massive demand of location-based mobile services is also reported in [4].

### 3.2 Physical and Communication Distance

The cloud-based Internet can be inefficient to serve the local information desired by mobile users. As a motivating example shown in Fig. 2a, assuming that a mobile user inside a shopping center intends to retrieve flyers of stores within the shopping center. To do this using the cloud-based Internet, the stores may need to first upload their flyers to a remote cloud server over Internet, and then direct mobile users to retrieve the desired information from the remote cloud server. In other words, although the *physical distance* between the mobile user (destination) and stores (original source) is short, using the remote cloud as the information depot, the actual *communication distance* can be far, e.g., from the cloud server to mobile user in this example.

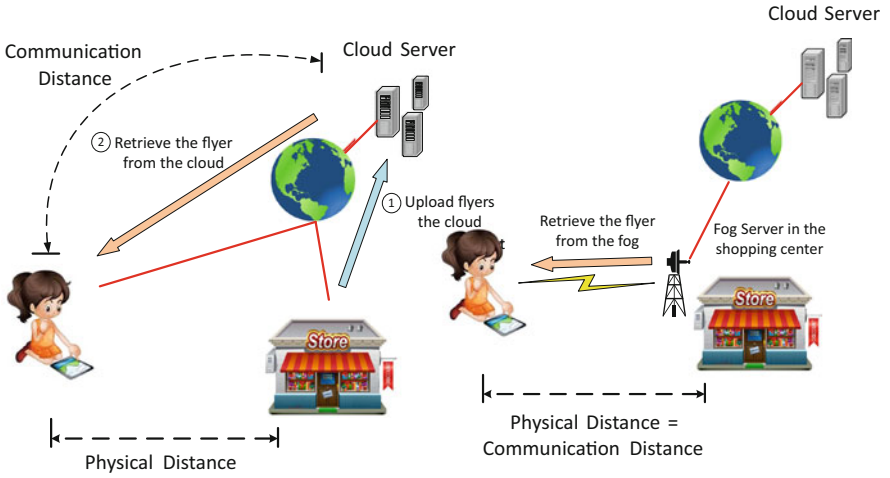


**Fig. 2** Example: download the flyer of a nearby store. (**a**) Retrieving the flyer from the cloud. (**b**) Retrieving the flyer from the fog

The Fog computing paradigm represents a practical and efficient solution to resolve the mismatch between physical and communication distances. As a remedy shown in Fig. 2b, a Fog server can be deployed inside the shopping center and to distribute the local store flyers to mobile users. As such, the physical distance is equal to the communication distance and users can acquire low-latency desirable services.

By minimizing the communication distance, the Fog computing therefore brings the following two advantages:

- **To mobile users**: compared to cloud, the Fog computing can provide enhanced service quality with much increased data rate and reduced latency and response time. Moreover, by reducing the bandwidth cost of data transmission in the backbone, the users can also be benefited from the reduced service cost.
- **To network**: by avoiding the duplicated back and forth traffic between cloud and mobile user, not only the backbone bandwidth can be significantly saved, the energy consumption of core networks can also be greatly reduced, which contributes to the sustainable development of networking.

## 4 Components of Fog Computing

The Fog thus behaves as a surrogate of Cloud or a private Cloud at the user's premises. This enables Fog servers to be more efficient to handle the localized computation requests. Therefore, Fog computing targets to deliver the localized and

location-based service applications to mobile users. In what follows, we showcase some examples of Fog computing implementation from this perspective, and discuss on the features of a Fog server as a comparison to Cloud server.

## *4.1 Exemplary Implementations*

### 4.1.1 Shopping Center

Assuming that a number of Fog servers are deployed inside a multi-floor shopping center, which collectively form an integrated localized information system. The Fog servers at different floors can pre-cache floor-related contents, such as the layout and ads of stores on a particular floor. The Fog servers can deliver engaged services including indoor navigation, ads distribution and feedback collections to mobile users through WiFi.

### 4.1.2 Parkland

The Fog computing system can be deployed in the parkland to provide localized travel services. For instance, Fog servers can be deployed at the entrance and other important locations of the park. The Fog server at the entrance can pre-cache information including park map, travel guide and local accommodations; other Fog servers at different locations inside the park can be incorporated with sensor networks for environment monitoring and provide navigation to travellers. By connecting the Fog servers to the park administration office and cloud, the Fog servers can be used as an information gateway to send timely alerts and notifications to travellers.

### 4.1.3 Inter-State Bus

Greyhound has launched "BLUE" [5], an on-board Fog computing system over inter-state buses for entertainment services. As an example illustrated in Fig. 3, a Fog server can be deployed inside the bus and provides on-board video streaming, gaming and social networking services to travellers using WiFi. The on-board Fog server connects to the Cloud through cellular networks to refresh the pre-catched contents and update application services. Using its computing facility, the Fog server can also collect and process user's data, such as number of travellers and their feedbacks, and report to cloud.

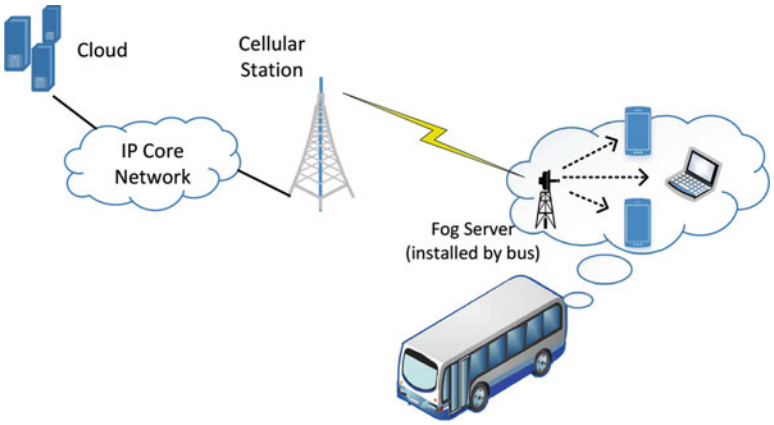


**Fig. 3** On-board fog computing system

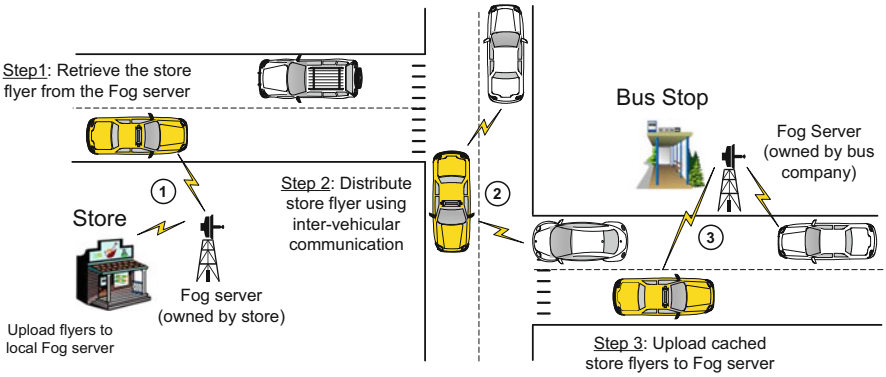


**Fig. 4** Fog computing for content distribution in vehicular ad hoc networks

### 4.1.4 Vehicular Fog Computing Networks

Luan et al. [6] present the application of Fog computing as an integrated large-scale network for localized content disseminations. Figure 4 shows a motivating scenario. Assuming that a store installs a Fog server at its parking lot with the purpose to distribute the store flyer. In step 1, the store uploads flyers to the Fog server via wireless connections, and the Fog server distributes the flyers wirelessly to vehicles driving through its coverage using wireless communications. With the vehicle moving to different locations, it can further disseminate the cached flyers to other vehicles using wireless communications, as depicted in Step 2. In Step 3, the flyers can also be retrieved and cached at other Fog servers deployed at different locations, e.g., bus stop, and further propagated in the network.

## 4.2 Comparison to Cloud Computing

Fog computing is featured by the dedication to serving the localized and location-based applications. To this end, a Fog server manages its on-board resources to fully explore the location information and predictable user demand with the following functions.

- **Wireless**: Fog computing is dedicated to serving the mobile users. Each Fog server typically has limited wireless coverage, e.g., 200 m using WiFi, and directly interacts with mobile users using the single-hop wireless connections.
- **Local Services**: Fog computing is dedicated to serving the localized information and providing location-based service applications. For example, the Fog computing system deployed in a specific park only provide the navigation services within the park.
- **Distributed Management**: A Fog computing system may typically be deployed and managed by the local business, with the purpose to deliver designated contents and services to specific user groups.

Using the example in Fig. 4 for illustration, the Fog servers provide localized content distribution using wireless communications, which matches the first two features. The Fog server deployed nearby the store may be installed and managed by the store owner for the distribution of store flyers; the Fog server at the bus stop may be managed by the bus company for the distribution of bus information, e.g., bus schedules, safety manual, etc., to mobile users waiting for the bus. The Fog computing system in [6] is therefore distributedly constructed with Fog servers distributed installed and managed by different entities to serve their own purposes, which matches the third feature.

Table 1 summarizes the differences between Fog computing and Cloud computing.

By targeting to different user groups at different locations, Fog computing extends Cloud computing to better serve local mobile traffics. As the system architecture shown in Fig. 1, the Fog servers deployed at different locations would be used to deliver engaged services specified by their owners. The Fog servers at different locations can connect to the same cloud and form an integrated Fog computing system in a wide region.

## 4.3 Components of Fog Computing

### 4.3.1 Storage

In a predefined service area, a Fog server predicts the mobile user's demand on information and pre-cache the desirable information accordingly using a proactive way in its storage. Such information can be either retrieved from the Cloud or uploaded by its owner. For example, the Fog server installed at a restaurant can

**Table 1** Comparison of fog computing and cloud computing

| | Fog computing | Cloud computing |
|---|---|---|
| Target user | Mobile users | General Internet users |
| Service type | Limited localized information services related to specific deployment locations | Global information collected from worldwide |
| Hardware | Limited storage, compute power and wireless interface | Ample and scalable storage space and compute power |
| Distance to users | In the physical proximity and communicate through single-hop wireless connection | Faraway from users and communicate through IP networks |
| Working environment | Outdoor (streets, parklands, etc.) or indoor (restaurants, shopping malls, etc.) | Warehouse-size building with air conditioning systems |
| Deployment | Centralized or distributed in reginal areas by local business (local telecommunication vendor, shopping mall retailer, etc.) | Centralized and maintained by Amazon, Google, etc. |

pre-cache the menu of the restaurant and dish recipes to serve the mobile users inside the restaurant. In another example, the Fog servers deployed in the airport can pre-cache the flight and local transportation information which is desirable to travellers in the airport. Therefore, the key design issue of Fog computing is to predict the user's demand and proactively select the contents to cache in the geo-distributed Fog servers based on the specific locations.

The Content Delivery Network (CDN) [7] represents the most mature catch networks and extensively pursued in both academic and industry. CDN is the Internet-based cache network by deploying cache servers at the edge of Internet to reduce the download delay of contents from remote sites. CDN mainly targets to serve traditional desktop Internet users, which have much broader interests and blur service demands that are more difficult to predict than those of mobile users. With precise service region, Fog computing has more clear target users of specific service demand. It is thus key for Fog servers to explore this feature to fully utilize its storage and computing resources to provide the most desirable services to mobile users.

Similar to Fog computing, the Information Centric Network (ICN) [8] is also a wireless cache infrastructure which provides content distribution services to mobile users with distributed cache servers. Different from the cache servers in ICN, the Fog servers are intelligent computing unit. Therefore, the Fog servers are not only used for caching, but also as a computing infrastructure to interact with mobile users and devices for real-time data processing. The Fog servers can be connected to the cloud and accordingly utilize the extensive computing power and big data tools for rich applications other than content distribution, such as internet of things, vehicular communications and smart grid applications [1].

Baştuğ et al. [9] also show that the information demand patterns of mobile users are predictable to an extent and propose to proactively pre-cache the desirable information before users request it. The social relations and device to device communications are leveraged. Unlike Fog computing, the proactive caching scheme in [9] is not explicitly used to serve local information services. As a more broad and generic paradigm, Fog computing can incorporate the proactive caching framework as described in [9].

### 4.3.2 Compute

A salient feature that differentiates Fog computing from the traditional cache networks is that Fog servers are intelligent compute system. This allow a Fog server to autonomously and independently serve local computation and data processing requests from mobile users. Satyanarayanan et al. [10] shows the applications of Fog computing in the cognitive applications. In another example, a Fog server inside the shopping mall or parkland can maintain an on-board geographic information system, and provide the real-time navigation and video streaming to connect mobile users.

Bridging the mobile and Cloud, a Fog server can also be conveniently used to collect the environmental data or demographic data from mobile users at the deployed spot, and transport the collected big data to Cloud for in-depth data analysis; the results can be provided to third party for strategic and valuable insights on business and government event planning, execution and measurement.

Despite of the high computing power, the Cloud is faraway from mobile users and can hardly support real-time computing intensive applications due to the bandwidth-constrained IP networks. The demand of real-time resource-intensive mobile applications, e.g., cognitive and internet-of-things applications, motivates the design of ubiquitous edge computing system [10, 11]. Cloudlets [3, 10] adopt the same framework of Fog computing, in which a Cloudlet server, similar to the Fog server, is deployed in the proximity of mobile users and processes the computing requests of mobile devices at real-time for video streaming and data processing. A comparison of processing delays using Cloudlets and Amazon Clouds is shown in http://elijah.cs.cmu.edu/demo.html. Transparent computing [11] is a highly virtualized system, which targets to develop the computing system transparent to users with cross-platform and cross-application support.

The Fog computing is a generic platform for edge computing and focuses on the localized service applications and computation requests. The prototype and techniques in [10, 11] can be incorporated in Fog computing framework.

### 4.3.3 Communication

Fog server can equip with different wireless interfaces, e.g., WiFi, Bluetooth and visible light communications [12] according to the specific application scenarios. The Fog computing differs from traditional radio access networks, e.g., WiFi and Femtocell networks, in two important ways.

**Cross-Layer Design** Unlike traditional WiFi access points, the Fog server manages an autonomous, all-inclusive network by providing both service applications and wireless communications to mobile users in the coverage. Therefore, a Fog server can work without Internet connections as that in [6]. Note that the Fog computing tailors its applications based on the specific deployment location and environment, and therefore is highly service-oriented. To this end, a Fog server can manage all the communication layers and effectively enables the cross-layer design [13] to provide the best service quality to users. For example, as in "BLUE" [5], a Fog server can cache a number of videos and deliver Youtube-like video streaming services to mobile users in the proximity. In this case, based on the context, wireless channel and video popularity information, the video services can be conveniently adapted towards the optimal performance via cross-layer adjustifications.

**Predictable Location-Based Service** The key of Fog computing is to provide the localized network and information applications to mobile users, whereas the traditional radio access networks focus on the provision of Internet applications and global information. With this distinguished feature, the design of Fog computing communications needs to consider the specific deployment environment and the features of mobile users in the considered scenario. For example, a Fog computing system deployed in the shopping mall needs to address the diverse mobilities of users, whereas the similar system deployed in the inter-state bus [5] only needs to consider static on-board passengers.

## 5 Case Study: Hybrid Data Dissemination in Fog Computing

In this section, we demonstrate a case study based on Fog computing and show how Fog computing can be incorporated into the 5G network towards improved performance to mobile users.

5G technique will make streaming applications becoming more and more popular, however the long latency may severely affect the user experience and is not tolerable. To address this issue, Fog computing can move Cloud services from remote Internet to the edge of networks and makes streaming content much closer to mobile users, which significantly decreases the streaming latency. On the contrary, the data dissemination from Cloud to every Fog servers can be expensive, which may takes the hudge 5G bandwidth resource. In addition, note that since the majority of streaming applications are video based, such as movies, teleplay and product advertisement, such contents are not always necessary to be strictly up-to-date and 1 or 2 days latency is affordable. Therefore, the Fog computing and delay tolerant network (DTN) techniques can be combined together to improve the performance of 5G network.

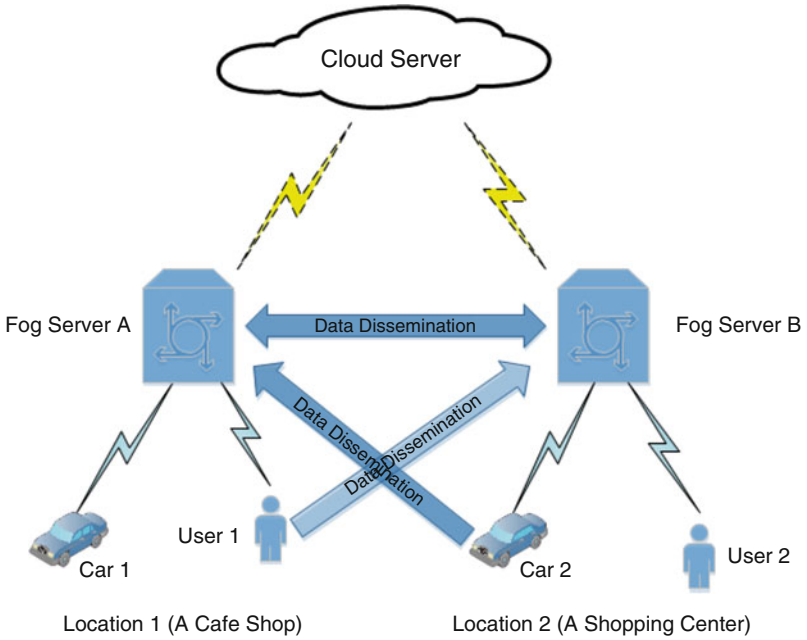


**Fig. 5** Data dissemination in fog computing based on delay tolerant network technique

## *5.1 System Model*

To have an efficient data dissemination in Fog computing, DTN is used to offload data among Fog servers. For example, in Fig. 5, if Fog server A has some small content, such as a new video ads, but Fog server B does not. In previous, Fog server B needs to get the update from a Cloud server directly. With DTN technique, user 1 could download this content when he has a coffee in this cafe shop. After a couple of hours, he goes to shopping center for shopping. When he moves into the transmission range of Fog server B, the content storied in user 1's mobile device is automatically upload to Fog server B and this store-carry-forward process is completed. For a large content, e.g. a high definition movie, transmitted from Fog server B to Fog server A, vehicle based DTN is used. In this example, if car 2 is parked in shopping center and within the Fog server B's transmission range, it downloads this movie into its local storage. Once this vehicle moves to the transmission range of Fog server A, this movie is uploaded to it and this data dissemination is completed.

In addition of the above DTN based data dissemination between mobile user and Fog server, direct data dissemination from Fog server to Fog server based on DTN technique is also available. As shown in Fig. 6, there is an on-board Fog server on a tourism bus, where all passengers on this bus can access its Fog server to watch movies or play games. If a passenger wants to find some new stuff, such as "just in"

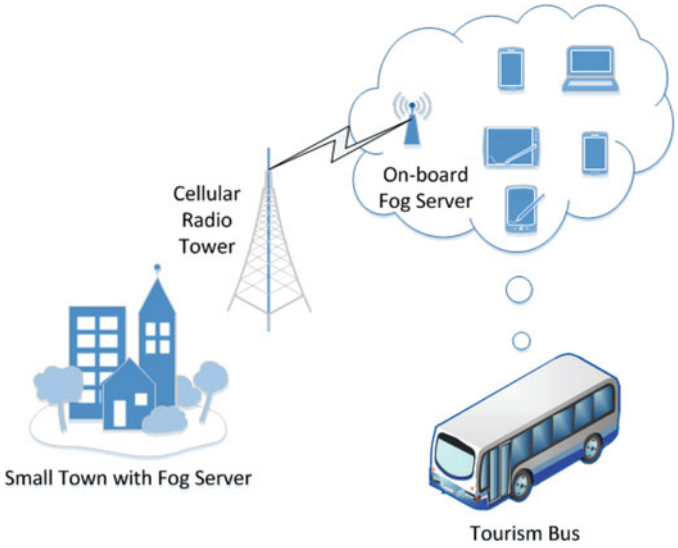


**Fig. 6** Data dissemination between fog servers

news, Fog server can use the cellular network to retrieve this content immediately. When this bus travel to a small town along its route, it can synchronize its content with the Fog server located in this small town to update both servers' content list. If the Fog server in this small town has some content while other small towns along the route do not have, this tourism bus could download these contents into its local storage and carry to those small towns where they need these contents.

With the above DTN based data dissemination techniques, we propose a hybrid data dissemination model, as shown in Fig. 7. This hybrid model not only includes the normal data dissemination between Cloud servers and Fog servers, but also involves large amount low-cost DTN based data disseminations, which can be used between Fog servers and mobile users and among Fog servers. To organize these data dissemination, we re-identify the function of Cloud servers. In this model, the main function of Cloud server is to act as the "control plane" to determine the Fog server needed to be updated with the required content and control data dissemination process, as shown in Fig. 8. Fog servers and part of Cloud servers are treated as "data plane" to provide data dissemination service.

This data dissemination model has three components, namely as *Data Structures*, *Protocol Messages* and *Algorithms*. Data structures use tables to store key information which is used to determine the path of data dissemination. Protocol messages use various tapes of messages to discover content and mobile devices associated with a Fog server, exchange content list, and other tasks to learn and maintain accurate information about the network. Algorithms are used to calculate the best data dissemination path.

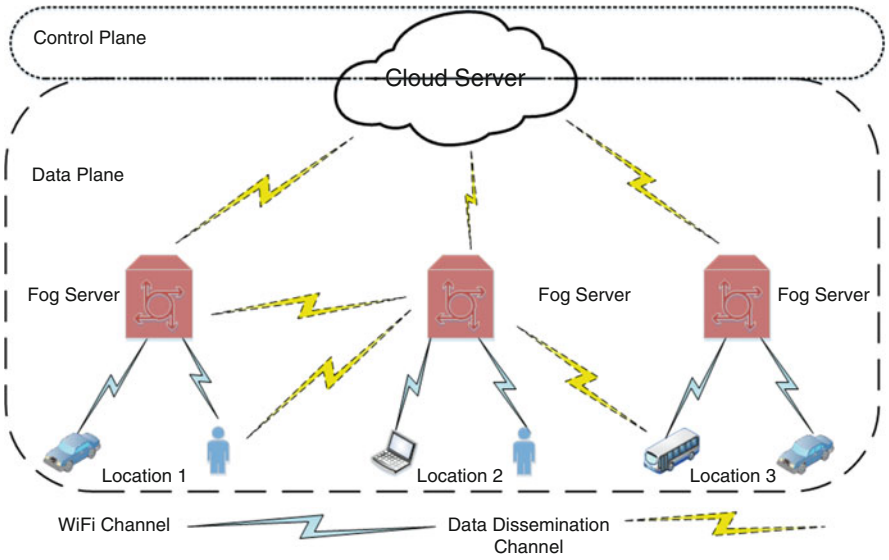


**Fig. 7** Hybrid data dissemination model in fog computing

Cloud server in this model needs to have an overall information and its data structures include the following tables:

- Fog Server List Table: a table to record all Fog servers, which are managed by Cloud server. This table includes Fog server's ID, content ID in each of Fog servers, mobile device ID associated with each of Fog servers.
- Global Content List Table: a table to record all public contents (not include these private content created by Fog server owner) in Fog servers or supposed to be in Fog servers. This table includes content ID, the size of each contents, Fog server's ID (for these Fog servers who have this content), date of update, validation time.
- Mobile Devices's Movement Pattern Table: a table to record mobile device's ID, Fog server ID (whom mobile device linked before), linked time, social attribute, geographic movement pattern.

A Fog server needs to have a table to record content ID, the size of this content, mobile device's ID which linked with this Fog server, the linked duration of this mobile device. For mobile devices, they need to record the content ID which they carry on, their movement path, time duration with a Fog server and its ID.

In order to collect and exchange the above information, several data messages are used in this model.

- Hello Message between Fog Servers and Cloud Server: an update message from a Fog server to its Cloud provider, which includes its content ID and associated mobile devices ID. This is a triggered message.

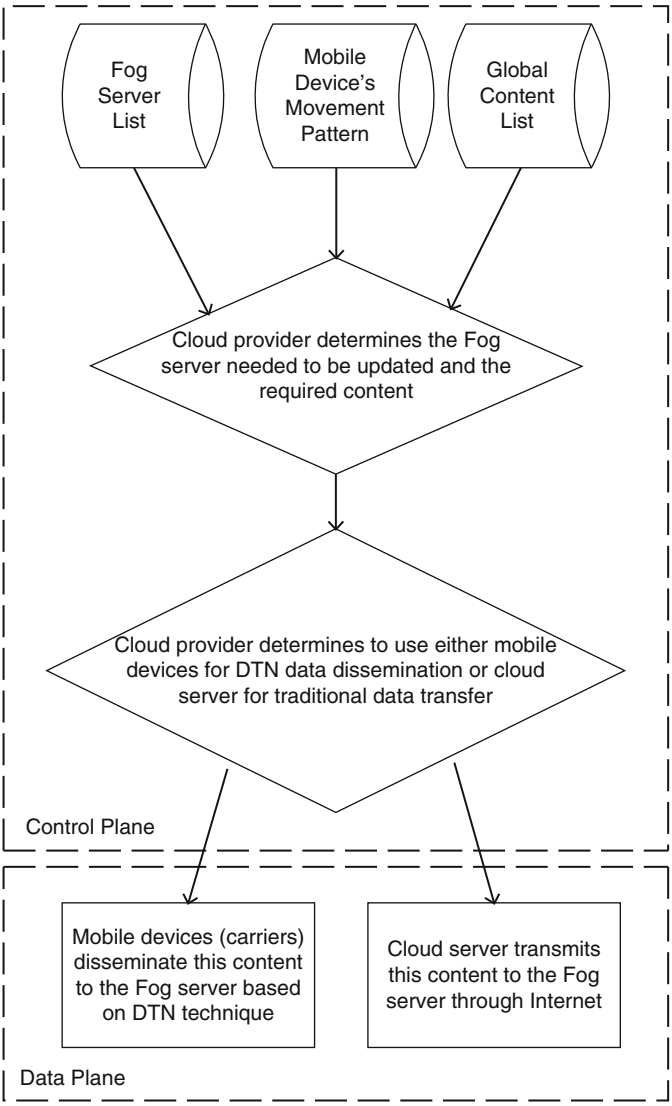


**Fig. 8** Control plane and data plane in this model

- DTN Data Dissemination Request Message: a message sent from Cloud provider to a Fog server. When a Cloud provider determines there is a content need to be updated for Fog server A, if Fog server B has this content and its associated mobile device has potential move to Fog server A, a DTN data dissemination request message is sent to Fog server B to ask it to disseminate the content through that mobile device.

- DTN Data Dissemination Accept Message: a message sent from Fog server to its Cloud provider. Once a Fog server received a DTN data dissemination request message and it sends the content to the corresponding mobile device, it sends this DTN data dissemination accept message to its Cloud provider to confirm that this content has been sent out.
- DTN Data Dissemination Decline Message: a message sent from Fog server to its Cloud provider. When a Fog server received a DTN data dissemination request message, it's associated carrier (mobile device), for some reason, does not receive the complete content before this carrier leave the current Fog server. In this case, the DTN data dissemination decline message is sent back to the Cloud provider.
- DTN Data Dissemination Acknowledgement Message: a message sent from Fog server to its Cloud provider. When a Fog server receives the assigned content from carrier, it sends this acknowledgement message to Cloud provider to confirm that this DTN data dissemination is completed.

Algorithms together with other operations in this hybrid data dissemination model are described in the following sub-section.

## *5.2 Data Dissemination*

Hybrid data dissemination is determined by control plane, as shown in Fig. 8, where Cloud provider in the control plane has the global information to control the data plane. The main data flow control algorithm conducted by Cloud provider is illustrated in Algorithm 1, where Cloud provider checks its global Fog server and content lists to determine the Fog server that needed to be updated, and the required content. It also checks which Fog server has this content. If none of Fog server has this content, Cloud provider sends updated content to that Fog server directly by using traditional Cloud based techniques, such as broadband and cellular networks [14]. Otherwise, the DTN based data dissemination is applied by using Algorithm 2 to choose mobile devices, which are connecting with these selected Fog servers, as carriers to provider DTN based data dissemination services.

The carriers selection is based on their delivery time and delivery probability (Algorithm 3). A pre-determined content delay threshold, $T_{delay}$, which is an attribute of this content and can also be treated as content delivery priority, is provided by the Cloud provider. Only those mobile devices (carriers) with a shorter delivery time compared with the pre-determined delay time, and a higher delivery probability are selected as potential DTN based data dissemination carriers.

Once these carriers are chosen, Cloud provider sends DTN data dissemination request message to each of the selected Fog servers to ask them send the required content to the Fog server which is needed to be updated. If the content is transmitted to the carried successfully, a DTN data dissemination accept message is sent back

**Algorithm 1** Data Flow Control Algorithm

**Step 1:** Cloud provider compares its "Global Content List" table with "Fog Server List" table to determine which Fog server is needed to be updated. In this case, Fog sever $F_d$ is determined and content $C$ is needed to be updated.
**if** There is no other Fog server has this content **then**
  Cloud provider sends this content to the Fog server directly, which is the same as traditional Cloud service and this dissemination process is finished.
**else**
  Move to the next step
**end if**
**Step 2:** Cloud provider determines a list of Fog servers, $< F_{c1}, F_{c2}, F_{c3}, \cdots >$, which have this content.
**Step 3:** Algorithm 2 is used to select $n$ most suitable carrier, $Carrier_n$, to provide this DTN dissemination service.
**if** $n$ greater than 0 **then**
  Move to the next step
**else**
  Cloud provider sends this content to the Fog server directly, which is the same as traditional Cloud service and this dissemination process is finished.
**end if**
**Step 4:** Cloud provider sends "DTN Data Dissemination Request" message to each of selected Fog servers to ask them send the content $C$ to $F_d$ by using the carrier (mobile device) determined in **Step 3**.
**Step 5:** Once a Fog server receives "DTN Data Dissemination Request" message, it sends the content $C$ along with the destination, $F_d$, to the selected carrier.
**if** Content $C$ is transferred to the selected carrier completely **then**
  This Fog server sends the "DTN Data Dissemination Accept" message to its Cloud provider and move to the next step.
**else**
  This Fog server send the "DTN Data Dissemination Decline" message to its Cloud provider. When Cloud provider receives this message, it repeats *Step 3* to get the "n+1" Fog server, if it has, and continue from *Step 4*.
**end if**
**Step 6:** Once $F_d$ receives the content $C$, it sends "DTN Data Dissemination Acknowledgement" message to Cloud provider.
**Step 7:**
**if** Cloud provider receives the "DTN Data Dissemination Acknowledgement" message within a pre-defined period, $T_{delay}$, in Algorithm 2 **then**
  It updates "Fog Server List" and "Global Content List" tables, and this data dissemination is finished
**else**
  It repeats from the *Step 1*
**end if**

to Cloud provider confirming this content has been sent out. Otherwise, a DTN data dissemination decline message is sent out. For example, a mobile device (carrier) left the Fog server's coverage area.

When the Fog server, who needs this content, receives the content, it sends a DTN data dissemination acknowledgement message to Cloud provider to confirm it has received the content and this DTN based data dissemination process is finished.

**Algorithm 2** DTN Data Dissemination Carrier Selection Algorithm

**Step 1:** Cloud provider determines the affordable delay time, $T_{delay}$, of this content. $T_{delay}$ Cloud be treated as the priority of this content.

**Step 2:** Cloud provider checks its "Fog Server List" and "Mobile Devices's Movement Pattern" tables to find the list of mobile users accessed $F_d$ before, $MobList_d(i)$, and their average connection time with $F_d$, $Time_d(i)$, where $i$ is the ID of connected mobile device.

**Step 3:** These mobile devices with a short connection time are filtered out, as they are not able to upload the content to the Fog server:

**for** each of mobile device in $MobList_d(i)$ **do**

**if**

$$\frac{Size_C}{Speed_i} > Time_d(i)$$

**then**

This mobile device is filtered out from $MobList_d(i)$ and a new list $MobList'_d(r)$ is formed, where $r$ is the number of mobile device meeting the above condition

**end if**

**end for**

**Step 4:** $MobList'_d(r)$ is further classified into two categories, *scheduled* and *non-scheduled* visit lists. Scheduled visit list stores these mobile devices which are pre-determined to visit a particular Fog server, such as airport shuttle bus. The rest of filtered mobile devices are classified into the non-scheduled visit list.

**Step 5:** For scheduled mobile devices, *Si*, as long as its delivery time, which is the time from now to its next scheduled visit time, is within the $T_{delay}$, it is added into the DTN data dissemination carrier list, $< Carrier_{S1}, Carrier_{S2}, \cdots, Carrier_{Sx} >$, where $x$ the total number of carriers selected to add into the carrier list.

**Step 6:** Non-scheduled mobile device list, *NSi*, is re-ordered by mobile devices' delivery probability to $F_d$ based on Algorithm 3. Cloud provider select the top $y$ mobile devices according their delivery probabilities to add them into the DTN dissemination carrier list, where the number of $y$ is the largest number to satisfy the following condition:

$$\frac{\sum_{i=1}^{x} DeliTime_{Si} + \sum_{i=1}^{y} DeliTime_{NSi}}{x + y} < T_{delay}$$

**End:** Now $x + y$ mobile devices are selected as carriers to provide DTN data dissemination service.

Otherwise, Cloud provider needs to re-select mobile nodes as carriers or directly sends the content using traditional method. Detailed hybrid data dissemination processed are illustrated in Algorithms 1–3, and notations used in these three algorithm are explained in Table 2.

## 6 Future Research Topics of Fog Computing in 5G

Based on the Mobile-Fog-Cloud hierarchy shown in Fig. 9, we envision potential research directions from the communication efficiency's viewpoint as follows.

**Algorithm 3** Mobile Device Delivery Probability

**Step 1:** For each of mobile devices, $m$, Cloud provider collects its contact frequency with $F_d$, $ConFre_m$, geographic locations and visit times of the three most recently visited Fog servers, $Loc_m < Lan, Lon, T >$.
**Step 2:** Based on the three most recent visited history, $Loc_m < Lan, Lon, T >$, and real distance from the map, average movement speed and direction of mobile device $m$ could be generated as $Speed_m$ and $Direction_m$.
**Step 3:** The expected delivery time from mobile device $m$ to Fog server $F_d$, $DeliTime_m$, is calculated by using $Speed_m$, $Direction_m$ and the geographic distance between both them.
**if** $DeliTime_m > T_{delay}$ **then**
The delivery probability of this mobile device, $DeliProb_m$, is marked as 0 and this algorithm is finished
**else**
Move to next step
**end if**
**Step 4:** Assume there are $n$ suitable mobile devices left in this step. The overall delivery probability of mobile device $m$ is calculated as:

$$DeliProb_m = \frac{ConFre_m}{\sum_{i=1}^{n} ConFre_i} \times (1 - \frac{DeliTime_m}{\sum_{i=1}^{n} DeliTime_i})$$

and each of them is added into the delivery probability list, $DeliProbList[n]$.
**Step 5:** Sort $DeliProbList[i]$ in ascending order
Set $u = 1, j = n$
**while** $u \leq n$ **do**
**while** $j > u$ **do**
**if** $DeliProbList[j-1] > DeliProbList[j]$ **then**
swap($DeliProbList[j-1]$, $DeliProbList[j]$)
**end if**
$j--$
**end while**
$u++$
**end while**
This delivery probability list is ready to be used for Algorithm 2.

## 6.1 Communications Between Mobile and Fog

Note that a Fog server manages 3-D resources including storage, computing and communication. The service quality acquired by users relies on the collective performance of resource utilization from all the three dimensions. Moreover, as Fog computing typically provides pre-defined application services and targets to specific user groups, the service-oriented resource allocation customized to the specific deployment environments is thus necessary. For example, considering the on-board Fog computing system inside the inter-state bus as in Fig. 3, three types of traffics may coexist including video streaming, gaming and web surfing delivered through the same Fog server. As such, a cross-layer MAC design at the Fog server can be devised based on the application's information. Considering that Fog servers have limited storage and deliver limited localized services only, another key design

issue is how to optimally select the desirable information contents to cache at each Fog server and determine the appropriate service applications which cause the least service failure rates to mobile users. The solution needs to consider the predictable pattern of mobile service requests, available storage and compute power of a Fog server.

The Fog computing can also be incorporated with the 5G cellular networks. In this case, by making the cellular base station a Fog server with on-board storage and compute facility, the entire Fog system can provide greater coverage and dedicated services to cellular users.

## 6.2 Communications Between Fog and Cloud

The cloud performs two roles in the integrated Fog computing system. First, the cloud is the central controller of Fog servers deployed at different locations. With each Fog server focusing on the service delivery to mobile users at specific locations, the cloud manages and coordinates the geo-distributed Fog server clusters at different regions. Second, the cloud is the central information depot. The Fog servers at different locations select the information contents from the cloud and then deliver the copied contents from its cache to the mobile users. With above two roles, the design goal of the communications between fog and cloud can be twofold: (1) how to enable the reliable and scalable control of Fog servers at the cloud; and (2) how to develop the scalable data routing scheme from cloud to Fog server for content updates.

Note that the dual functions of cloud as stated above well match the architecture of a software-defined networking (SDN) [15, 16], which decouples the traffic routing to the control plane and data plane. It is thus promising to apply the SDN scheme for the control of Fog computing.

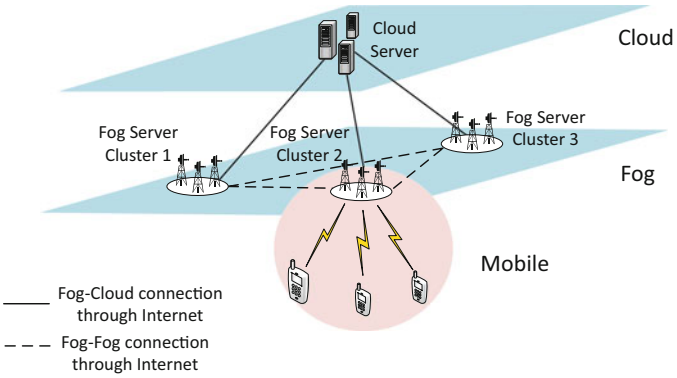


**Fig. 9** Mobile-fog-cloud architecture of fog computing

**Table 2** Notation used in Sect. 5.2

| Notation | Definition |
|---|---|
| $F_d$ | A fog server needed to be updated (the destination of data dissemination) |
| $C$ | A content needed to be updated and disseminated |
| $F_i$ | The fog server $i$ |
| $Carrier_i$ | A content carrier (mobile device) $i$ to provide DTN based data dissemination service |
| $T_{delay}$ | The maximum affordable delay time of a content. It is a time period from now to its must updated time, such as a shopping center promotion video must be released on every Wednesday |
| $MobList_d(i)$ | The *ith* mobile device (carrier) in the mobile list attached with fog server $d$ |
| $Time_d(i)$ | An average connection period between mobile device $i$ and fog server ($d$). It is a maximum time window used to upload/download a content to a fog server |
| $Size_C$ | The size of content $C$ |
| $Speed_i$ | The wireless transmission speed of mobile device $i$ |
| $Loc_i < Lan, Lon, T >$ | The geographic location vector of mobile device $i$ visited at latitude $Lan$ and longitude $Lon$ on time $T$ |
| $Direction_i$ | The expected movement direction of mobile device $i$ |
| $DeliTime_i$ | An average content delivery time of Carrier $i$ to the destined fog server |
| $ConFre_m$ | A contact frequency between the mobile device $m$ and its destined fog server |
| $DeliProbList[n]$ | A list to store all carriers (mobile devices) delivery probability to the destined fog server |

### *6.3 Internet-of-Things Applications*

As Fog servers are deployed at the physical spot close to mobile users and can be equipped with sensors, it is convenient to incorporate the Fog computing with the Internet-of-things applications. Bonomi [1] and Stojmenovic and Wen [17] present the examples of adopting Fog computing in the applications of smartgrid, vehicular networks and sensor networks.

## 7 Conclusion

This book chapter introduced the Fog computing under 5G environment.

This article presents Fog computing, a new networking frontier dedicated to serving mobile users. By deploying reserved compute and communication resources at the edge, Fog computing absorbs the intensive mobile traffic using local fast-rate connections and relieves the long back and forth data transmissions among cloud and mobile devices. This significantly improves the service quality perceived by mobile users and, more importantly, greatly save both the bandwidth cost

and energy consumptions inside the Internet backbone. Therefore, Fog computing represents a scalable, sustainable and efficient solution to enable the convergence of cloud-based Internet and the mobile computing. The purpose of this article is to investigate on the major motivation and design goals of Fog computing from the networking's perspective. We emphasis that the emergence of Fog computing is motivated by the predictable service demands of mobile users, and Fog computing is thus mainly used to fulfill the service requests on localized information. As a Fog server possesses hardware resources in three-dimensions (storage, compute and communications), the 3-D service-oriented resource allocations are therefore the key of Fog computing. Moreover, with the three-tier Mobile-Fog-Cloud architecture and rich potential applications in both mobile networking and Internet-of-things, the Fog computing also opens broad research issues on network management, traffic engineering, big data and novel service delivery. Therefore, we envision a bright future of Fog computing.

## References

1. F. Bonomi, R. Milito, J. Zhu, S. Addepalli, Fog computing and its role in the internet of things, in *Proceedings of ACM MCC* (2012), pp. 13–16
2. R. Balan, J. Flinn, M. Satyanarayanan, S. Sinnamohideen, H.I. Yang, The case for cyber foraging, in *Proceedings of ACM SIGOPS* (2002), pp. 87–92
3. M. Satyanarayanan, P. Bahl, R. Caceres, N. Davies, The case for vm-based cloudlets in mobile computing, No. 99 (2011)
4. Location-Based Mobile Services Have Huge Untapped Potential Worldwide (2012)
5. G.M. Relations, Greyhound launches BLUE, an exclusive Wi-Fi enabled onboard entertainment system. https://www.greyhound.com/en/newsroom/viewrelease.aspx?id=528&year=2013. Accessed on Dec 2014
6. T.H. Luan, L.X. Cai, J. Chen, X. Shen, F. Bai, Vtube: towards the media rich city life with autonomous vehicular content distribution, in *Proceedings of IEEE SECON* (2011), pp. 359–367
7. G. Peng, CDN: content distribution network (2004). arXiv preprint cs/0411069
8. B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, B. Ohlman, A survey of information-centric networking. IEEE Commun. Mag. **50**(7), 26–36 (2012)
9. E. Baştuğ, M. Bennis, M. Debbah, Living on the edge: the role of proactive caching in 5G wireless networks. IEEE Commun. Mag. **52**, 82–89 (2014)
10. M. Satyanarayanan, Z. Chen, K. Ha, W. Hu, W. Richter, P. Pillai, Cloudlets: at the leading edge of mobile-cloud convergence, in *Proceedings of MobiCASE* (2014)
11. Y. Zhang, Y. Zhou, Transparent computing: a new paradigm for pervasive computing, in *Ubiquitous Intelligence and Computing: Third International Conference* (Springer, Berlin/Heidelberg, 2006), pp. 1–11
12. T. Komine, M. Nakagawa, Fundamental analysis for visible-light communication system using LED lights. IEEE Trans. Consum. Electron. **50**(1), 100–107 (2004)

13. C.X. Lin, X. Shen, J.W. Mark, L. Cai, Y. Xiao, Voice capacity analysis of WLAN with unbalanced traffic. IEEE Trans. Veh. Technol. **55**(3), 752–761 (2006)
14. K. Zheng, F. Hu, W. Wang, W. Xiang, M. Dohler, Radio resource allocation in lte-advanced cellular networks with m2m communications. IEEE Commun. Mag. **50**, 184–192 (2012)
15. H. Kim, N. Feamster, Improving network management with software defined networking. IEEE Commun. Mag. **51**, 114–119 (2013)
16. K. Zheng, L. Hou, H. Meng, Q. Zheng, N. Lu, L. Lei, Soft-defined heterogeneous vehicular network: architecture and challenges. CoRR (2015). abs/1510.06579
17. I. Stojmenovic, S. Wen, The fog computing paradigm: scenarios and security issues, in *Proceedings of FedCSIS* (2014)

# A Conceptual 5G Vehicular Networking Architecture

## The Driving Force to Connected Autonomy

**Konstantinos Katsaros and Mehrdad Dianati**

**Abstract** This chapter presents a thorough investigation on current vehicular networking architectures (access technologies and overlay networks) and their (r)evolution towards the 5G era. The main driving force behind vehicular networking is to increase safety, with several other applications exploiting this ecosystem for traffic efficiency and infotainment provision. The most prominent existing candidates for vehicular networking are based on dedicated short range communications (DSRC) and cellular (4G) communications. In addition, the maturity of cloud computing has accommodated the invasion of vehicular space with cloud-based services. Nevertheless, current architectures can not meet the latency requirements of Intelligent Transport Systems (ITS) applications in highly congested and mobile environments. The future trend of autonomous driving pushes current networking architectures further to their limits with hard real-time requirements. Vehicular networks in 5G have to address five major challenges that affect current architectures: congestion, mobility management, backhaul networking, air interface and security. As networking transforms from simple connectivity provision, to service and content provision, fog computing approaches with caching and prefetching improve significantly the performance of the networks. The cloudification of network resources through software defined networking (SDN)/network function virtualization (NFV) principles, is another promising enabler for efficient vehicular networking in 5G. Finally, new wireless access mechanisms combined with current DSRC and 4G will enable to bring *the vehicles in the cloud*.

## 1 Introduction

Connected vehicles enable a broad range of applications for safety, traffic efficiency and infotainment, and in the future facilitate connected autonomous vehicles. However, it is shown that the current communication systems are not sufficiently

K. Katsaros (✉) • M. Dianati
Institute for Communication Systems, University of Surrey, Guildford GU2 7XH, UK
e-mail: k.katsaros@surrey.ac.uk; m.dianati@surrey.ac.uk



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_22

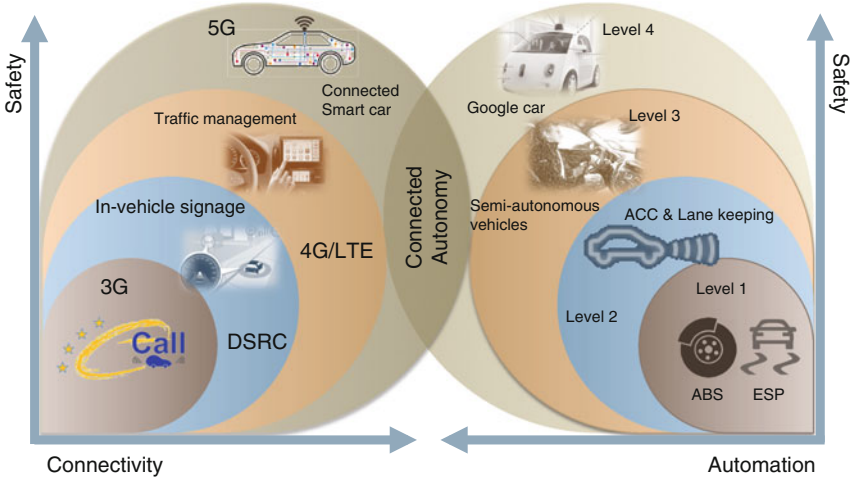


**Fig. 1** Evolution of vehicular safety systems with connectivity and automation (NHTSA classification [50]: Level 1: Function specific automation; Level 2: Combined function automation; Level 3: Limited self driving automation; Level 4: Full self-driving automation)

reliable for the market [59] and something that raises concerns to the public when vehicles are travelling through traffic automatically with no driver behind the wheel. On the other hand, academia and industry are capitalizing on the application requirements and weaknesses of current communication systems to research and design the fifth generation (5G) networks.

There are presently two parallel paths in automotive industry that 5G will seek to converge; *connected* and *autonomous* vehicles (Fig. 1). Safety has always been one of the driving forces of innovation in automotive industry. From passive and active safety systems such as seatbelts and sensors, which alleviate the damage of an accident and alert drivers, we are currently at what is known as *connected mobility*. Intelligent Transportation Systems (ITS) are the building block for connected mobility. They are the outcome of coupling transportation systems with wireless communications offering users several innovative and effective services. It is envisaged that the future vehicles will be able to form ad-hoc networks between them, or connect to infrastructure in order to exchange important traffic and safety related information in highway and urban environments. There is a strong demand from the automotive industry, driven by market forces and consumer needs, to transform future cars into smart, connected systems, *a smartphone on wheels*, that offer personalised services and have the capability to adapt to the users' preferences and needs. Several communication technologies have been proposed and investigated for vehicular network applications from academia, industry and governmental bodies. Each of these technologies have certain properties that make them suitable for particular type of applications, mostly dictated by the end-to-end delay requirements, and the dissemination mode i.e. broadcast, geocast, or unicast.

However two technologies, namely, dedicated short range communication (DSRC), e.g., IEEE WAVE, ETSI ITS G5, and cellular, e.g., 3GPP LTE, are the most promising potential candidates.

What is more, state-of-the-art cars in the market currently provide some simple autonomous features such as emergency braking function, adaptive cruise control, and lane assist system. In addition, advanced navigation and infotainment systems take advantage of the Internet connectivity via cellular networks to provide enhanced and dynamic navigation services. Research in robotics, inspired by unmanned space vehicles, has given birth to driverless concept cars such as Google robotic car[1] empowered by proximity sensing and on-board autonomous processing.

Given that by 2030–2050 nearly all cars are expected to be autonomous [42], it is emerging that a combination of connected vehicles and autonomy can significantly boost the performance, safety, and reliability of new generations of autonomous transportation system. The binding component to achieve this, is 5G communications. Connected autonomous vehicles will lead to vastly improved safety and fewer collisions. KPMG estimates that by 2030, connected autonomous cars could save over 2500 lives and prevent more than 25,000 serious accidents in the UK. Production and dissemination of traveller information will be improved with the exploitation of Internet of Things (IoT) and Big Data analytics facilitated by low-latency, fully connected vehicles and cloud infrastructure. Travelling experience will be enriched for autonomous vehicles' occupants with augmented-reality infotainment systems. In this direction, Ericsson has proposed the Connected Traffic Cloud—a managed cloud concept that enables the sharing of real-time traffic and road condition data between connected vehicles and road traffic authorities. Future cars will be highly autonomous, smart systems that relieve the drivers from tedious tasks and reduce the negative impacts of human errors on safety of passengers and pedestrians. They will be connected to the internet and their surrounding environment (other cars and infrastructures), having access to a massive amount of information and processing power from the cloud and awareness about their surroundings environment. This will improve safety, efficiency, environmental friendliness, and comfort level of future transport systems. Such significant increase in digital capabilities of future cars is expected to have enormous positive economic and societal impacts.

Following this brief introduction to the vehicular networking ecosystems and their potential benefits, Sect. 2 provides a more detailed description of the vehicular applications and the specific requirements they have from a networking perspective. Then, in Sect. 3 current vehicular communication systems and their limitations are presented, demonstrating different access technologies as well as networking architectures. Based on the requirements for current and future vehicular application, as well as the limitations of current vehicular communication systems, a conceptual vehicular networking architecture towards 5G is proposed in Sect. 4. Finally, Sect. 5 concludes this chapter.

[1]Google Self-Driving Car Project [online]: https://plus.google.com/+GoogleSelfDrivingCars.

## 2 Vehicular Applications and Requirements

In this section, current and futuristic use cases of vehicular networks and their specific communication requirements are reviewed. These include latency limits, reliability, as well as security and privacy requirements, focusing only on off-board communications.[2] The basic requirements and use cases for 5G networking have been presented in Chap. "Spectrum Analysis and Regulations for 5G". and can be summarized in the following list compared to current 4G:

- Increase in mobile data volume (bps/km$^2$) and peak data rates by a factor of x100 and x1000, respectively;
- Reduce the end-to-end delay to 1 ms;
- Increase reliability and support higher mobility; and
- Increase energy efficiency by 10 %.

Latency requirements for the different categories of ITS applications are specified by ETSI [66]. Ultra low latency (e.g., 1 ms for the round trip delay) will be required in order to ensure the exchange of kinematic information for distributed control algorithms [23] for future ITS. Furthermore, the architecture of ITS systems is shifting towards a cloud-based system in order to off-load heavy calculations and storage to the cloud, facilitate vehicle maintenance and deliver new services to customers with a short time-to-market [16].

There is a long list of desirable applications considered for the short-term deployment of ITS [66] that can be grouped into four categories:

- Safety applications focusing on reducing accidents and their impact, e.g., road hazard warning [19];
- Traffic efficiency applications aiming at more efficiency and thus greener traffic, e.g., green light optimal speed advisory (GLOSA) [38];
- Comfort and infotainment applications, e.g., automatic map updates or video streaming [64]; and
- Manufacturer oriented applications, e.g., automatic software updates [27].

However, new application types have been foreseen for future deployments of automated and connected vehicles, e.g. augmented reality navigation, co-operative electronic stability control [23].

The following subsections analyse some of the use cases and their requirements, e.g. for safety applications, and general requirements for vehicular networking, e.g. security.

[2] Vehicle to Vehicle and Vehicle to Infrastructure.

### 2.1 Safety Applications

Safety-related applications aim to enhance driver's road perspective by analysing received messages such as Cooperative Awareness Message (CAM), Basic Safety Message (BSM) or Decentralized Environment Notification Message (DENM) broadcast by other vehicles and road side units (RSUs) in the communication range. These messages carry the status of the transmitting vehicle like kinematic information, type of vehicle etc, or information related to specific event, e.g. an accident. They are transmitted with frequencies ranging from 1 to 10 Hz. ETSI specifies the latency limit for the reception of these messages to 100 ms [66]. Currently, these applications are realised by in-vehicle systems that display warnings or emit sound to inform the driver. In the future, these warnings can be displayed on augmented dashboards as overlay information on top of what a driver is seeing through the front window, e.g. MINI's augmented reality.[3] It is also very important that cars can communicate not only with each other but also with other participants in the city, such as pedestrians or cyclists.

### 2.2 Comfort and Infotainment Applications

Comfort and infotainment applications are mainly constrained by the available network capacity. For example, the trend in video streaming is to support "4K" content. Netflix, a content provider, currently encodes 4K streams at ∼15 Mbps and with projection to get it down to 10–12 Mbps for 30 fps streams. Even though such download speeds may be available with 4G technology, the challenge here is how to scale to multiple users with high mobility, as future users will expect to continue their good quality connection independent of their location and speed. In terms of latency requirements, ETSI has given a higher limit for this type of applications at 500 ms [66], but some types of applications, such as VoIP, are also constraint by jitter.

### 2.3 IoT/Monitoring Cloud-Assisted Applications

Internet of Things (IoT) is gaining momentum and it is foreseen to reach more than 50 millions connected devices by 2020 [12], with 11 millions being embedded cellular-based telematics [4]. A recent survey[4] showed that application related to

[3]MINI's augmented-reality glasses allow drivers to see through the body of their car [Online]:http://goo.gl/oQUwal.

[4]The 10 most popular Internet of Things applications right now [online]: http://iot-analytics.com/10-internet-of-things-applications/.

connected vehicles, like smart parking and remote-controlled car, are in the 10 most popular applications. In addition, the electrification of vehicles and its interconnection with smart-grid has spurred new use-cases that will require very good scalability and ability to efficiently manage heterogeneous networks [9]. Currently, each vehicle has an average of 60–100 sensors on board. This is projected to reach 200 by 2020 as vehicles become "smarter". All these sensors travel together with the vehicle like a swarm, which has lead to the development of several applications. One of the most futuristic and highly computational expensive use cased of IoT has been proposed in the work of Jain et al. [35]. It exploits several inputs from vehicle sensors to pre-assess the driver behaviour and anticipate future manoeuvres (3.5 s before they occur) only within ∼3.6 ms computation time. In the future, individual driver input shall be fuzzed with the aid of cloud infrastructure and big-data analytics with surrounding drivers to provide co-operative assessment of manoeuvres with the same computational time limit.

## *2.4 Connected Automated Driving*

There is a plethora of automated functions currently implemented on-board, like active suspension systems, that interconnect sensors and actuators around the vehicle through a ultra-low latency wired network. However, future co-operative automated functions should support equivalent characteristics while fusing sensor information from other vehicles. The computational capabilities of the vehicles are currently restricted, which limits the supported functionalities. Recent literature has revealed that the performance of the control systems can be significantly enhanced if the on-board information and processing is augmented by the information and processing power from the cloud [73]. In the industry, Electrobit[5] has collaborated with Infineon and NVIDIA to develop an automated driving platform that will increase the computing power and intelligence of vehicles required for automated driving. On the other hand, ZF Friedrichshafen AG has launched a cloud-based smart urban electric vehicle[6] With two semi-automated driver assistance functions they aim to increase comfort and safety of the passengers. Concretely, Smart Parking Assist maneuvers the vehicle into virtually any small parking space—remote-controlled at the push of a button using mobile devices such as a smartphone or smartwatch. Comfortable, efficient motoring is possible with the concept vehicle thanks to the cloud-based PreVision Cloud Assist assistance function. Where necessary, the system reduces the drive torque, for instance, in good time before entering the bend and thus throttles back the speed without any mechanical braking. Most

[5]Elektrobit (EB) collaborates with Infineon and NVIDIA to deliver first-of-its-kind automated driving platform—Press Release [online] https://goo.gl/LxxaLk.

[6]ZF Press Release "Collective Knowledge: PreVision Cloud Assist in the Advanced Urban Vehicle from ZF", July 2015.

of the research on automated functions is focused on light vehicles. However, trucks and buses have also been enhanced with autonomous functions. For example Daimler's research on autonomous trucks has produced functions such as the new anti-jackknifing system, which Daimler engineers have developed to ensure that extra-long articulated buses remain controllable in difficult situations. All these characteristics of automated driving make it a challenging topic as it requires ultra-reliable, fault-tolerant networks, low end-to-end delay among vehicle and with cloud-based infrastructure, and it must work everywhere. In addition, driver-less vehicles should be able to communicate with other "vulnerable" road users such as bicycles and pedestrians.

## *2.5 Security and Privacy*

This section provides an overview of the security and privacy requirements for vehicular networking. Achieving high level of security with low communication cost is challenging. In addition, concerns regarding privacy breaches introduces another challenge to bring this type of technology to the masses. As cars become smarter, more and more vehicle components interconnect through the onboard vehicular network. This trend has been reflected in the number of lines of code running on all the electronic control units (ECUs). A modern high-end car features around 100 million lines of code, and this number is planned to grow to 200–300 millions in the near future. This creates new threats as external adversaries can potentially intrude and take control of critical vehicle functions [36]. A classification of the threats in modern vehicles can be found in [46]. The functionalities related to security and privacy that have to by provided by the communication system include the followings.

- *Data origin authentication* ensures the validity of the origin of a message. Vehicles commonly use digital certificates to perform authentication. A digital signature typically contains information such as name of a vehicle or its role, a public key, validity, and signature.
- *Data Integrity* ensures that the contents of messages are not altered on transit by an adversary, which could undermine the safety of users of the system and lead to fatalities. This can be useful in multi-hop message relay scenarios where messages are passed on from one node to the other. Integrity checks can also help prevent replay attacks in which location and time stamp information can be altered to deceive message recipients.
- *Network Availability* is a key requirement of vehicular communication system for implementation of critical applications such as emergency message dissemination. Although provisioning of total and unconditional availability is not feasible in practice, it is crucial for network engineers and application developers to have a clear understanding of the availability constraints and implement proper

**Table 1** Vehicular application requirements

| Application type | Latency limit (ms) | Bandwidth | Security requirements | Examples use cases |
|---|---|---|---|---|
| Infotainment | 500 | ✓✓✓ | ✓ | Video streaming |
| Traffic efficiency | 200 | ✓ | ✓✓ | e-Toll collection |
| Safety | 100 | ✓ | ✓ | Collision warning |
| Reliable M2M | 20 | ✓ | ✓✓ | Auto-cruise |
| Real-time virtual reality | 5 | ✓✓ | ✓✓ | Augmented reality navigation |
| Control | 1 | ✓✓ | ✓✓✓ | Co-operative electronic stability control |

mechanisms to combat potential Denial of Service (DoS) attacks and mitigate the risk associated with the availability of the network.

- *Privacy* in vehicular communication systems refers to the capability that ensures personal identifying information of the drivers is not disclosed to third parties. Hence, the network often and by default is required to ensure that activities of drivers cannot be traced by the adversaries. However, conditional privacy stipulates that it should be possible to reveal the identity of offending drivers for revocation and criminal prosecution. *Anonymity* is a technique of hiding the physical identity of a vehicle such as IP address and electronic number plates and thus providing the required privacy. The Vehicular Public Key Infrastructure (VPKI) should ensure a means to offer conditional anonymity to drivers by separating vehicular information from personal information about drivers.
- *Confidentiality* provision of the VPKI ensures that messages cannot be easily read or understood by third parties even if they are intercepted. However, for the reasons of safety, message confidentiality should not be implemented for emergency messages. Hence, broadcast applications such as DENM and CAM do not typically require confidentiality.

A summary of application requirements in terms of latency limits, bandwidth and security is presented in Table 1. Applications are grouped in five types and example use cases are given for each type.

## 3 Current Vehicular Communication Systems

This section presents the state-of-the-art in vehicular networking architectures and their ability to meet the application requirements specified in Sect. 2. To this end, the most prominent existing wireless access technology candidates for vehicular communications are dedicated short-range communications (DSRC) based on IEEE 802.11p standard and cellular communications with 3GPP LTE. In addition, hybrid solutions which integrate DSRC and LTE have emerged that aim to benefit from advantages of both technologies. There are also higher layer (overlay) networking

architectures alternatives like information centric networking (ICN), which decouples locator and identifier of contents, and vehicular cloud computing. Content-based applications have promoted the introduction of ICN, with its most popular candidate the Named-Data Networking (NDN), in the vehicular networking sphere. Vehicular cloud extends the concepts of mobile cloud computing in the vehicular ecosystem taking intro account the characteristics of the vehicular applications. These five vehicular networking architectures will be analysed in the remainder of the section.

## *3.1 Dedicated Short Range Communication (DSRC)*

Dedicated short range communication (DSRC) is based on a two-way wireless system operating in the 5.9 GHz licensed spectrum band. In the US, Federal Communication Commission has allocated 75 MHz spectrum, whereas in Europe, ETSI allocated 30 MHz in the same band. Both in Europe and the US, the principal physical and MAC layers are based on the IEEE 802.11p standard. In Europe, that has been implemented within the ETSI ITS G5 [21], with ITS-G5A being dedicated to safety related applications and ITS-G5B to non-safety applications. On the other hand, in US the implementation follows the IEEE WAVE standard [32]. The main differences of the two implementations are located on upper layers and particularly the network as shown in Fig. 2. Both implementations promote a dual stack architecture, splitting traffic to ITS-specific applications and generic IP-based applications. The later are served by TCP/UDP or other transport protocol over an IPv6 network. These are non-safety applications and are carried by ITS G5 B or UMTS in the European stack. It is possible to translate IPv6 to GeoAddress and use the GeoNetworking layer [68]. For ITS-specific applications, ETSI has proposed the Basic Transport Protocol (BTP), a UDP-like transport protocol, operating on top of the GeoNetworking layer. GeoNetworking provides different methods of packet forwarding (broadcast, geocast and unicast) based on geographic routing. On the other hand, IEEE WAVE architecture proposed the Wave Short Message Protocol (WSMP) as the main network layer for ITS-specific traffic. Contrary to ETSI GeoNetworking, WSMP is based purely on single-hop broadcast, hence no routing protocol is employed as no application requiring forwarding has been designed in WAVE architecture.

### 3.1.1 DSRC Performance in Vehicular Environments

There have been several performance evaluations of DSRC networks either for benchmarking or for comparison with optimised protocol proposals. In Fig. 3, we present a baseline evaluation of the performance of a standard DSRC (IEEE 802.11p-based) network in highway and urban scenarios with respect to packet delivery ratio and delay. All vehicles periodically broadcast single-hop packets

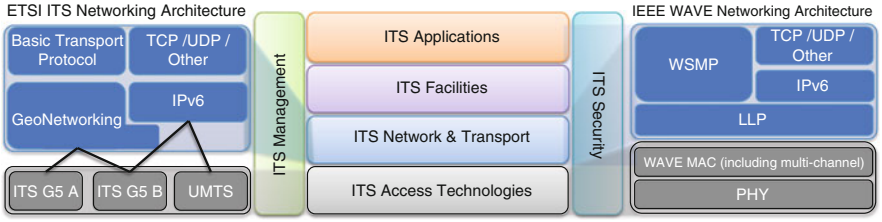


**Fig. 2** DSRC protocol stack in Europe (*left*) [20] and US (*right*) [32]

of same type (e.g. CAMs) and same size (300 bytes) with frequency 10 Hz. For our evaluations we have assumed a vehicle density of 1400 and 1800 veh/h/lane with three lanes per direction following a Poisson arrival and vehicles travel with speeds varying from 60–120 km/h[7] in highway scenarios. In urban scenario, the vehicle density is 900 veh/h/lane with two lanes per direction again following a Poisson arrival. Vehicles travel in a Manhattan grid road topology with speeds from 50–60 km/h. The path loss model in both scenarios considers the large scale and shadowing effects and in the urban scenarios buildings obstruct line-of-sight communications.

Evaluations from state-of-the-art works like Wu et al. [74] and Mir [48] conform with the baseline results in Fig. 3. In principle DSRC can meet the strict latency constrains for active safety in most cases, but as the vehicle density and vehicle speed increase the percentage of packets received over the delay limits increase. For DSRC networks, the delay increases when the transmission frequency and number of vehicle increase. Sparse networks exhibit better performance, however, as the network load increases, denser networks with higher transmission frequencies result in higher end-to-end delays. The end-to-end delay increases as the vehicle speed increases. Vehicles spend more time in contending for the common channel as the result of larger neighborhood densities and because of the fact that they need to access the channel more frequently for higher transmission frequencies.

The previous evaluation only covered the single-hop broadcast scenarios supporting safety and traffic efficiency applications. For infotainment and cloud-based applications, unicast multi-hop communications are required. By design DSRC cannot meet the data rate requirements for high throughput infotainment applications such as 4K TV, as the high data rate of 27 Mbps configuration of IEEE 802.11p is only achievable with a very good channel quality. Such situation is obtained only with static nodes and line-of-sight communication, while lower rates (6–12 Mbs) that are more robust, are selected for highly mobile cases. Further, connection to RSUs and from there to the internet is intermittent with an average duration of a few seconds depending on vehicle speed. Handing-off between RSUs introduces significant delay for address re-configuration and authentication in the new access point [83]. Clearly the ultra-low latency of future real-time cloud-based control applications can not be met in current DSRC systems.

[7]The geography of transport systems—Highway Speed, Flow and Density [online] http://goo.gl/biXc6d.

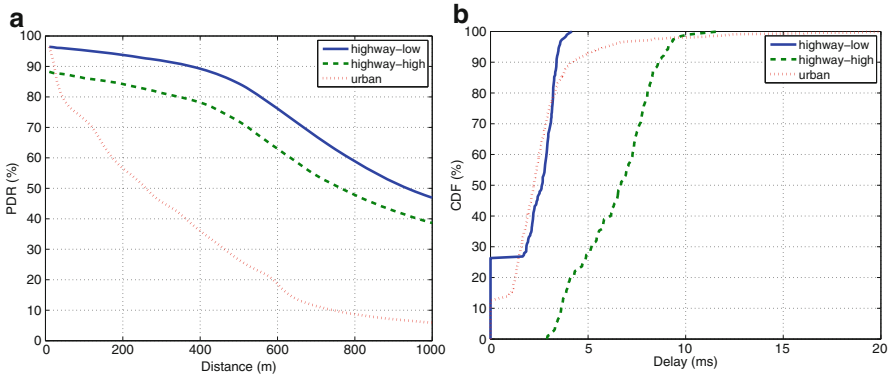


**Fig. 3** DSRC Performance in highway and urban scenarios. (**a**) Packet delivery ratio. (**b**) CDF of single-hop delay

## *3.2 Cellular Network Communications*

Cellular communications have long been used in vehicular networking, from GSM up to the most recent LTE networks, due to the wide penetration of the technology. One representative use case of cellular networking for vehicular applications is the European initiative *e-Call* that will be mandatory for all new vehicles shipped after April 2018. In case of a crash, e-Call equipped vehicles will automatically call the nearest emergency centre using Europe's single emergency number (112), as well as send a 'minimum set of data', which includes the location of the vehicle. ETSI has proposed a framework to integrate mobile networks in cooperative ITS (C-ITS) architectures [67] as depicted in Fig. 4. The framework identifies related functional requirements on the ITS architecture, amendments and modifications of existing standards on C-ITS in order to enable usage of public mobile cellular networks. One of the most important requirements in the architecture is a network server[8] responsible to the re-broadcasting of messages as LTE design does not allow Vehicle-to-Vehicle broadcast. A recently emerged technology that enables direct vehicle-to-vehicle communications when they are in proximity is Device-to-Device (D2D). A feasibility study of D2D in a vehicular network context has been performed in [11] that demonstrated promising results in boosting spectrum utilization. LTE has already been trialled by manufacturers for emulating non-safety critical V2V applications, such as exchanging non-time critical information including hyper-local weather, road conditions, and traffic data directly between vehicles.

[8]Reflector as dubbed in CoCarX project [15].

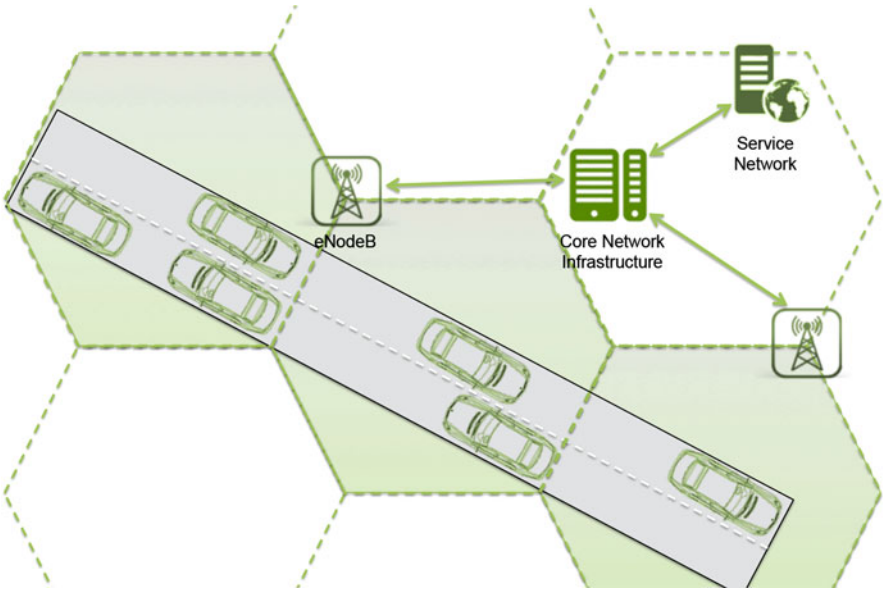


**Fig. 4** Architecture for mobile networks in cooperative ITS

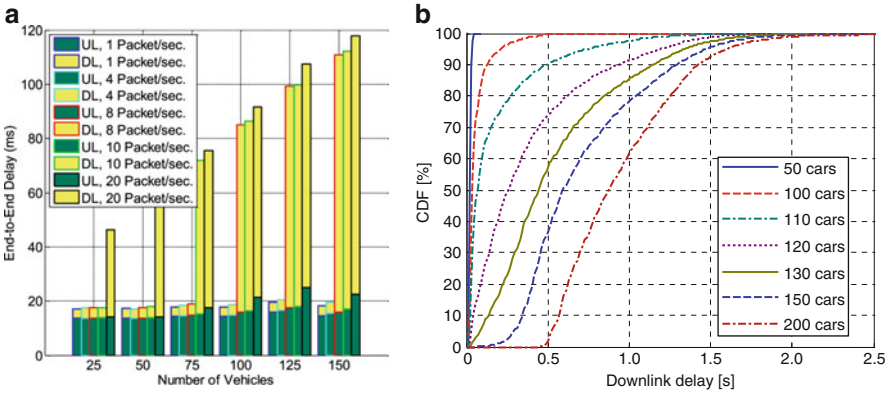


**Fig. 5** LTE Performance. (**a**) End-to-End delay [48]. (**b**) CDF Downlink delay [67]

### 3.2.1 Cellular Networks Performance in Vehicular Environments

The design of LTE networks targets a user plane latency of 20 ms. These figures exclude the latency introduced by the backhaul network (i.e., RAN edge-Core edge, including security gateway), which is currently about 10 ms [65]. Unfortunately, the actual end-to-end delay is much greater than the target specifications as reported by OpenSignal for US mobile network providers [24].

Performance evaluations of LTE networks in the vehicular context exist in the literature [6, 48, 55, 67] and samples are presented in Fig. 5. One common assump-

tion followed by all these works is that only vehicular traffic exists on the network, hence their results are generally over optimistic. Their results can be summarised in the following. LTE can meet the latency requirements for co-operative and active safety applications in limited scenarios (e.g., low loads, low number of vehicles). However, it can provide better service to infotainment applications due to better mobility support and higher available bandwidth. The end-to-end delay for the LTE network with vehicular traffic application only (i.e., no background traffic flows) is analysed as follows (Fig. 5a). As the network size and transmission frequency increase, the end-to-end delay increases. For higher network load, the total delay however gradually increases. While the uplink delays largely remain intact, the downlink delays increase significantly as the number of vehicles and transmission frequency increases. This can be also seen in Fig. 5b, where the CDF of downlink delays is analysed. Further, the increase of participating vehicles and the frequency of broadcast messages has a significant effect on the latency. The evaluation of LTE for CAM and DENM dissemination in [55] exhibits such behavior. The use of eMBMS to broadcast messages can potentially reduce the delay and increase the number of supported nodes/cell [6]. Moreover, even though LTE provides some service differentiation through the Evolved Packet switched System (EPS) Bearer mechanism, none of them can guarantee the ultra-low delay requirements of real-time control applications.

### *3.3 Hybrid Networking Architecture*

Hybrid solutions that integrate DSRC with cellular networks, in loose or tight coupling, can alleviate some of the burdens and extend the range of supported scenarios for both LTE and DSRC usage, but still there are limitations. For example, the works in [79, 82] propose mechanisms that assist routing in DSRC systems with dissemination of data through cellular network, looking at the tradeoff between higher delivery ratio and lower latency. However, these techniques still cannot meet the latency requirements in Table 1. Another use of cellular systems in conjunction with ad-hoc networks is to facilitate efficient clustering mechanisms [57, 63]. In these works, only the cluster heads are connected to the cellular networks and data is disseminated using DSRC within the clusters. Clustering mechanisms are useful for reducing effective network size and potentially reduce latency within each cluster however the highly dynamic nature of vehicular networks requires very efficient and low overhead mechanisms. Further, tight coupled hybrid networking architectures facilitate with off-loading traffic from cellular networks to DSRC either for capacity improvements or for economic reasons [10, 39, 40]. Vehicular traffic could select either networks depending on the type of traffic (e.g. signalling or data traffic as in [39]), or based on a utility function taking into account the cost, signal quality and other indicators. However, several challenges arise in such architectures that relate to the offloading strategies, the practicality of DSRC infrastructure deployments that need to be tackled in order to fully benefit from these networks.

### 3.4 Information Centric Networking

There is an increasing trend in networking research to decouple locator and identifier of content, such as the IP addresses, from the actual content. This networking architecture is known as Information Centric Networking (ICN) and the most notable implementation is the Named-Data Networking (NDN) architecture [80]. Other ICN architectures include Data Oriented Network Architecture (DONA) [41] with a clean-slate redesign of Internet naming and name resolution, and Content Centric Networking (CCNx) [34] addressing scalability, security and flexibility challenges of modern Internet. In principle, ICN architectures are based on the publish-subscribe paradigm, where intermediate nodes/routers can cache the information and serve future consumers. This is depicted in a "narrow waist" approach Fig. 6, where the IP-based networking architecture on the left is compared with ICN-based architecture on the right. As mentioned above, ICN decouples content from the identifiers, and the whole network architecture is now constructed around the content itself, rather than the IPs. The characteristics of vehicular content, such as geography—and time-dependency of certain events, and the fact that it targets a large set of vehicles independent of their IPs, makes ICN a suitable candidate networking architecture. In addition, ICN provides support of content caching, which is beneficial for the intermittent DSRC-based networks. Recent works on ICN related to the vehicular networks ecosystem include [8, 29, 71, 75].

### 3.5 Vehicular Cloud Networking

Several automotive companies are promoting cloud connectivity services through cellular systems in their latest high-end models such as the Ford Sync[9] or the Apple CarPlay.[10] Such services are designed to deliver applications in a car for information, entertainment, making service appointments and other needs. In the future, remote diagnostics and operating system upgrades could be feasible. Model Predictive Control (MPC) techniques that are foreseen in future autonomous vehicles and require significant computing resources could be provided as a cloud service as well [25]. Content-based networking presented in Sect. 3.4 has promoted further the use of cloud-based architectures. Vehicular cloud networking shares the principles of mobile cloud networking where each component of the network is considered *a service* [28, 30, 44, 62, 73, 81]. A basic architecture for vehicular cloud networking is presented in Fig. 7. Cloud-based services, ranging from ITS-specific to generic content services, can be accessible either through cellular access networks

[9]Ford SYNC technology [Online]:www.ford.com/technology/sync.

[10]Apple CarPlay [Online]:www.apple.com/ios/carplay.

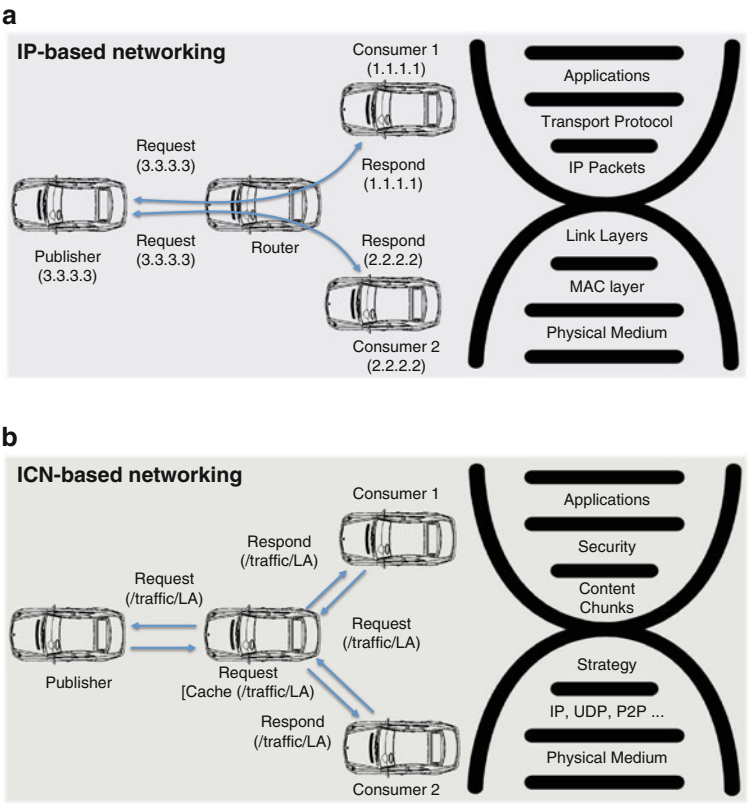


**Fig. 6** ICN-based compared to IP-based vehicular architecture. (**a**) IP-based networking. (**b**) ICN-based networking

or DSRC-based. However, such architectures introduce significant delays as current cloud-based services are quite centralised with small number of spatially distributed servers.

## 4 Vehicular Networking in 5G

This section aims to analyse the challenges and candidate 5G technologies that have been recently under discussion and their implications on vehicular networking. More advanced networked-car visions will rely on capabilities that come with the 5G technology and perhaps will augment today's DSRC and 4G networks. For example, BMW anticipates that 5G mobile networks could be vital in providing the mission-critical reliability as it seeks to deploy self-driving cars onto city streets.

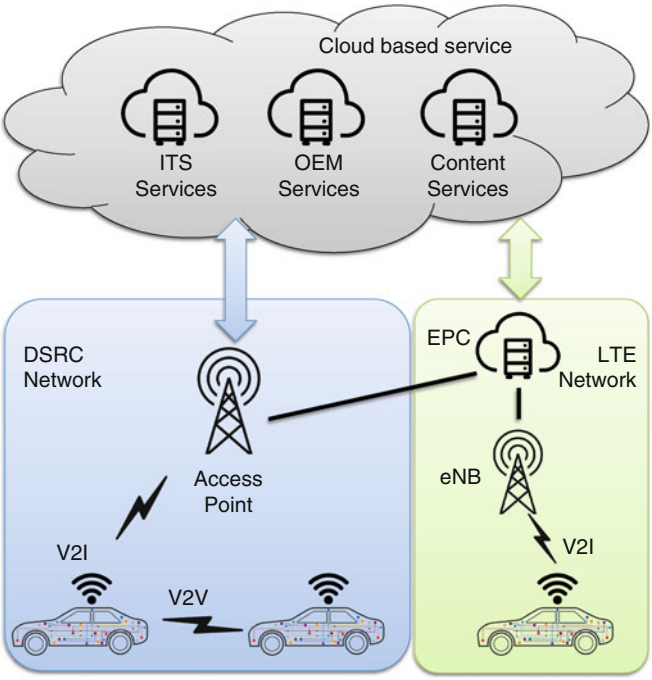


**Fig. 7** Vehicular cloud networking architecture

## *4.1 Challenges*

In Sect. 3, current vehicular networking architectures were reviewed, determining their limitations to meet applications' requirements, due to congestion, mobility and other bottlenecks. Existing research reveals that nominal, congestion free, access network latency in the range of 1–2 ms is feasible [23] for communications between cars and the edge of the cloud. But, progress toward this objective requires novel research into wireless transmission, network architectures and functions/protocols to reduce serialisation, queuing, and network processing latencies. For reliability and security, additional degrees of freedom in 5G communication systems (eg, spatial diversity in dense base station deployments) as well as physical layer encryption techniques have great potentials. These challenges and existing solutions are identified in the remainder of the section.

### 4.1.1 Congestion

The performance analysis both of DSRC and LTE systems, in Sects. 3.1.1 and 3.2.1 respectively, has outlined a common limitation on both architectures due to

congestion. The effective network size and traffic load play a significant role to the performance of latency. Due to the limited available resources, competition among entities to access the shared radio resource introduces latency. In order to avoid congestion caused due to large network size or heavy load, multiple solutions have been proposed.

First, congestion can be avoided by introducing clustering which groups the large amount of users into clusters with fewer users and competition. Clustering-based algorithms [53] aim to seek opportunities to minimize the consumption of radio resources via integrating information, working as one station, and reusing radio resources in different clusters. Second, reuse of radio resources via smart allocation of multiple non-overlapping channels helps reduce the contention level, e.g., multichannel MAC design [78]. It is critical to have a technology that works even when the network operator is not there, as they are out of range for example. Device-to-device (D2D) communication is a potential solution for such situations. D2D communications [11, 56] also target to tackle the problem by reducing contention level in the system via encouraging local p2p communications. Meanwhile, D2D communications easily avoid redundancy in terms of signalling for controlling and maintenance as in the traditional cellular networks. Coordination (e.g., among Base Stations) on top of D2D communication systems expedites radio resource utilization.

For DSRC systems, ETSI has proposed several decentralized congestion control (DCC) mechanisms [69]. For example, the transmit power control or the DCC sensitivity control mechanisms work on the same principle as clustering, multi-channel MAC and D2D that aim to reduce the effective network size, thus reducing the congestion levels. On the other hand, transmit data rate control, transmit access control mechanisms and the proposed TCP-like congestion control [60] tackle the congestion from another perspective, that of the offered load. For the future vehicular networking, it is unlikely to use a single radio resource management scheme to satisfy the stringent requirements of reliability and latency but more like a combination of efficient MAC schemes. A novel design of the integration of the above mentioned solutions potentially helps reduce the latency caused by congestion.

### 4.1.2 Mobility Management

Vehicular networks are characterised by nodes that are constantly moving on semi-predictive but predefined trajectories over the underlying road topology, contrary to the stationary IoT networks or the low random mobility of pedestrians. This level of node mobility results also to the need of network mobility as vehicles travel under the coverage of different RSUs. It has already been mentioned in Sect. 3.1 that the mobility and sparse RSU deployments result in intermittent connections for DSRC networks and increased delays due to hand-off procedures. However, similar challenge is envisioned in future 5G networks, where the densification of

the network with the use of small cells will increase the probability of hand-off. Therefore, optimising hand-off procedure is a crucial task in the process of reducing delay and having more reliable connections.

Solutions based on Network Mobility (NEMO) and Mobile IP (MIPv6) are aiming to reduce the handoff time utilizing predictive techniques to perform registration in advance [3, 43, 61]. However, the centralized architecture of MIPv6 is inefficient for highly dynamic vehicular networks and a distributed mobility management approach would be more appropriate [45]. Other solutions include Location Identity Separation Protocol (LISP) [22] that natively supports mobility with route optimization, multi-homing dual stack and network mobility. LISP is also compatible with the ICN principles of locator/identifier separation. Further, the utilization of Stream Control Transmission Protocol's (SCTP) inherent multi-homing support and Media Independent Handover (IEEE 802.21 standard) have also been investigated for efficient handover [52]. Efficient and simultaneous use of all available access networks combined with minimal signalling and predictive techniques ("make before brake") would help reduce the delay introduced by the mobility of the user.

### 4.1.3 Backhaul Network

Mane services have been moved to a cloud based architecture from data storages e.g. dropbox, up to complete operating systems e.g. Chrome OS. Particularly for ITS services, the use of cloud has been demonstrated in Sect. 3.5 with current infotainment services and futuristic control. However, the location of the server would significantly impact the performance of the service, as control techniques can tolerate very low latency and content has to be available seamlessly.

Currently, cloud-based services are centralised in proprietary data centres and the connection of the network provider with that data centre will impact the delay. The locality of the remote host with respect to the end-user (vehicle) is a major source of latency for applications that require such connections. The latency will depend on the type of links and distance of the network gateway and the remote host, as well as the routing rules, which may contradict with the Network Neutrality that FCC recently vote in favor. Thus, bringing the remote host closer to the end-user, what is dubbed as fog computing [5, 14], reduces the backhaul delay. Fog computing within a vehicular network architecture entitles spatially distributing service components and caching content within a flat network provider core and the vehicles [37, 72].

### 4.1.4 Air Interface

As mentioned in the beginning of this section, 1 ms end-to-end delay is feasible in congestion free networks and a proposed division of that delay in different components is demonstrated in Fig. 8. Such architecture will require 0.2 ms for the

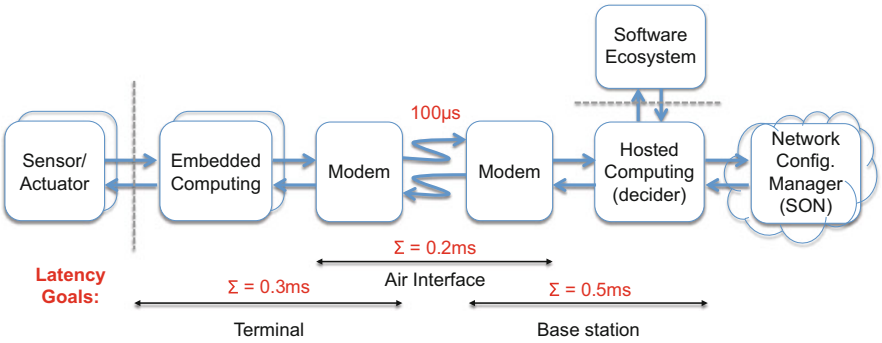


**Fig. 8** Proposed end-to-end delay budget in 5G [23]

air interface, that accounts for ∼15 timeslots[11] the current IEEE 802.11p standard. Current access categories (ACs) with low priorities in the IEEE 802.11p can not satisfy this delay limit and only AC0 and AC1 can fulfil this requirement in theory, under the assumptions of very low contention level and no collisions. Moreover, in current LTE systems, the duration of a single frame is 10 ms consisting of 10 sub-frames with two slots per sub-frame. Large packets would require segmentation to over multiple sub-frames, which would increase the transmission delay of those packets. Further, from the 10 sub-frames, maximum 6 can be used for MBMS, while the rest are left to unicast traffic. This impacts the delay of broadcast packets, which for ITS applications is the majority of traffic.

It is clear that in order to support 0.2 ms air-interface delay in DSRC-based networks, the backoff procedure and relevant mechanisms has to be redesigned. Potential solutions include a mixture of TDMA and CSMA access in a super-frame cycle. Ultra-high AC would have deterministic delay through the TDMA part of the frame cycle and lower AC's would content on the CSMA cycle. Such access mechanisms are employed for on-board networks based on FlexRay [26]. Several air interface solutions specifically designed for reliable vehicular communications have been proposed and analysed in METIS 2020 project [47]. Most notable, the coded slotted Aloha (CSA) MAC technique enables reliable ad-hoc communications that operate in nearly double the network sizes of standard IEEE 802.11p networks. In terms of cellular technologies, the current LTE frame structure is not able to support the 0.2 ms limit. There is a proposal for flexible TDD frame structure [51] that targets to provide the overall latency of 1 ms able to support future ITS use cases such as cooperative autonomous driving, tactile Internet and real time control. Mechanisms to reduce transmission time interval (TTI) and retransmission processes as well as provision of better QoS differentiation have enabled the reduction of latency nearly five times [47].

[11] Assuming each timeslot duration is 13 µs.

### 4.1.5 Security

Security in vehicular communication systems is vital as explained in Sect. 2.5. However, maintaining a secure network (e.g. exchanging keys, encryption) and authenticating source of a message introduces significant latency [54], both in terms of communication and processing. The processing time increases remarkably in dense networks [7]. High beacon rate (e.g., 10 Hz for CAMs) also leads to significant computational overhead in terms of authentication, apart from the congestion reported in Sect. 4.1.1.

The majority of the delay related to security is spent on the processing/checking the legitimation of messages e.g., 43.2 % of the CPU time is devoted to the message check [58]. In order to reduce this delay, the authentication/checking mechanism has to be redesigned. Lightweight filter security schemes have the potential to reduce overall processing time. Such techniques reduce the amount of information to be checked each time or allow fewer checks according to the "credit" for each user. Other solutions aim to reduce the overhead introduced by security mechanisms. They appropriately design effective security and privacy-enhancing schemes such as group signature schemes, hybrid pseudonym schemes that can protect legitimate users as effectively as unsecured systems [7]. However, even those techniques are not fast enough to reach the 1 ms delay budget for the most demanding applications. In order to perform faster computations for security, physical layer security performed by system on chip (SoC) components or exploiting propagation randomness to establish secret keys [49]. Finally, the VPKI which is currently employed to manage security key and pseudonyms ([70]) has to be redesigned as it is developed around DSRC systems since most applications are using V2V broadcast messages. However, in the 5G architecture, DSRC and cellular are expected to be tightly coupled, thus the security protocols used should be shared or at least compatible so that vehicles can communicate seamlessly through any technology.

## *4.2 Conceptual Vehicular Architecture in 5G*

Having considered the requirements and potential solutions to realise secure, ultra-low latency, high-reliability vehicular networks, a *conceptual* vehicular networking architecture in 5G is presented in Fig. 9. This follows a revolutionary approach, rather than simply evolution from the current 4G architecture, empowered by intelligence of the cloud and cooperation among cars and infrastructures. The architectural design is divided into four layers: (a) the generic cloud/internet service layer, (b) the core network cloud layer, (c) the radio access network layer and finally (d) the vehicle and RSU space. Cloud computing technologies are mature enough to be adopted by mobile network providers, hence each of these layers is represented as a cloud. This ecosystem shall be able to support current and future ITS applications in the most demanding scenarios. Following a top-down approach, these layers and functionalities within each of them are described in more details in the remainder of the section.

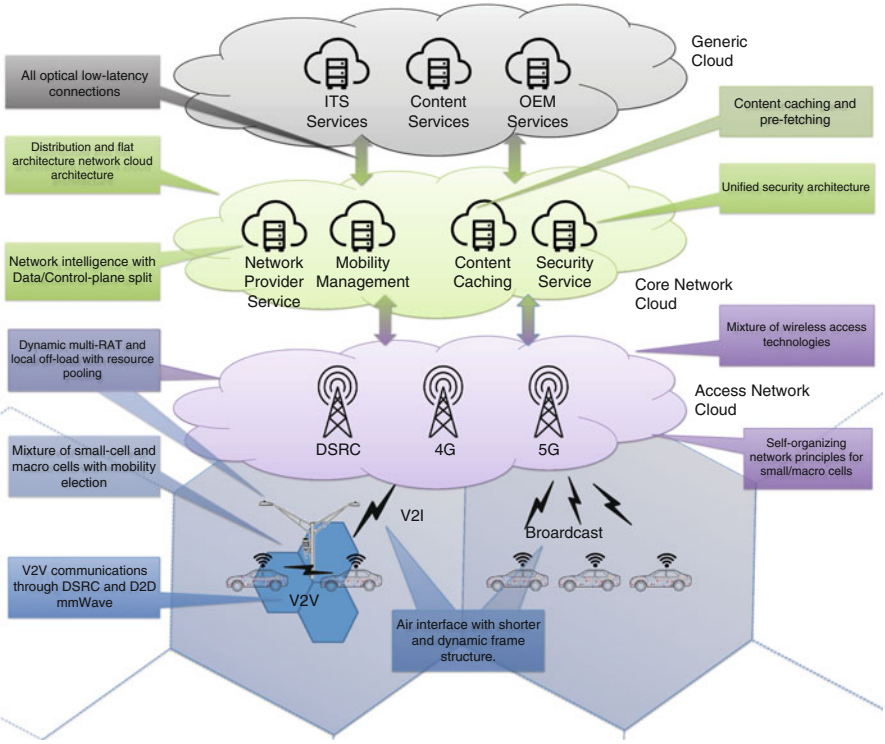


**Fig. 9** Conceptual 5G architecture

### 4.2.1 Generic Cloud

The generic cloud represents the current cloud infrastructure where public authorities and OEMs provide their ITS-specific services. In addition, generic content providers are included in this layer. Following the basic design considerations for 5G architecture, the services have to be spatially distributed even in the generic cloud. The connections of these data-centres with the core network have to be all-optical in order to provide low-latency and high reliability.

### 4.2.2 Core Network Cloud

The core network architecture should be completely revamped. Even though state-of-art LTE-A evolved packet core (EPC) architecture is simpler than the previous 3G core architecture, by reducing network entities, supporting an all-IP network and smart off-loading functions, it still introduces significant delays. In the conceptual 5G architecture of Fig. 9, most of the current EPC functionalities are represented by the *Network Provider Services*. Their implementation is based on the principles of

network function virtualization (NFV) with distributed and flat architecture forming clusters of service functionalities. This allows to reduce signalling traffic by as much as 70 %, and thus reduce end-to-end delays as reported in [31]. This flat cluster-based distributed architecture in the core is also reflected on the access network cloud that will be described later.

As mobility is a distinctive characteristic of vehicular networks and has been shown to be one of the most important challenges for 5G networks, there is a lot of research activity around it. Particularly, the work of Yegin et al. [77] proposes a distributed mobility technique that, contrary to the centralised Mobile IP approaches, pushes the intelligence on the mobile terminal. This provides better scalability and improves efficiency as the terminal is in better position to identify its own flows than the core network. A software-defined network approach is also employed by Yazici et al. [76] to provide the connectivity management as a service (CMaaS). Similarly to [77], high level intelligence (layer 4+) is responsible to maintain the network connectivity through protocols such as MP-TCP, SCTP and SIP. However, a hierarchical network control with different levels of complexity that can provide service differentiation at connectivity level is proposed in CMaaS.

One major enabler for low-latency communication is the ability to cache content within the network, either in the core or even closer to the end user. As multimedia accounts for more than two thirds of global mobile traffic [13], pre-fetching and caching popular content in intermediate nodes can accommodate users of the same content faster. Such an approach of mobile content caching within the EPC as well as the radio access network is described in [72]. It is shown that this technique can reduce the delay by 80 % depending on the amount of cached content in addition to increase the economic benefits for the network provider. Apart from content caching, functionality distribution throughout the core and access network can resolve problems related to congestion and latency. This is described as *fog computing* [14] and provides compute, storage, and network services at the network edge. It also adds intelligence in the network platform itself to orchestrate the underlying resources in real-time.

Last important component of the core network is the security management service. It has been mentioned previously that in order to meet the delay constraints physical layer security mechanisms have to be employed. However, the core network should implement a sophisticated distributed VPKI. With the densification of the radio access network and the use of multiple radio access technologies, the probability of a vehicle being required to handover and re-authenticate increases. Similar to the network mobility solutions, a SDN-based approach for authentication is proposed in [18]. In this technique user-related security context including identity, location, direction and physical layer characteristics, is shared by the SDN controller to the predicted next cell access point. This accelerates the authentication process and increases the levels of tolerance to network failures as the security context is pre-shared among access points.

### 4.2.3 Access Network Cloud

Significant redesigns in the wireless access network will appear in the future 5G with the enabling technologies described in Part II "Transmission and Design Techniques for 5G Networks" of this book. These will characterise the 5G access, however current technologies such as 4G and DSRC will not eclipse. Within the 5G access network cloud, several technologies have to work in synergy in a multi-tier architecture from macro cells providing a ubiquitous service coverage to D2D communications through mm-Wave and DSRC for inter-vehicle connections.

Using the principles of software defined radio (SDR), hardware and software are decoupled. This enables network intelligence to flexible control the available radio resources either from the network or the user perspective. Operators can dynamically configure the network capacity faster, which increases the flexibility for network expansion and bottleneck elimination. However, the resulting increase in operational complexity calls for advanced self-organizing techniques for the underlying networks. Exploitation of big data analytics of control, signalling and contextual data within the core network provides unprecedented amounts of system-level intelligence. This network intelligence can be fed into proactive self-organised network (SON) engines to enable better resource orchestration as reported in [33].

The dynamic multi-RAT access will also stimulate further off-load from macro cells to small cells or DSRC in order to increase performance and reduce the cost of communications. There are a lot of research works on how and when a user shall switch from one technology to another. At the moment, SNR-based measurements are the preferred method. However Andreev et al. [2] have shown that load-aware user-centric schemes, which augment SNR measurements with additional information about network loading, can improve the performance compared to only SNR-based selection.

### 4.2.4 Vehicle Network Cloud

The last layer in this conceptual architectural design is the vehicle itself and the RSUs. The principles of fog computing can also be employed at this layer, which is the edge of the access network as depicted in Fig. 10. Such a technique is presented in [62], where RSUs form a cloud employing SDN to dynamically instantiate, replicate, and/or migrate services. This improves the reconfigurations costs and infrastructure delay when compared to purist service installations.

It has been already mentioned that the number of electronic control units on a vehicle has increased. These interconnected ECUs along with the assisting sensors and actuators are generally classified as Cyber-Physical System (CPS). Within the CPS a vehicle can be interpreted as a resource for sensing, data storage, computing, data relaying, and a means for locating other objects [1]. However, it is a trade-off between sophisticated sensors and computational on-board and the QoS communications capabilities. Less sophisticated on-board systems would require more cloud-based assistance hence higher QoS communications and vice versa. The focus

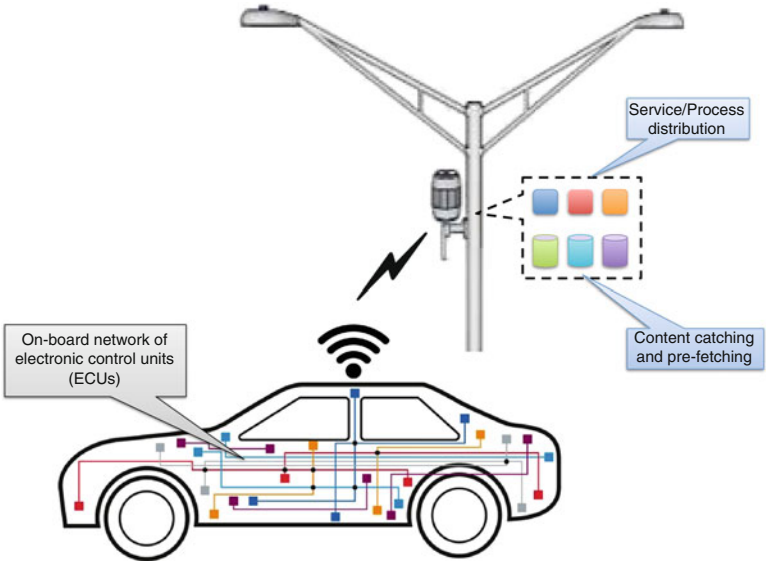


**Fig. 10** Vehicle and RSU cyber-physical ecosystem

should be shifted from bringing "the internet in the vehicle" to bringing "the vehicle in the internet", converting the vehicles from a simple content consumers to a content generators. This will also increase the need for high-speed links in up-link transmission for both vehicle-to-vehicle and vehicle-to-infrastructure. In the absence of infrastructure, either due to early-state of deployment or due to network malfunction, vehicles can also form a vehicular cloud [17]. For example, parked vehicles can be used as a temporary network and storage infrastructure which can increase the resilience and reliability of the vehicular network.

## 5 Summary

The fifth generation of mobile communications systems is evangelising intelligent systems that work together to give the impression of unlimited data capacity, impressive speeds and impeccable efficiency compared to current 4G networks. In addition, these kind of services would also require guaranteed performance all the way from the cloud that runs the service to the vehicle, and back again. These characteristics will enable new types of applications such as connected autonomous vehicles, which are at the moment infeasible due to constraints such as latency or capacity limitations of current vehicular communication systems. The future connected autonomous vehicles will lead to fewer accidents and improve fuel efficiency as huge volumes of data can be fuzzed in the cloud and provide

real-time information to large amount of vehicles. Wideband connectivity will enable operating system updates, driving software, and support next-generation applications.

Key enablers of the future vehicular networking in 5G are analysed in this chapter. Cloud computing and SDN are two of the driving forces for the 5G core network where resources can be dynamically orchestrated and re-configured fast. In the access network current DSRC and 4G technologies should be optimised and cooperate with the new 5G radios. Vehicles and RSUs will also play a significant role in the network operations with service distribution and content caching, apart from being just the end-user of those services and contents. Finally, what is required for 5G is a series of trade-offs to enable a balance between the IP approach and the ICN/NDN approach possibly as a regionalised NDN approach with classical IP-based policies between local domains.

**Acknowledgements** The authors would like to thank Jerry Foster, 5G systems architect at 5GIC, for sharing his wisdom during the conceptual 5G architecture definition.

## References

1. S. Abdelhamid, H. Hassanein, G. Takahara, Vehicle as a resource (VaaR). IEEE Netw. **29**(1), 12–17 (2015). doi:10.1109/MNET.2015.7018198
2. S. Andreev, M. Gerasimenko, O. Galinina, Y. Koucheryavy, N. Himayat, S.P. Yeh, S. Talwar, Intelligent access network selection in converged multi-radio heterogeneous networks. IEEE Wirel. Commun. **21**(6), 86–96 (2014). doi:10.1109/MWC.2014.7000976
3. R. Baldessari, A. Festag, J. Abeille, Nemo meets vanet: a deployability analysis of network mobility in vehicular communication, in *7th International Conference on ITS Telecommunications, 2007. ITST '07* (2007), pp. 1–6. doi:10.1109/ITST.2007.4295897
4. S. Bayless, A. Guan, *Connected Vehicle Technical Insigts - Vehicle Applications and Wireless Interoperability*. Technology Scan Series 2011–2015 (ITS America, Washington, DC, 2015)
5. F. Bonomi, R. Milito, J. Zhu, S. Addepalli, Fog computing and its role in the internet of things, in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, MCC '12* (2012), pp. 13–16. doi:10.1145/2342509.2342513
6. J. Calabuig, J. Monserrat, D. Gozalvez, O. Klemp, Safety on the roads: LTE alternatives for sending ITS messages. IEEE Veh. Technol. Mag. **9**(4), 61–70 (2014). doi:10.1109/MVT.2014.2362272
7. G. Calandriello, P. Papadimitratos, J.P. Hubaux, A. Lioy, On the performance of secure vehicular communication systems. IEEE Trans. Dependable Secure Comput. **8**(6), 898–912 (2011). doi:10.1109/TDSC.2010.58
8. M. Chen, D.O. Mau, Y. T. Zhang, T. Taleb, V.C. Leung, VENDNET: VEhicular Named Data NETwork. Veh. Commun. **1**(4), 208–213 (2014). doi:10.1016/j.vehcom.2014.002
9. X. Cheng, X. Hu, L.L. Yang, I. Husain, K. Inoue, P. Krein, R. Lefevre, Y. Li, H. Nishi, J. Taiber, F.Y. Wang, Y. Zha, W. Gao, Z. Li, Electrified vehicles and the smart grid: the ITS perspective. IEEE Trans. Intell. Transp. Syst. **15**(4), 1388–1404 (2014). doi:10.1109/TITS.2014.2332472
10. N. Cheng, N. Lu, N. X.S. Zhang, Shen, J.W. Mark, Vehicular wifi offloading: challenges and solutions. Veh. Commun. **1**(1), 13–21 (2014)
11. X. Cheng, L. Yang, X. Shen, D2D for intelligent transportation systems: a feasibility study. IEEE Trans. Intell. Transp. Syst. **PP**, 1–10 (2015). doi:10.1109/TITS.2014.2377074

12. Cisco Systems Inc., The internet of things - how the next evolution of the internet is changing everything. White paper (2011)
13. Cisco Systems Inc., Cisco visual networking index: global mobile data traffic forecast update, 2012–2017. Technical Report (2013)
14. Cisco Systems Inc., Fog computing. Cisco Technology Radar (2014)
15. CoCar Consortium, CoCarX Coperative Cars eXtended ITS services and communication architecture. Deliverable D3 (2011)
16. COST IC 1004, Scientific challenges towards 5G mobile communications. White paper (2013)
17. F. Dressler, P. Handle, C. Sommer, Towards a vehicular cloud - using parked vehicles as a temporary network and storage infrastructure, in *International Workshop on Wireless and Mobile Technologies for Smart Cities, WiMobCity '14* (2014), pp. 11–18. doi:10.1145/2633661.2633671
18. X. Duan, X. Wang, Authentication handover and privacy protection in 5g hetnets using software-defined networking. IEEE Commun. Mag. **53**(4), 28–35 (2015). doi:10.1109/MCOM.2015.7081072
19. T. ElBatt, S.K. Goel, G. Holland, H. Krishnan, J. Parikh, Cooperative collision warning using dedicated short range wireless communications, in *Proceedings of the 3rd International Workshop on Vehicular Ad Hoc Networks* (2006), pp. 1–9. doi:10.1145/1161064.1161066
20. EN 302 665, Intelligent Transport Systems (ITS) - Communications Architecture. European standard, ETSI (2010)
21. ES 202 663, Intelligent Transport Systems (ITS) - European profile standard for the physical and medium access control layer of Intelligent Transport Systems operating in the 5GHz frequency band. European standard, ETSI (2009)
22. D. Farinacci, V. Fuller, D. Meyer, D. Lewis, The Locator/ID Separation Protocol (LISP). RFC 6830 (Experimental) (2013). http://www.ietf.org/rfc/rfc6830.txt
23. G. Fettweis, S. Alamouti, 5G: personal mobile internet beyond what cellular did to telephony. IEEE Commun. Mag. **52**(2), 140–145 (2014). doi:10.1109/MCOM.2014.6736754
24. Fierce Wireless, 3G/4G wireless network latency: comparing Verizon, AT&T, Sprint and T-Mobile in February 2014 (2014). http://goo.gl/OkiHl6
25. D. Filev, J. Lu, D. Hrovat, Future mobility: integrated vehicle control with cloud computing. ASME Dyn. Syst. Control Mag. **1**(1), 18–24 (2013)
26. FlexRay Consortium, Flexray communications system-protocol specification. Technical Report (2005)
27. S. Gaudin, Ford uses Microsoft cloud to seamlessly update cars (2015), http://goo.gl/1dWhrL. Accessed: 15 May 2015
28. M. Gerla, E.K. Lee, G. Pau, U. Lee, Internet of vehicles: from intelligent grid to autonomous cars and vehicular clouds, in *2014 IEEE World Forum on Internet of Things (WF-IoT)* (2014), pp. 241–246. doi:10.1109/WF-IoT.2014.6803166
29. G. Grassi, D. Pesavento, G. Pau, L. Zhang, S. Fdida, Navigo: interest forwarding by geolocations in vehicular named data networking. ArXiv e-prints (2015)
30. L. Gu, D. Zeng, S. Guo, Vehicular cloud computing: a survey, in *2013 IEEE Globecom Workshops (GC Wkshps)* (2013), pp. 403–407. doi:10.1109/GLOCOMW.2013.6825021
31. H. Hawilo, A. Shami, M. Mirahmadi, R. Asal, NFV: state of the art, challenges, and implementation in next generation mobile networks (vepc). IEEE Netw. **28**(6), 18–26 (2014). doi:10.1109/MNET.2014.6963800
32. IEEE, IEEE Guide for Wireless Access in Vehicular Environments (WAVE) - Architecture. IEEE Std. 1609.0-2013 (2014), pp. 1–78. doi:10.1109/IEEESTD.2014.6755433
33. A. Imran, A. Zoha, Challenges in 5g: how to empower son with big data for enabling 5g. IEEE Netw., **28**(6), 27–33 (2014). doi:10.1109/MNET.2014.6963801
34. V. Jacobson, D.K. Smetters, J.D. Thornton, M.F. Plass, N.H. Briggs, R.L. Braynard, Networking named content, in *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies, CoNEXT '09* (2009), pp. 1–12. doi:10.1145/1658939.1658941

35. A. Jain, H.S. Koppula, B. Raghavan, A. Saxena, Know before you do: anticipating maneuvers via learning temporal driving models. arXiv preprint arXiv:1504.02789 (2015)
36. W. Jones, BMW fixes software flaw that would've let hackers unlock doors. IEEE Spectrum (2015). http://goo.gl/Io9Yht
37. K. Kanai, T. Muto, H. Kisara, J. Katto, T. Tsuda, W. Kameyama, Y.J. Park, T. Sato, Proactive content caching utilizing transportation systems and its evaluation by field experiment, in *2014 IEEE Global Communications Conference (GLOBECOM)* (2014), pp. 1382–1387. doi:10.1109/GLOCOM.2014.7037001
38. K. Katsaros, R. Kernchen, M. Dianati, D. Rieck, C. Zinoviou, Application of vehicular communications for improving the efficiency of traffic in urban areas. Wirel. Commun. Mob. Comput. **11**(12), 1657–1667 (2011)
39. K. Katsaros, M. Dianati, L. Le, Effective implementation of location services for VANETs in hybrid network infrastructures, in *International Conference on Communications (ICC) Workshops* (2013), pp. 521–525. doi:10.1109/ICCW.2013.6649289
40. K. Katsaros, M. Dianati, R. Tafazolli, G. Xiaolong, End-to-End delay bound analysis for location-based routing in hybrid vehicular networks. IEEE Trans. Veh. Technol. **PP**(99), 1–1 (2015). doi: 10.1109/TVT.2015.2482362
41. T. Koponen, M. Chawla, B.G. Chun, A. Ermolinskiy, K.H. Kim, S. Shenker, I. Stoica, A data-oriented (and beyond) network architecture. SIGCOMM Comput. Commun. Rev. **37**(4), 181–192 (2007). doi:10.1145/1282427.1282402
42. KPMG, Connected and autonomous vehicles—the uk economic opportunity. Technical Report (2015)
43. J.H. Lee, T. Ernst, N. Chilamkurti, Performance analysis of PMIPv6-based network mobility for intelligent transportation systems. IEEE Trans. Veh. Technol. **61**(1), 74–85 (2012). doi:10.1109/TVT.2011.2157949
44. E. Lee, E.K. Lee, M. Gerla, S. Oh, Vehicular cloud networking: architecture and design principles. IEEE Commun. Mag. **52**(2), 148–155 (2014). doi:10.1109/MCOM.2014.6736756
45. P. Matzakos, J. Härri, B. Villeforceix, C. Bonnet, An IPv6 architecture for cloud-to-vehicle smart mobility services over heterogeneous vehicular networks, in *3rd International Conference on Connected Vehicles & Expo ICCVE 2014* (2014)
46. C. McCarthy, K. Harnett, A. Carter, Characterization of potential security threats in modern automobiles - a composite modeling approach. DOT HS 812 074, NHTSA (2014)
47. METIS Consortium, Proposed solutions for new radio access. Deliverable D2.4 (2015)
48. Z.H. Mir, F. Filali, LTE and IEEE 802.11p for vehicular networking: a performance evaluation. EURASIP J. Wirel. Commun. Netw. **2014**(89) (2014). doi:10.1186/1687-1499-2014-89
49. A. Mukherjee, S. Fakoorian, J. Huang, A. Swindlehurst, Principles of physical layer security in multiuser wireless networks: a survey. IEEE Commun. Surv. Tutorials **16**(3), 1550–1573 (2014). doi:10.1109/SURV.2014.012314.00178
50. NHTSA, Preliminary statement of policy concerning automated vehicles (2013)
51. Nokia, 5g radio access system design aspects. White paper (2015)
52. N. Omheni, F. Zarai, M.S. Obaidat, K.F. Hsiao, L. Kamoun, A novel media independent handover-based approach for vertical handover over heterogeneous wireless networks. Int. J. Commun. Syst. **27**(5), 811–824 (2014). doi:10.1002/dac.2628
53. B. Peng, T. Peng, Z. Liu, Y. Yang, C. Hu, Cluster-based multicast transmission for device-to-device (d2d) communication, in *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*, pp. 1–5 (2013). doi:10.1109/VTCFall.2013.6692134
54. J. Petit, Z. Mammeri, Authentication and consensus overhead in vehicular ad hoc networks. Telecommun. Syst. **52**(4), 2699–2712 (2013). doi:10.1007/s11235-011-9589-y
55. M. Phan, R. Rembarz, S. Sories, A capacity analysis for the transmission of event and cooperative awareness messages in lte networks, in *ITS World Congress* (2011)
56. J. Qiao, X. Shen, J. Mark, Q. Shen, Y. He, L. Lei, Enabling device-to-device communications in millimeter-wave 5g cellular networks. IEEE Commun. Mag. **53**(1), 209–215 (2015). doi:10.1109/MCOM.2015.7010536

57. G. Remy, S. Senouci, F. Jan, Y. Gourhant, Lte4v2x: Lte for a centralized vanet organization, in *2011 IEEE Global Telecommunications Conference (GLOBECOM 2011)*, pp. 1–6 (2011). doi:10.1109/GLOCOM.2011.6133884
58. N. Ristanovic, P. Papadimitratos, G. Theodorakopoulos, J.P. Hubaux, J.Y. Le Boudec, Adaptive message authentication for multi-hop networks, in *2011 Eighth International Conference on Wireless On-Demand Network Systems and Services (WONS)* (2011), pp. 96–103. doi:10.1109/WONS.2011.5720206
59. P.E. Ross, Cars that talk need wireless that works. IEEE Spectrum (2014). http://goo.gl/juFrIA
60. S. Ruehrup, P. Fuxjaeger, D. Smely, TCP-like congestion control for broadcast channel access in VANETs, in *International Conference on Connected Vehicles and Expo (ICCVE)* (2014)
61. S. Ryu, K.J. Park, J.W. Choi, Enhanced fast handover for network mobility in intelligent transportation systems. IEEE Trans. Veh. Technol. **63**(1), 357–371 (2014). doi:10.1109/TVT.2013.2272059
62. M. Salahuddin, A. Al-Fuqaha, M. Guizani, Software-defined networking for rsu clouds in support of the internet of vehicles. IEEE Internet Things J. **2**(2), 133–144 (2015). doi:10.1109/JIOT.2014.2368356
63. R. Sivaraj, A. Gopalakrishna, M. Chandra, P. Balamuralidhar, Qos-enabled group communication in integrated vanet-lte heterogeneous wireless networks, in *2011 IEEE 7th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)* (2011), pp. 17–24. doi:10.1109/WiMOB.2011.6085417
64. F. Soldo, C. Casetti, C. Chiasserini, P. Chaparro, Video streaming distribution in vanets. IEEE Trans. Parallel Distrib. Syst. **22**(7), 1085–1091 (2011). doi:10.1109/TPDS.2010.173
65. STOKE, Latency considerations in LTE, Implications to security gateway. White Paper (2014)
66. TR 102 638, Intelligent Transport Systems (ITS) - Vehicular communications - Vehicular communications; basic set of applications; definitions. Technical Report, ETSI (2009)
67. TR 102 692, Intelligent Transport Systems (ITS) - Framework for public mobile networks in Cooperative ITS (C-ITS). Technical report, ETSI (2012)
68. TS 102 636-6-1, Intelligent Transport Systems (ITS) - Vehicular communications - GeoNetworking - Part 6: internet integration - Sub-part 1: Transmission of IPv6 packets over GeoNetworking protocols. Technical specification, ETSI (2011)
69. TS 102 687, Intelligent Transport Systems - Decentralized congestion control mechanisms for intelligent transport systems operating in the 5GHz range; access layer part. Technical specification, ETSI (2011)
70. TS 102 940, Intelligent Transport Systems - Security - ITS communications security architecture and security management. Technical specification, ETSI (2012)
71. L. Wang, R. Wakikawa, R. Kuntz, R. Vuyyuru, L. Zhang, Data naming in vehicle-to-vehicle communications, in *2012 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (2012), pp. 328–333. doi:10.1109/INFCOMW.2012.6193515
72. X. Wang, M. Chen, T. Taleb, A. Ksentini, V. Leung, Cache in the air: exploiting content caching and delivery techniques for 5G systems. IEEE Commun. Mag. **52**(2), 131–139 (2014). doi:10.1109/MCOM.2014.6736753
73. M. Whaiduzzaman, M. Sookhak, A. Gani, R. Buyya, A survey on vehicular cloud computing. J. Netw. Comput. Appl. **40**, 325–344 (2014). doi:10.1016/j.jnca.2013.08.004
74. X. Wu, S. Subramanian, R. Guha, R.G. White, J. Li, K.W. Lu, A. Bucceri, T. Zhang, Vehicular communications using DSRC: challenges, enhancements, and evolution. IEEE J. Sel. Areas Commun. **31**(9), 399–408 (2013). doi:10.1109/JSAC.2013.SUP.0513036
75. Z. Yan, S. Zeadally, Y.J. Park, A novel vehicular information network architecture based on named data networking (ndn). IEEE Internet Things J. **1**(6), 525–532 (2014). doi:10.1109/JIOT.2014.2354294

76. V. Yazici, U. Kozat, M.O. Sunay, A new control plane for 5g network architecture with a case study on unified handoff, mobility, and routing management. IEEE Commun. Mag. **52**(11), 76–85 (2014). doi:10.1109/MCOM.2014.6957146
77. A. Yegin, J. Park, K. Kweon, J. Lee, Terminal-centric distribution and orchestration of ip mobility for 5g networks. IEEE Commun. Mag. **52**(11), 86–92 (2014). doi:10.1109/MCOM.2014.6957147
78. X. Yin, X. Ma, K. Trivedi, A. Vinel, Performance and reliability evaluation of BSM broadcasting in DSRC with multi-channel schemes. IEEE Trans. Commun. **63**(12), 3101–3113 (2014). doi:10.1109/TC.2013.175
79. L. Zeng, Y. Zhu, 3g-assisted routing in vehicular networks, in *2012 IEEE Sensors* (2012), pp. 1–4. doi:10.1109/ICSENS.2012.6411481
80. L. Zhang, A. Afanasyev, J. Burke, V. Jacobson, K. Claffy, P. Crowley, C. Papadopoulos, L. Wang, B. Zhang, Named data networking. SIGCOMM Comput. Commun. Rev. **44**(3), 66–73 (2014). doi:10.1145/2656877.2656887
81. N. Zhang, N. Cheng, A. T. Gamage, K. Zhang, J.W. Mark, X.S. Shen, Cloud assisted HetNets toward 5G wireless networks. IEEE Commun. Mag. **53**(6), 59–65 (2015). doi:10.1109/MCOM.2015.7120046
82. Q. Zhao, Y. Zhu, C. Chen, H. Zhu, B. Li, When 3g meets vanet: 3g-assisted data delivery in vanets. IEEE Sensors J. **13**(10), 3575–3584 (2013). doi:10.1109/JSEN.2013.2265304
83. K. Zhu, D. Niyato, P. Wang, E. Hossain, D.I. Kim, Mobility and handoff management in vehicular networks: a survey. Wirel. Commun. Mob. Comput. **11**(4), 459–476 (2011). doi:10.1002/wcm.853

# Communications Protocol Design for 5G Vehicular Networks

**Francesco Chiti, Romano Fantacci, Dino Giuli, Federica Paganelli, and Giovanni Rigazzi**

**Abstract** This chapter provides an overview on existing standards in vehicular networking and highlights new emerging trends towards an integrated infrastructure based on the interworking of heterogeneous technologies. Next-Generation Mobile Vehicular Networks are first characterized by providing an insight on relevant stable standards in wireless communications technologies, with a special focus on Heterogeneous Vehicular Networks. Furthermore, the chapter discusses a general framework supporting opportunistic networking scheme and outlines novel application and use cases based on social- and context-awareness paradigms.

## 1 Introduction

The automotive industry is a challenging domain for the integration and evolution of heterogeneous Information and Communication Technologies (ICTs). Indeed, in order to provide drivers with a more pleasant and safer driving experience, vehicles are gradually endowed with sensing, actuation, computing, user-interaction and communications capabilities. Boosted by marketing and commercial competition, vehicles are thus becoming *intelligent nodes* capable of interconnecting, cooperating and even autonomously adapting to the surrounding environment. Since drivers usually spend a non-negligible fraction of time in vehicles and their behaviour may have a strong impact at a collective scale (e.g., pollution and noise in a city), the adoption of advanced ICT approaches for achieving a more safe, green and effective mobility management is still an open research issue. Moreover, the familiarity with the *social networking* paradigm makes people used to share information on-line with (in)direct *contacts* and use it in their everyday life depending on the degree of trust.

F. Chiti • R. Fantacci (✉) • D. Giuli • G. Rigazzi
Department of Information Engineering, University of Florence, Florence, Italy
e-mail: francesco.chiti@unifi.it; romano.fantacci@unifi.it; dino.giuli@unifi.it; giovanni.rigazzi@unifi.it

F. Paganelli
National Interuniversity Consortium for Telecommunications, Research Unit at the University of Florence, Florence, Italy
e-mail: federica.paganelli@unifi.it



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_23

It is worth noticing that this vision inherently deals with multimedia information, as it aggregates, process and disseminates sensed data (originating both on board and on the road side) integrated with warning/alarms, navigation suggestions, and context details that come from a specific area or even the Internet.

The widespread adoption of existing social-based traffic and navigation applications, paves the way for introducing novel vehicular services and applications, that are typically location-based as well as community-driven. To this purpose, wireless communications may enable the effective exchange of *mobile* information among vehicles in a participatory way to enhance open information sharing and knowledge exchange processes. The actual trends in mobile networking are: (1) to extend services offered by traditional providers with unlimited peer-to-peer capabilities, and (2) to dynamically integrate available communications segments to exploit all communication opportunities. In this context, the crucial point is how to guarantee high-capacity and reliable connectivity without affecting the system scalability and cost-effectiveness. Due to the attractive benefits in terms of enhanced data rate and ultra-low latency, 5G represents the most promising solution to enable novel and efficient Intelligent Transport System (ITS) applications and to support the internetworking with existing technologies.

In this chapter, we provide an overview on existing standards in vehicular networking and highlight new emerging trends towards an integrated infrastructure based on the interworking of heterogeneous technologies. Section 2 introduces Next-Generation Mobile Vehicular Networks by providing an insight on relevant stable standards in wireless communications technologies, and reviewing emerging 5G vehicular communication solutions, with a special focus on Heterogeneous Vehicular NETworks (HetVNET). As a preliminary result towards this integrated networking scheme, we propose a general framework supporting opportunistic networking scheme as overlay of an LTE-A system. In Sect. 3 a review on game theory inspired approaches for clustering schemes is provided. Finally, Sect. 4 provides an overview on novel application and use cases based on social- and context-awareness that are enabled by the emerging technological paradigms in Vehicular Networking.

## 2 Next-Generation Mobile Vehicular Networks

Enabling inter-vehicle communications paves the way to a plethora of novel applications for Intelligent Transportation Systems (ITS), where the main goal is not only improving safety and efficiency of transportation systems, but also providing multimedia contents and information to the mobile users without the support of cellular networks. Based on the application area, vehicular applications can be classified as follows:

1. *Traffic management applications*. To address traffic and environmental related issues, as bottlenecks and fuel consumption, traffic management applications aim at improving the vehicle traffic flow, traffic coordination and traffic assistance, and at providing updated local information, maps and information of relevance

bounded in space and time. This type of time-to-live traffic application is usually used in vehicle-to-backoffice or vehicle-to-roadside scenarios. They may use beacons or alerts, a multihop position-based communication regime, and their latency cannot be higher than 400 ms, whereas the packet delivery ratio cannot be lower than 95 %.

2. *Safety applications* Minimizing the probability of traffic accidents and avoiding collisions situations that most frequently occur between vehicles and other objects such as animals, trees, and pedestrians, are the end goals of safety applications. This type of applications relies on real-time information and uses a vehicle-to-vehicle communication scheme. They provide information and assistance to drivers to avoid traffic accidents, which are mainly caused by human errors [32]. Vehicles and roadside units share information, which is then used to predict a dangerous situation. Moreover, this information is used to locate dangerous locations on roads. They use beacon messages, a single-hop position-based or fast-bidirectional communication regime, and their latency cannot exceed 100 ms, whereas the packet delivery ratio cannot be lower than 99 %. Some examples of safety applications include: intersection collision warning, lane change assistance, overtaking vehicle warning, head on collision warning, and emergency vehicle warning.
3. *Driver assistance and infotainment applications* These applications provide services such as comfort and driving assistance. This class of applications attempts to support all features needed by drivers and passengers for a convenient travel. Driver assistance applications provide information about repair notifications, remote diagnostics, context information, navigation information, and alerts. These applications usually use vehicle-to-backoffice or vehicle-to-roadside communication. They utilize normal messages and bidirectional communication; their latency cannot be higher than 400 ms, whereas the packet delivery ratio cannot be lower than 95 % [48]. Infotainment applications also are known as in-car comfort entertainment, and they usually do not use inter-vehicular communications. These applications are usually found inside vehicles or at vehicle-to-roadside settings. They use alerts, a multihop position-based communication scheme, and their latency cannot be higher than 400 ms, whereas the packet delivery ratio cannot be lower than 95 % [48]. Applications in this category include cooperative local services and global Internet services.

## *2.1 Standardization Trends and Perspectives*

Wireless communications are experiencing an impressive evolutionary trend, making users able to independently collect and share information through heterogeneous devices and networks. Specifically, the simultaneous availability of multi-radio technologies compels their integration and management towards enhanced *social inspired* networking opportunities.

An interesting case study is represented by the so called Vehicular Ad-hoc Networks (VANETs), which aim at revolutionising the travelling experience especially

within a smart city [34]. The underlying concept is indeed to provide each vehicle with a *wireless identity* which is not simply a *global* identifier, as it happens in IPv6, but it is strictly related with user profiles (drivers and passengers) and the context as well. This allows to increase the interoperability and to *extend the horizon*, thus increasing the safety and optimizing the resources.

However, this vision strictly depends upon the presence of a communications infrastructure providing connectivity to all the vehicles. In the following we briefly overview the candidate technologies [24], spanning from available approaches towards an enhanced paradigm of *ad hoc* interaction that relies upon next generation networks.

### 2.1.1 DSRC-WAVE

The available technology for supporting the Intelligent Transportation Systems (ITS) applications in the short-range communications is represented by Wireless Access in Vehicular Environments (WAVE), also referred to as Dedicated Short Range Communication (DSRC). It is an approved amendment to the IEEE 802.11 standard, whose protocol suite results from the combination of IEEE 802.11p and IEEE 1609. The former focuses on physical and MAC layers for vehicular environments, while resource management, security, networking and multichannel operation are handled according to the IEEE 1609 protocol suite.

IEEE 802.11p is based on the popular IEEE 802.11 WiFi, exploiting the same medium access scheme—based on the exchange of Request to Send/Clear to Send packets (RTS/CTS)—to alleviate packet collisions in case of unicast communications. However, this technique might not be enabled when a packet is sent in broadcast and the packet collision probability is consequently increased. Furthermore, the quality of service (QoS) support is provided by using the enhanced distributed channel access (EDCA) scheme, which cannot guarantee high packet delivery probability due to the huge number of collisions generated by high priority packets, characterized by a small and fixed Contention Window (CW) size. The entire 75-MHz bandwidth is divided into seven channels: one channel, i.e., Control Channel (CCH), is only used to convey emergency and safety traffic, while six Service Channels (SCH) are occupied by infotainment and multimedia traffic transmission. Further, a channel switching scheme allows the devices to tune to CCH during the CCH time interval and then switch to a specific service channel during the SCH interval.

### 2.1.2 ISO CALM

Continuous Air-interface Long and Medium range (CALM) is the protocol suite standardized by ISO/TC 204 WG for Europe [30]. The CALM CI (Communication Interface) provides all the functionalities to manage the physical and link layers and can support diverse wireless technologies, ranging from infrared communications to cellular and satellite communications. Moreover, the CALM networking layer

coordinates network and transport layers and can either employ IPv6 mobility support protocols or the CALM FAST protocol to perform unicast and broadcast transmissions and store and forward operations.

### 2.1.3 C2C-CC

Car2Car Communication Consortium (C2C-CC) represents an European open industrial standard mostly focused on (non) safety applications development. It proposed the C2CNet architecture which is not completely IPv6 oriented, while supporting multi-hop communications through ad hoc designed routing protocols and vehicles beaconing [30]. Besides, it is inherently conceived for a multi-radio environments supporting both IEEE 802.11a/b/g/n, IEEE 802.11p and cellular technologies.

### 2.1.4 B3G Cellular Networks

Wide-area cellular networks, especially 4G Long Term Evolution-Advanced (LTE-A), are expected to directly support vehicular applications by means of self-organizing femto-cells, traffic shaping and *vertical* roaming across different radio access technologies (RATs). Despite limitations on the radio access flexibility and on the resource allocation efficiency, especially in the case of small data transmission, LTE-A networks offer benefits in terms of spectral efficiency, cost reduction and capacity. Furthermore, 4G systems intrinsically guarantee long operative life, scalability and higher performance, thus representing a remarkable solution to support VANET-like applications. As a consequence, cellular systems are being adopted by numerous car manufacturers with the aim of offering diverse services, e.g., remote vehicle monitoring, infotainment or assisted driving experience.

Along with the Machine Type Communication (MTC) definition, the Third Generation Partnership Project (3GPP) Release 12 introduces in LTE-A the support to Device-to-Device (D2D) or direct-mode communications, enabling P2P transmission between devices in proximity [5]. This functionality fosters a new generation of devices able to communicate with each other, without any kind of human intervention and without involving the cellular infrastructure for the user plane. On the other hand, D2D communications pose also several challenges. First of all, direct mode communications should be established without affecting traditional communications via base station, i.e., by avoiding possible *interferences* with other devices. Secondly, peer and service discovery functionalities need to be introduced, since a device is typically not aware of other terminals in the proximity. Following the paradigm of Multi-hop Cellular Networks (MCN) [28, 33, 35], D2D also provides more flexibility in a cellular network, as direct links enable the communication among the terminals using multiple hops.

The evolutionary trend, presented in Fig. 1, points out a convergence of the existing VANET standards, i.e., WAVE, CALM and C2C-CC, towards an integrated

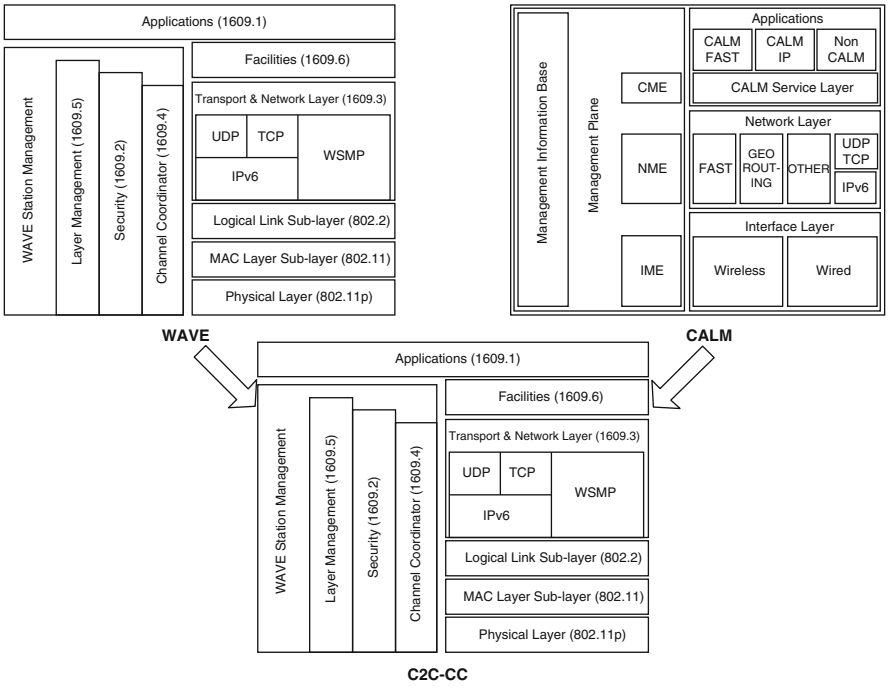


**Fig. 1** Convergence of the existing VANET standards towards an integrated protocol stack, as pointed out in [30]

protocol stack, as it is anticipated in [30]. In particular, it is evident an increasing degree of interoperability with different air interfaces both at the Physical and MAC layers, while the higher layers are expected to be unified by the adoption of IPv6 oriented protocols, according to the Internet of Things (IoT) paradigm.

Mobile network evolution is expected to converge by 2020 into the novel 5G system effectively integrating both wired and wireless segments to accomplish expanded use cases and applications [7, 44]. In particular, new services are expected such as the massive sensor and vehicular-to-anything communications. The latter aspect is involved with the so called safety critical applications, requiring extremely low set-up times and transmission delays.

5G is going to be conceived as an unplanned *ultra dense* system, where nodes are able to self organize and to improve cell performance. This trend is clearly pointed out by the emphasis 3GPP working group is progressively posing on *small* cells, whose limit is represented by the D2D technology, provided an optimal coordination/cooperation among devices/cells. As the system is assumed to be inherently heterogeneous, a higher degree of flexibility is required for both radio access and backhaul networks. At the light of the above considerations, it can be expected that VANET will represent a relevant case study for 5G architecture, effectively integrating existing WAVE, CALM and C2C-CC functional stacks, as highlighted in Fig. 1.

## 2.2 Internetworking with 5G Systems

Next 5G communication systems are expected not only to accommodate the massive growth of the mobile data demand generated by the huge number of future connected devices, but also to support diverse wireless applications and satisfy a new user expectation of the network. Accordingly, 5G will lead to higher data rate and network capacity necessary to cope with the extreme network densification, and will guarantee lower latency (i.e., 1 ms), higher reliability (i.e., 99.999 %) and better energy efficiency [22]. To satisfy all the requirements, close integration of vehicles into cellular networks will be of paramount importance and novel heterogeneous networking solutions will assume vehicles to be able to help the network deliver the traffic as well as generate/consume data.

### 2.2.1 Communication Modes

Due to significant evolution in vehicle technology, future cars will represent the integration of a plethora of emerging information processing, communication, control and electronic systems. Vehicles will be equipped with diverse technologies, including light detection and ranging (LIDAR), radar, advanced cameras and GPS. Intelligent cars will then collect the information about the surrounding environment generated by each sensor and elaborate decisions to make the driving experience more secure and comfortable [47]. In this context, 5G will enable cooperative behavior applications by providing low-latency and high reliable V2X communication, thus disseminating relevant information related to traffic and road condition or sending warning messages in the event of accidents [2]. As described in the next section, vehicles will leverage multi-hop D2D communications, where short-range direct connectivity allows to exchange critical data with minimum delay, while every mobile terminal is also in charge of forwarding the packets in order to communicate with farther terminals without relying on cellular base stations. Moreover, D2D communications will be established on the licensed spectrum, thus guaranteeing high reliability, improved security and easy integration with the cellular network. Fig. 2a shows the D2D links established among close vehicles exchanging information generated by on-board sensors.

Due to the reduced battery power constraint and the easy multiple antenna deployment, vehicular communications also represent an attractive solutions to extend cellular coverage and enhance network capacity for cellular users in the proximity [11]. As shown in Fig. 2b, vehicles can act as mobile Small Cells (mSC) providing better quality of experience for passengers and pedestrians and helping small cells to accommodate the traffic. As a result, drastic network densification can be obtained if effective interference management techniques are employed and vehicle density is sufficient to guarantee adequate data offloading.

Vehicle Content Distribution Networks (VCDN) aim at enabling high-bandwidth content distribution in vehicular networks by opportunistically communicating with

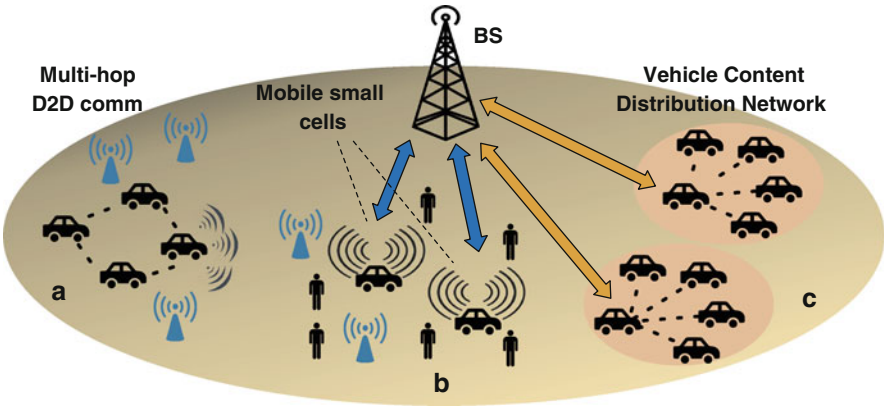


**Fig. 2** 5G vehicular communication solutions: multi-hop D2D communication (**a**), mobile small cell network (**b**) and vehicle content distribution networking (**c**)

close Access Points (AP) to download multimedia content [38]. Fig. 2c illustrates a VCDN within a cellular network. A clustering approach can be applied to select a cluster-head vehicle, which downloads data from the BS and disseminates into the cluster via short-range D2D connectivity. In turn, this approach leads to a reduction in the number of connections vehicle-BS, even though clusters are largely affected by the traffic conditions and vehicle density.

### 2.2.2 Heterogeneous Vehicular Networking

Heterogeneous Networks (HetNet) are a popular and attractive solution to increase the network capacity and enable high user data rates in cellular systems [15]. Usually, HetNets refer to the coexistence of various cells with different coverage, namely, macro, small and femto-cells, which require sophisticated interference cancellation techniques to combat the interference originated from nearby cells. In the context of vehicular networking, Heterogeneous Vehicular NETworks (HetVNET) identify complex systems integrating diverse radio technologies, spanning from IEEE 802.11p DSRC to 3G/4G cellular networks [16]. As an example, an LTE network can guarantee global connectivity and ubiquitous coverage and can help to send multimedia traffic demanding high bandwidth, while DSRC is mostly useful to deliver data generated by safety and early warning applications due to the reduced latency. However, the major challenges associated with the adoption of multi-radio technologies lie in the need for (1) a dynamic support of multiple network configurations, (2) a flexible and effective radio resource management strategy and (3) an efficient mechanism to satisfy the QoS requirements depending on the vehicular application. Inspired by the solutions presented in [48], we envision a 5G vehicular communication system able to perform the joint management of multiple radio technologies and the resource allocation according to the specific

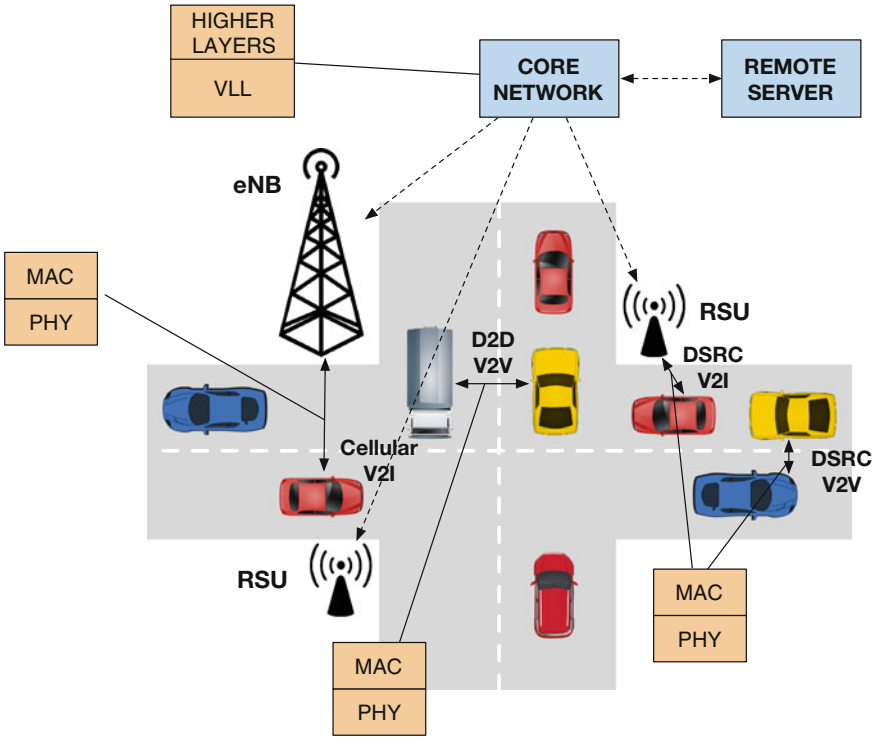


**Fig. 3** Proposed 5G architecture for the integration DSRC-LTE in vehicular applications

application. The proposed 5G architecture is illustrated in Fig. 3. Specifically, we suppose that the core network handles key functionalities, ranging from security functions to packet aggregation, being directly connected to remote servers offering diverse services for vehicular users. Furthermore, the HetVNET RAN consists of both cellular LTE eNBs and DSRC RSUs, which guarantee the communication vehicle-to-infrastructure. We also assumed two possible communication modes:

1. V2I, which can be performed through either the cellular link toward a BS or the DSRC interface by exploiting RSUs deployed along the road.
2. V2V, enabling the direct communication between vehicles via LTE D2D connections or DSRC links.

The radio access technology adaption is then accomplished by including a Virtual Link Layer (VLL) into the core network protocol stack. This layer encompasses all the functionalities required to handle different radio interfaces and allows the load balancing among the different systems involved. Moreover, efficient radio interface coordination helps to guarantee the QoS support for safety and multimedia applications. To this end, we assume that network virtualization methods are conveniently applied to obtain virtual resources [9], which abstract the radio resources and the physical layers of each wireless system within the HetVNET.

## 2.3 Multihop D2D Paradigm

In this section, we address the feasibility issues of an integrated framework able to support opportunistic networking scheme as *overlay* of an LTE-A system as a first step toward the integrated networking envisaged by 5G systems. In particular, we adopted a network-assisted solution to manage (1) network and services discovery, (2) users scheduling and (3) resources allocation, depending on the network configuration. Accordingly, the proposed integrated design consists of three subsequent phases:

1. a D2D *cluster* discovery phase, where the eNB identifies *potential* D2D users,
2. a *graph* formation phase, which, depending on the specific services to be activated, deals with the selection of the optimal routing scheme,
3. a resource allocation algorithm aware of the D2D network configuration previously established.

It could be noticed that the overall process has to be updated depending on the mobility pattern which effects connectivity and service definition. The reference scenario is depicted in Fig. 4, where an eNB serves both pedestrian and vehicular

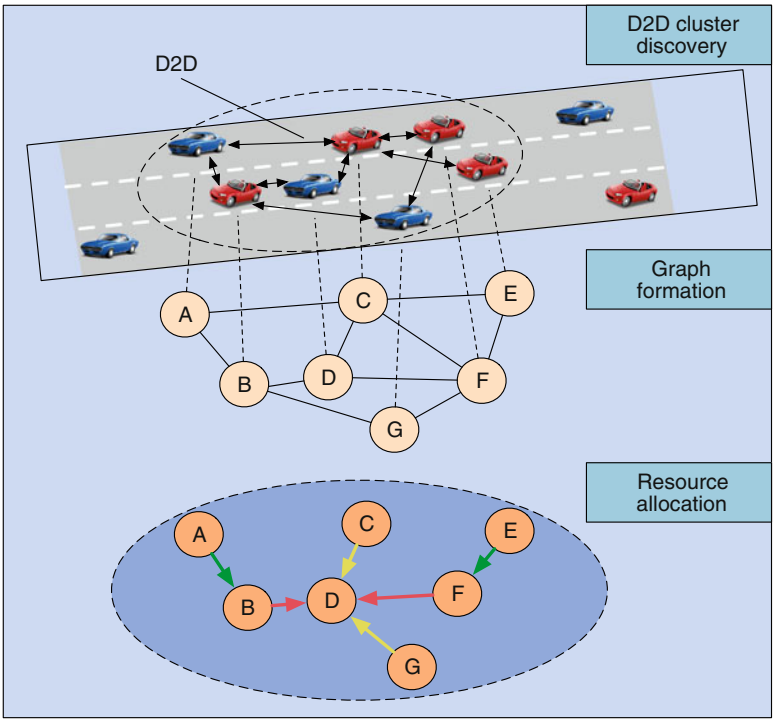


**Fig. 4** Proposed integrated network management frameworks across the specific phases

users. We assume all the devices capable of operating both in *basic* cellular mode, as well as in *direct* mode (MD2D), i.e., by directly exchanging data. We focus on a specific use case, according to which all the node send traffic information to a *unique* collector[1] in charge of data aggregating and transmitting to eNB.[2] Furthermore, due to transmit power constraints, we assume no direct communication between sensing nodes and the collector node, thus implying a multi-hop communication scheme and an optimized routing scheme. For the sake of simplicity, we consider an *overlay* resource allocation, where resources dedicated to cellular and D2D users are mutually orthogonal.

### 2.3.1 D2D Cluster Discovery

As previously mentioned, a network assisted approach, i.e., via the eNB, is a valid solution to support the cluster and services discovery phase since the selection of the D2D candidates can be performed by the network entity. The eNB can decide how clusters are formed by taking into account the context information provided by incoming vehicles (e.g. location, direction, speed, final destination). For instance, vehicles moving towards opposite directions can be partitioned into two different clusters, even though they are close. Moreover, the eNB could also maintain a registry of service and resources made available by cluster members in order to facilitate the dynamic instantiation of *composite* services within a cluster through the steps of service discovery and provider selection. Once a D2D cluster is discovered by collecting the availability of each UE to cooperate and share particular services, an authentication procedure is applied in order to allow direct communication between devices in LTE-A networks. Besides, a direct beaconing is necessary to measure and disseminate the channel state information (CSI) about *all* the *potential* links between cluster members. The eNB is also in charge of assisting a vehicle joining an already formed cluster by taking into account the context information delivered by the vehicle: when a vehicle is approaching an ENB it can, indeed, advertise its position, resource and service capabilities, application and user interests and availability to join D2D clusters. Based on this information, the eNB can select the active clusters that best match the context information and extend these D2D clusters accordingly.[3]

[1] The collector selection can be performed by the eNB or by the D2D users, according to a centralized or distributed approach, respectively.

[2] We assumed that D2D communications can be performed only during the *uplink* frame.

[3] It can be noticed that this procedure might require the cooperation of the vehicle itself and additional eNBs due to roaming.

### 2.3.2 Graph Construction

Once the information regarding the cooperating nodes has been collected by the eNB, the network topology can be selected, depending on the specific use case, as explained din Sect. 4.3. Then this topology is optimized according to a particular target function. It is worth noting that in VSNs all the possible topologies are inherently data centric, since the kind of information affects the message passing scheme and the *logical* connectivity as well. For instance, data aggregation could be effectively performed via a *tree* topology where a single node is in charge of collecting and refining data sensed by other ones and to eventually deliver to the eNB. To this end, Routing Protocol for Low-power and Lossy networks (RPL) is a widely adopted solution to establish a convenient loop free route toward a given destination [1].

### 2.3.3 Network-Aware Resource Allocation Algorithm

Once the topology has been established, a serious drawback is represented by potential bottlenecks, which for a tree topology involve nodes with lower rank value, i.e., nodes closer to the root (collector).[4]

An effective resource allocation strategy should take into account the rank of each node, providing more resources to the nodes closest to the collector. A heuristic strategy could assign resources according to a *proportional fair* scheduling approach, by setting the scheduling weights $w_i$ for the $i$-th node taking into account the its estimated rate $r_i$[5] as well as the rate of the sub-tree $\mathscr{T}_i$ rooted at the $i$-th node ($\mathscr{T}$ represents the overall tree). In particular, we have:

$$w_i \doteq \frac{r_i + \sum_{j \in \mathscr{T}_i} r_j}{\sum_{k \in \mathscr{T}} r_k + \sum_{j \in \mathscr{T}_k} r_j} \tag{1}$$

### 2.3.4 Performance Analysis

We investigated the performance achievable by the proposed approach by focusing on a low-to-moderate mobility scenario, where only *one* Resource Block (RB) is reserved for VANET traffics with a minimum power transmission and a robust Modulation and Coding Scheme (MCS).[6] We referred to an anycast data dissemination, as explained in Fig. 8, where, depending on the mobility patterns, several topologies are possible.

[4]This problem is commonly addressed by means of the *min-cut graph analysis*.

[5]This information is collected by the eNB during the services discovery phase.

[6]In particular, we adopted the MCS 5, according to LTE standard.

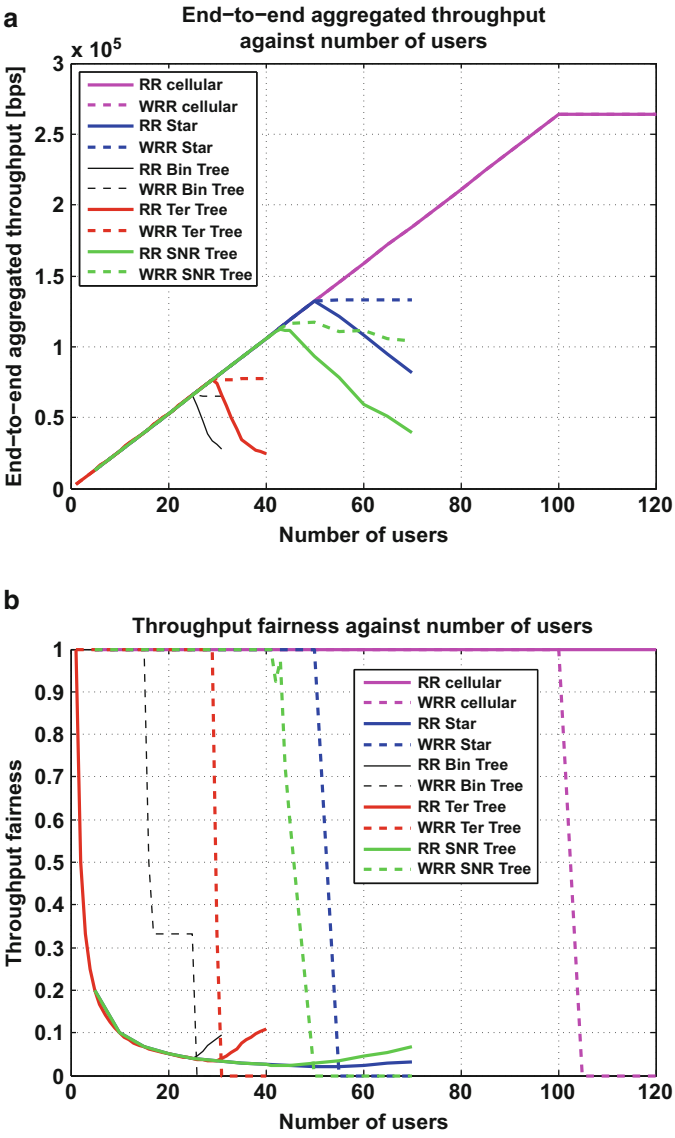


**Fig. 5** End-to-end aggregated throughput and fairness comparisons for a practical case study

In Fig. 5 the performance in terms of throughput and fairness, respectively, are investigated for different network topologies, including basic cellular mode, single hop D2D (star), multihop D2D with binary, ternary or stochastic trees. Besides, the network unaware (RR) and aware (WRR) scheduling approaches have been compared. It can be noticed a *linear* throughput increasing for all the topologies, while in saturation conditions WRR scheduling scheme achieves better performance

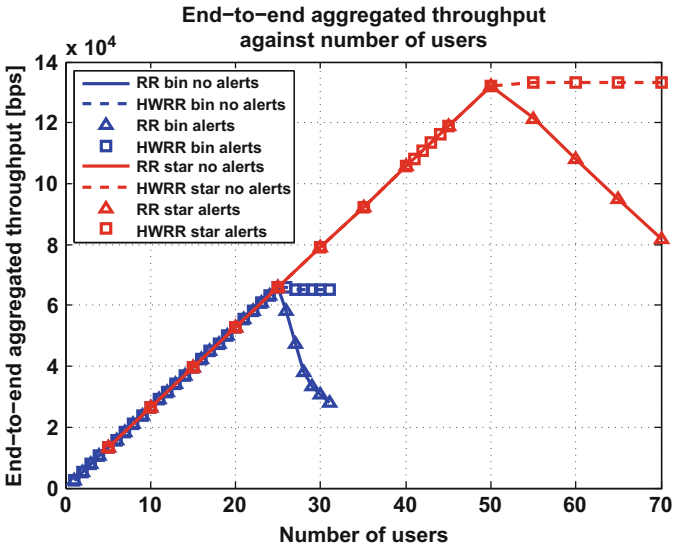


**Fig. 6** Throughput performance in the case of a star and a binary tree topologies with and without alarm flow for WRR and HWRR algorithms

than RR. In addition, it is worth pointing out that the availability of one RB allows the effective management of at most 100 vehicles arranged in a star topology, while the *effective* cluster size decrease for tree topology. Further, WRR is more fair than RR, even though in saturation WRR implies the *starvation* effect.

We investigated also a different scenario where two traffic flows are considered, a background *data* and a sporadic *alarm* flows, with different rates (specifically data flows is 100 times higher than the other one) and priorities. In Fig. 6, the throughput performance are investigated in the case of a star a binary tree topologies with and without alarm flow for WRR and HWRR algorithms, pointing out an unchanged network capacity. However, in Fig. 7, the delivery delay performance are also evaluated for WRR and HWRR algorithms, respectively, highlighting that the latter approach is more capable of handling flow priority and matching time constraints.

## 3 Advanced Communications Protocol Design

### *3.1 Small-World Networking*

The topic of D2D communications has been recently investigated in several possible applications, as data offloading [3], cell coverage extension [18], content sharing [37], and disaster area alerting [33] with a specific focus mainly on downlink

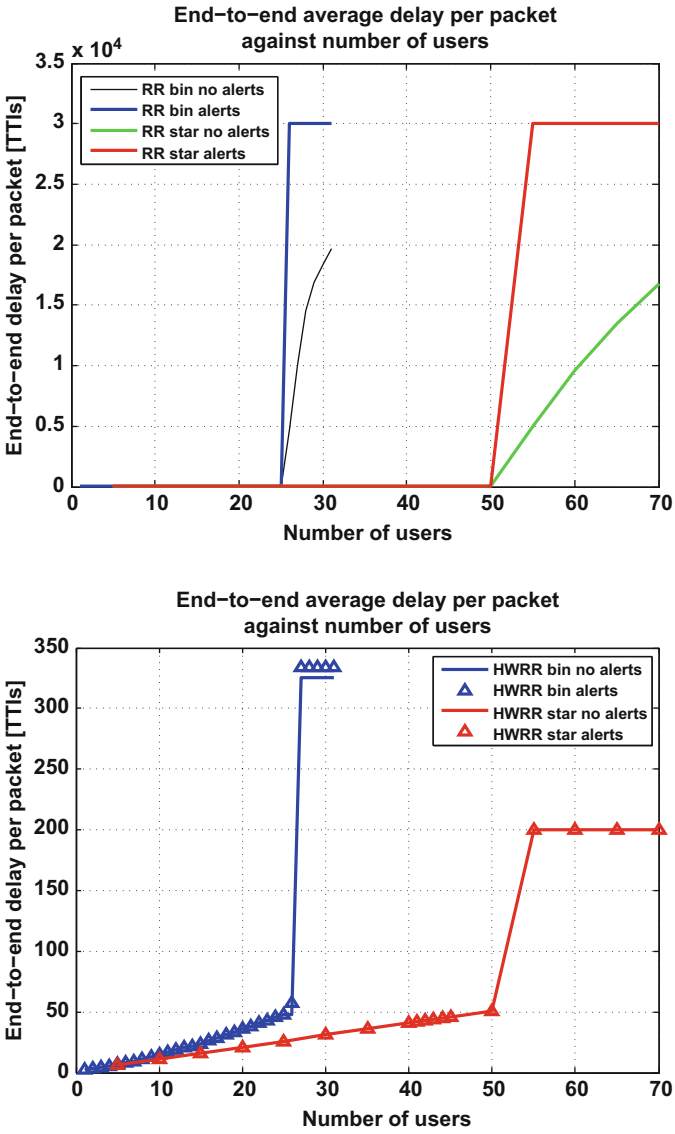


**Fig. 7** End-to-end delay per packet performance in the case of a star and a binary tree topologies with and without alarm flow for WRR and HWRR algorithms

issues. However, these contributions are limited to the case of single hop scheme, while the more general case of multihop D2D communications still deserves more efforts. In [12] network-assisted D2D communication is addressed with an analysis on power control limited to a more traditional two-hop scenario. Conversely,

multihop D2D communication is considered in [35] for end-to-end Machine-to-Machine connectivity, deriving a topology aware resource allocation scheme for overlay networks and avoiding co-channel interference.

Despite these contributions, the development of suitable protocols which effectively support information sharing is still an open issue. Recent research on large scale self-organized distributed networks, as well as in social communities, suggests an alternative perspective for routing protocol design [25]. It relies on the evidence of scale-free properties for typical VANET deployments [31], which implies a *small-world* network structure. The analysis of the so called *network clustering coefficient*[7] quantifies the clustering properties of a specific community of devices and motivates the introduction of clustering protocols.

This is particularly relevant for the design of a joint clustering and routing scheme, which dynamically organises nodes into groups, referred to as clusters, where each cluster is coordinated by one of the vehicles, referred to as the cluster head (CH) and the rest of vehicles are referred to as the ordinary nodes (ONs) [45]. The optimal CH selection relies on a combination of several parameters, such as the ID, degree, energy level, position, speed and direction, and is typically NP-hard to solve. Several heuristic policies [10, 42] have been proposed in ad hoc networks, such as the lowest-ID, highest-degree and node-weight heuristics. However heuristics are always not optimal.

## 3.2 Game Theory Inspired Approaches

Generally speaking, Game theory (GT) has been widely considered as a promising approach to model the interaction among independent nodes in self-organizing, decentralized and autonomic networks. In particular, non-cooperative models address the interaction among individual rational decision makers, called players. In such models players are assumed selfish, and select their strategies in order to maximize their own performance without caring about the global network performance, leading to steady-state equilibrium that are socially undesirable. As examples of application the case of a medium access control based on the CSMA/CA is considered in [8], while the packet forwarding problem is analyzed in [17, 41]. Likewise, Cooperative GT, focuses on the study of coalition of players, that act like a single entity. In particular, [36] proposes a novel classification of coalitional games, whose main class is that of the canonical games. In this case the superadditivity property holds, meaning that the formation of large coalitions is never detrimental to any of the involved players. The objectives of a canonical game are to study the stability and fairness of the grand coalition, i.e., the coalition of all the players, where stability means finding a payoff allocation which guarantees that

[7] In particular, it represents the probability that two nodes connected to a common node are also connected between themselves.

no group of players have an incentive to leave the grand coalition. In [26] the authors use the canonical games model to tackle the problem of how to fairly allocate the transmission rates between a numbers of users accessing a wireless Gaussian MAC channel, guaranteeing the stability under the envy-free fairness criterion. Finally in [21], canonical coalition games are used to solve an inherent problem in packet forwarding ad hoc networks.

Only recently, GT and coalitional games have been successfully applied also to a few D2D communications, by first analysing open issues [40]. In [27] the problem of uplink radio resource allocation is studied when multiple D2D pairs and cellular users share the available resources. In addition, GT is applied to ensure energy-efficient D2D resource allocation [49], and a simple coalitional game is studied for energy-efficient D2D communications in Long Term Evolution (LTE) public safety networks [46].

The aforementioned limitations of centralized optimization and heuristic solutions, especially for clustering schemes, have led to an interesting body of literature that deals with the use of coalitional graph games, where a specific graph interconnects the players to maximize their individual payoffs [14]. The coalitional graph games principle has been applied to several wireless-oriented applications, such as wireless physical layer security improvement [6], network formation for maximizing data exchange [4], and decentralized joint relay selection and power allocation [43]. Despite their potential, such game-theoretical approaches have their own shortcomings, such as requiring the knowledge of other players' actions, the solution equilibrium is evaluated from each single player's point of view, which doesn't fit the setting of two distinct sets of players (CHs and ONs) in the VANET clustering problem. Recently, matching theory has becoming a promising technique which can overcome some limitations of game theory and optimization [20].

## 4 Novel ITS Applications Based on Context and Social Awareness

Vehicular Social Networks (VSNs) [29] are an emerging type of network which allows the exchange of information among drivers, passengers, as well as vehicles. The exploitation of social relationships can improve and extend the added value provided by applications of vehicular networks (e.g. navigation safety applications, navigation efficiency, entertainment, participatory and urban sensing, emergency [19]). Indeed, leveraging the social dimension in vehicular networking can bring several advantages:

- ease the dissemination of information by exploiting users common interests and preferences thus avoiding to deliver information to uninterested parties;
- ease the on-time and continuous production of up-to-date and capillary information on the road conditions and traffic status by turning users and vehicles into producers of information;

- enhance the exchange of information, media content and recommendations for entertainment and tourism;
- support new mechanisms for reputation building and establishment of trust among nodes in Vehicular Ad-Hoc Networks (VANETs) [24], (for instance the relationship degree in online social networks could contribute to define the degree of trust with an unknown node entering the network).

With respect to on-line and mobile social networks, VSNs pose several challenges:

- the topology of a VSN changes very often and it is formed by peers that move at high speed;
- contact duration of vehicles on the move is short (in the order of tens of seconds [29]);
- the technological infrastructure is made by heterogeneous components, such as On-Board Units (OBU), vehicle sensors and actuators, mobile phones, Road Side Units (RSU) and remote services.

The nature of VSNs is inherently dynamic and dependent on context changes, as discussed hereafter.

## *4.1 Context-Awareness in VSNs*

One of the first and more widely adopted definition of context was provided by Dey [13]: "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves".

Context-awareness was defined as the capability of a system to use context for providing relevant information and/or services to the user, depending on the user's task. In the last decade, a lot of research has been conducted in the field of context-aware systems to propose novel context representation and reasoning models and context management frameworks and also to widen the scope of application of context-aware adaptation mechanisms, beyond information provisioning to end users, to include several application domains, including the automotive one.

Context-awareness plays a relevant role in VSNs for several reasons:

- context information can be used to dynamically detect and build communities. Relations in VSNs can be built by taking into account highly variable information such as common geographical location, final destination and/or part of the route, but also rarely changing information, such as habits (e.g., daily routes to work), friendship, family relationships, common interests and trust;
- members of VSNs can exchange context updates, such as change of speed, occurrence of congestion events or accidents, PoI recommendations;

- provided services can adapt to current context, e.g., navigation services can suggest a new route, alerts are propagated only to nodes that can be affected by the alerted event.

It is expected [19] that the evolution of Vehicular Networks will intersect that of smart phones and Wireless Sensor Networks (WSN), thus extending the range of context sources and the type of context information that can be gathered in the vehicular environment and opening up the opportunity of novel context-aware services. Ad-hoc networking mechanisms provided by wireless and cellular technologies are key enablers for VSNs since they enable the establishment of opportunistic networks relaying on the cooperation of dynamic groups of peers (i.e., smart phones, vehicle OBUs, smart objects along the road infrastructure).

Moreover, these opportunistic networks can be promptly put in place or reconfigured when the situation requires a boost in the degree of social interactions in order to face unexpected events. For instance, as soon as an incident occurs, a network can be setup to alert the vehicles approaching that location.

Context in vehicular/automotive environments can be represented by referring to the following main categories:

- *User*
- *Vehicle*
- *Social Context*
- *Physical context*
- *Application*
- *Network*
- *Device.*

This context information can be used to trigger appropriate adaptation at different scopes (in-vehicle, vehicle-to-vehicle, vehicle-to-remote services) in order to enhance the quality of travel for drivers and passengers and assure their safety. Table 1 provides some main examples of context-aware adaptation actions.

## *4.2 Service Overlay*

The appropriate use of context-aware adaptation paradigms in the operation of application, network and communication services is considered a key enabler not only for VSN but also, on a wider scope, for 5G technology [39]. Indeed, 5G infrastructure is envisioned as a *Neural bearer* where *Everything* (applications, processing resources, network, data, etc.) can be provided as a service and most of intelligence, including context-aware adaptation logic, will be distributed at the edge, i.e., in the aggregation and access segments up to the end user premises [39]. As a consequence, it is possible to envisage an overlay of heterogeneous service capabilities pervasively distributed and interconnected and deeply integrated through the 5G network infrastructure that can be dynamically composed and

**Table 1** Context-aware adaptation in VANETS

| Context | | Context-aware adaptation | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Context Category* | *Description* | Warning (collision, security distance, etc.) | Prevent the propagation of irrelevant information | Suggest new route to avoid delay due to incident/congestions | Location and direction-aware delivery of information | Keep information "alive" in certain locations and over certain intervals of time | Adapt information delivery and input from user to current user activity and level of attention | Efficient message routing | Select of the optimal transmission rate, | Dynamic building g of groups | Adaptive users scheduling and resource allocation |
| User | User's profile and preferences, user habits, user activity, driver behavior. | X | X | X | | | X | | | | |
| Vehicle | Location, direction, speed. | X | | X | X | X | | | | X | |
| Social | People/Vehicles nearby, distance among vehicles, short- and long-term social relationships. | X | X | | | X | | X | | X | X |
| Physical context | Time, weather information, type of road and traffic density, congestion, incident. | X | | X | X | X | | | X | X | X |
| Application | Navigation applications, tourist guides, Driving assistance systems, PoI recommendation | X | X | X | X | X | X | X | | X | |
| Network | Information characterizing VANETs, such as topology, bandwidth, traffic performance and link quality | | | | | | | X | X | | X |
| Device | Information characterizing devices with user interfaces (e.g. smart phones, OBU, etc.), such as device hardware and software configuration, battery power, memory consumption, received signal strength of available access networks, etc. | | | | | | X | X | | | |

adapted to context changes. In this perspective, it is worth mentioning the IEEE NGSON standard [23] since it specifies a service-overlay network over IP-based infrastructure that provides context-aware and dynamically adaptive service control and delivery functions. The NGSON functional architecture specifies a set of functional entities (FEs) for main NGSON capabilities: service composition, service discovery and negotiation, service routing, context information management, content delivery, and service policy decision to negotiate and enforce Quality of Service (QoS) at service and transport levels with underlying networks.

The envisioned 5G-enabled VANET scenario poses new challenges for the physical deployment of the NGSON distributed overlay. Indeed, to accommodate the requirement of pervasive intelligence and adaptation capabilities, we envision a NGSON physical deployment where FEs would be distributed at least across three levels: locally (on the vehicle and in a proximity-based group of vehicles), at the edge (access/aggregation network) and remotely (network operators and service providers data centers). Thanks to opportunistic communication, hw and sw resources hosted by nodes, such as smart-phones and vehicle OBUs, can be exposed as a service and dynamically composed across different nodes for providing added-value services. For instance a video captured by a cam on a vehicle can be processed by a video compressor provided by a smart-phone and then diffused to the other peers of the network.

## *4.3 Application Scenarios and Communications Use Cases*

Cooperation among vehicles can enable a wide range of application services for the benefit of several categories of end users: drivers of private vehicles, drivers of public or collective transport vehicles, transportation authorities, public

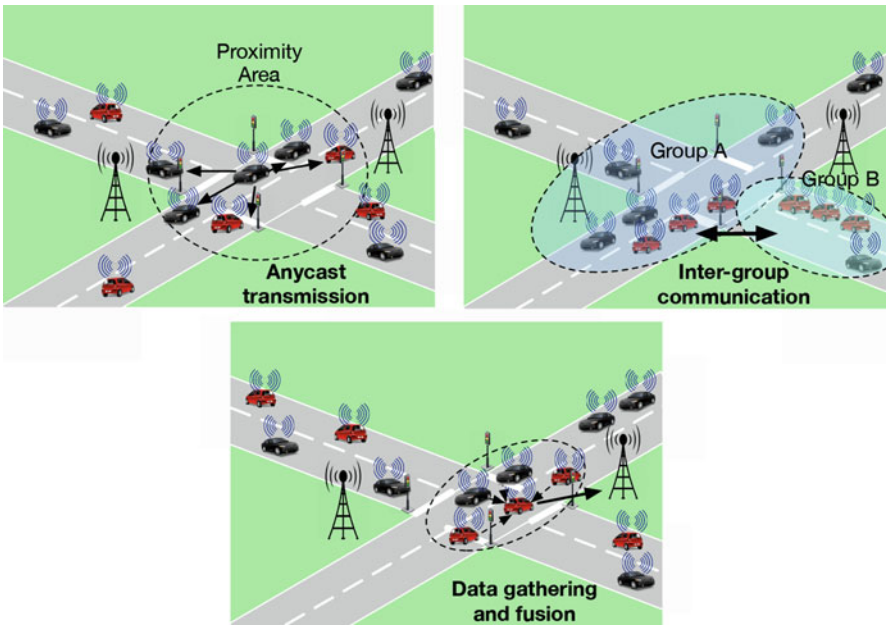


**Fig. 8** Potential use cases from a data exchanging and networking perspective

administrations, etc. Hereafter we provide some scenarios to illustrate how context-aware applications can be empowered by multi-hop D2D communication. To this purpose, we preliminary focus on three different networking use cases, starting from (1) critical information diffusion performed by a single vehicle towards a potential group of neighbours, up to (2) data exchange among convoys when a communication opportunity arises, and (3) data aggregation and fusion performed by a vehicle acting as gateway to the Internet, as depicted in Fig. 8. As soon as communications become less sporadic and context aware, an ad hoc networking scheme is required: in particular we indicated (1) a basic flooding scheme or more advanced solutions relying on (2) disruption-tolerant networking (DTN) or classical mobile ad hoc networking (MANET).

### 4.3.1 Safety and Early Warning Applications

Safety and early warning applications are one of the most investigated types of applications in vehicular systems. Leveraging the dynamic exchange of context information among OBUs and RSUs, dynamic communities can be built that group vehicles that are in the same geographical area and eventually share a significant part of their route or have the final destination in common. For example, events of interest are: an abrupt deceleration of a vehicle or a group of vehicles, an obstacle on the road, a sudden change of weather conditions, a breakdown of a vehicle.

This information can be provided by drivers or passengers (as already enabled by existing applications, e.g. Waze) or detected and automatically *posted* by vehicles so to speed up the time needed for generating and disseminating the alert. For instance a vehicle, elected as coordinator of a group of vehicles, can collect context information provided by group members and execute a task for analysing this information and infer the occurrence of events of interests. For instance, a sudden deceleration and change of direction by a few vehicles may be interpreted as an obstacle on the roadway and dispatched to other interested peers.

### 4.3.2 Autonomic Transport Systems in Smart Cities

The paradigm of *opportunistic* communication is one of the main ICT enablers in the fields of Smart Cities. Indeed, adaptive and resilient communications are expected to play a key role in the achievement of smart city goals (i.e. sustainable economic growth, quality of life, rational use of resources through the active involvement and participation of citizens). Especially V2X communication can enable novel scenarios of cooperation among people, vehicles and urban infrastructure (e.g. traffic lights, street lights, message boards, etc.) so to achieve the best compromise between private interests (e.g. minimize travel time to work) and public goals (e.g. reduce CO2 and noise levels). People, vehicles and smart urban objects can act as sensors and exchange context information. Distributed or centralized decision approaches can be implemented that based on this information can dynamically adapt the use of urban resources (e.g., change traffic light policies, allow or avoid access to a road). For instance, priority at a crossing road could be dynamically decided according to current situation: e.g. number of vehicles on a lane, a congestion after the crossroad, an approaching ambulance, etc. Drivers can be appropriately alerted through road messages and in vehicle alerts. Furthermore, with the introduction of autonomous vehicle control systems, vehicles could cooperatively agree on a traffic flow policy and autonomously enforce the decision (e.g. a group of vehicles approaching a traffic light can smoothly decelerate in order not to stop at the traffic light).

## 5 Summary

In this chapter we have analyzed main current trends and perspective in Vehicular Networking in 5G. We first provided an overview on existing standards and highlighted expected benefits enabled by the interworking with 5G systems. We proposed a 5G vehicular communication systems able to perform the joint management of multiple radio technologies and the resource allocation according to specific application requirements. We pointed out main open issues regarding the protocol design and we discussed the adoption of game theory inspired methodologies in

the VANET clustering problem. Finally, we provided an insight on recent trends in context- and social-aware vehicular applications enabled by the envisaged evolution in Vehicular Networking.

## References


1. A. Aijaz, A. Aghvami, Cognitive machine-to-machine communications for internet-of-things: a protocol stack perspective. IEEE Internet Things J. **2**(2), 103–112 (2015)
2. M. Amoozadeh, A. Raghuramu, C. Chen-Nee, D. Ghosal, H.M. Zhang, J. Rowe, K. Levitt, Security vulnerabilities of connected vehicle streams and their impact on cooperative driving. IEEE Commun. Mag. **53**(6), 126–132 (2015)
3. S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, Y. Koucheryavy, Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands. IEEE J. Sel. Areas Commun. **33**, 67–80 (2015)
4. E. Arcaute, R. Johari, S. Mannor, Network formation: bilateral contracting and myopic dynamics. IEEE Trans. Autom. Control **54**(8), 1765–1778 (2009)
5. D. Astely, E. Dahlman, G. Fodor, S. Parkvall, J. Sachs, LTE release 12 and beyond. IEEE Commun. Mag. **51**(7), 154–160 (2013)
6. S. Bayat, R.H.Y. Louie, Z. Han, B. Vucetic, Y. Li, Physical-layer security in distributed wireless networks using matching theory. IEEE Trans. Inf. Forensics Secur. **8**(5), 717–732 (2013)
7. F. Boccardi, R. Heath, A. Lozano, T. Marzetta, P. Popovski, Five disruptive technology directions for 5g. IEEE Commun. Mag. **52**(2), 74–80 (2014)
8. M. Cagalj, S. Ganeriwal, I. Aad, J.-P. Hubaux, On selfish behavior in CSMA/CA networks, in *IEEE INFOCOM* (2005)
9. L. Chengchao, F.R. Yu, Wireless network virtualization: a survey, some research issues and challenges. IEEE Commun. Surv. Tutorials **17**(1), 358,380 (2015)
10. F. Chiti, R. Fantacci, R. Mastandrea, G. Rigazzi, A.S. Sarmiento, E. López, A distributed clustering scheme with self nomination: proposal and application to critical monitoring. Springer Wirel. Netw. **21**(1), 329–345 (2015)
11. COST IC1004, White Paper on scientific challenges towards 5G mobile communications (2013)
12. J. da Silva, G. Fodor, T. Maciel, Performance analysis of network-assisted two-hop d2d communications, in *IEEE Globecom 2014 Workshops* (2014), pp. 1050–1056
13. A.K. Dey, Understanding and using context. Pers. Ubiquit. Comput. **5**(1), 4–7 (2001)
14. L. Dong, Z. Han, A.P. Petropulu, H.V. Poor, Improving wireless physical layer security via cooperating relays. IEEE Trans. Signal Process. **3**(58), 1875–1888 (2010)
15. B. Dongwoon, P. Cheolhee, L. Jungwon, N. Hoang, J. Singh, A. Gupta, P. Zhouyue, K. Taeyoon, L. Chaiman, K. Min-Goo, K. Inyup, LTE-advanced modem design: challenges and perspectives. IEEE Commun. Mag. **50**(2), 178–186 (2012)
16. F. Dressler, H. Hartenstein, O. Altintas, O. Tonguz, Inter-vehicle communication: quo vadis. IEEE Commun. Mag. **52**(6), 170–177 (2014)
17. M. Felegyhazi, L. Buttyan, J.-P. Hubaux, Ash equilibria of packet forwarding strategies in wireless ad hoc networks. IEEE Trans. Mob. Commun. **5**(5), 463–476 (2006)
18. G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, Z. Turanyi, Design aspects of network assisted device-to-device communications. IEEE Commun. Mag. **50**, 170–177 (2012)
19. M. Gerla, L. Kleinrock, Vehicular networks and the future of the mobile internet. Comput. Netw. **55**(2), 457–469 (2011)
20. Y. Gu, W. Saad, M. Bennis, M. Debbah, Z. Han, Matching theory for future wireless networks: fundamentals and applications. IEEE Commun. Mag. **53**(5), 52–59 (May 2015)

21. Z. Han, V. Poor, Coalitional games with cooperative transmission: a cure for the curse of boundary nodes in selfish packet-forwarding wireless networks. IEEE Trans. Commun. **57**, 203–213 (2009)
22. ICT-317669 METIS Project, Requirements and General Design Principles for New Air Interface. Deliverable D2.1, August 2013, https://www.metis2020.com/documents/deliverables
23. IEEE Standard 1903, The functional architecture of next generation service overlay networks (2011)
24. G. Karagiannis, O. Altintas, E. Ekici, G. Heijenk, B. Jarupan, K. Lin, T. Weil, Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions. IEEE Commun. Surv. Tutorials **13**(4), 584–616 (2011)
25. J.M. Kleinberg, Navigation in a small world. Nature **406**, 845 (2000)
26. R. La, V. Anantharam, A game-theoretic look at the gaussian multi-access channel, in *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 66, pp. 87–106 (2003)
27. Y. Li, D. Jin, J. Yuan, Z. Han, Coalitional games for resource allocation in the device-to-device uplink underlaying cellular networks. IEEE Trans. Wirel. Commun. **13**, 3965–3977 (2014)
28. Y.-D. Lin, Y.-C. Hsu, Multihop cellular: a new architecture for wireless communications, in *INFOCOM 2000. Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3 (IEEE, 2000), pp. 1273–1282
29. F. Mezghani, R. Dhaou, M. Nogueira, A.-L. Beylot, Content dissemination in vehicular social networks: taxonomy and user satisfaction. IEEE Commun. Mag. **52**(12), 34–40 (2014)
30. S. Mohammad, A. Rasheed, A. Qayyum, Vanet architectures and protocol stacks: a survey, in *Communication Technologies for Vehicles*, ed. by T. Strang, A. Festag, A. Vinel, R. Mehmood, C. Rico Garcia, M. Röckl. Lecture Notes in Computer Science, Springer Berlin Heidelberg, vol. 6596 (2011), pp. 95–105
31. R. Monteiro, S. Sargento, W. Viriyasitavat, O.K. Tonguz, Improving VANET protocols via network science, in *IEEE Vehicular Networking Conference (VNC)* (2012)
32. National Highway Traffic Safety Administration (NHTSA), National Motor Vehicle Crash Causation Survey. U.S. Department of Transportation, Report DOT HS 811 059 (2008)
33. H. Nishiyama, M. Ito, N. Kato, Relay-by-smartphone: realizing multihop device-to-device communications. IEEE Commun. Mag. **52**(4), 56–65 (2014)
34. A. Rasheed, H. Zia, F. Hashmi, U. Hadi, W. Naim, S. Ajmal, Fleet and convoy management using VANET. J. Comput. Netw. **1**(1), 1–9 (2013)
35. G. Rigazzi, F. Chiti, R. Fantacci, C. Carlini, Multi-hop d2d networking and resource management scheme for m2m communications over LTE-a systems, in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)* (2014), pp. 973–978
36. W. Saad et al., Coalitional game theory for communication networks. IEEE Signal Process. Mag. **26**(5), 77–97 (2009)
37. J. Seo, T. Kwon, V. Leung, Social groupcasting algorithm for wireless cellular multicast services. IEEE Commun. Lett. **17**, 47–50 (2013)
38. U. Shevade, Y.C. Chen, L. Qiu, Y. Zhang, V. Chandar, M.K. Han, H.H. Song, Y. Seung, Enabling high-bandwidth vehicular content distribution, in *Proceedings of the 6th International Conference* (ACM, 2010), p. 23
39. D. Soldani, A. Manzalini, Horizon 2020 and beyond: On the 5g operating system for a true digital society. IEEE Veh. Technol. Mag. **10**(1), 32–42 (2015)
40. L. Song, D. Niyato, Z. Han, E. Hossain, Game-theoretic resource allocation methods for device-to-device communication. IEEE Wirel. Commun. **21**(3), 136–144 (2014)
41. V. Srinivasan et al., Cooperation in wireless ad hoc networks, in *IEEE INFOCOM* (2003)
42. D. Tian, Y. Wang, H. Xia, F. Cai, Clustering multi-hop information dissemination method in vehicular ad hoc networks. Intell. Transp. Syst. **7**(4), 464–472 (2013)
43. B. Wang, Z. Han, K.J.R. Liu, Distributed relay selection and power control for multiuser cooperative communication networks using stackelberg game. IEEE Trans. Mob. Comput. **7**(8), 975–990 (2009)

44. C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, E. Hepsaydir, Cellular architecture and key technologies for 5g wireless communication networks. IEEE Commun. Mag. **52**(2), 122–130 (2014)
45. T.L. Willke, P. Tientrakool, N.F. Maxemchuk, A survey of inter-vehicle communication protocols and their applications. IEEE Commun. Surv. Tutorials **11**(2), 3–20 (2009)
46. E. Yaacoub, O. Kubbar, Energy-efficient device-to-device communications in LTE public safety networks, in *IEEE Globecom 2012 Workshops* (IEEE, 2012), pp. 391–395
47. S.H. Yu, O. Shih, H.M. Tsai, N. Wisitpongphan, R. Roberts, Smart automotive lighting for vehicle safety IEEE Commun. Mag. **51**(12), 50–59 (2013)
48. K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, Y. Zhou, Heterogeneous Vehicular Networking: A Survey on Architecture, Challenges, and Solutions. IEEE Commun. Surv. Tutorials, **17**(4), 2377–2396, (2015) Fourthquarter
49. Z. Zhou, M. Dong, K. Ota, R. Shi, Z. Liu, T. Sato, Game-theoretic approach to energy-efficient resource allocation in device-to-device underlay communications. IET Commun. **9**(3), 375–385, (2015)

# Next-Generation High-Efficiency WLAN

**Nan Cheng and Xuemin (Sherman) Shen**

**Abstract** Currently, the exponential growth of mobile data traffic has put an increasingly heavy burden on the cellular network, and results in severe overload problem. As a cost-effective Internet access solution, WiFi networks consume a major portion of the global Internet traffic, and greatly offload the cellular network. However, with the increasing demands for WLAN and the deployment of carrier-WiFi networks, the number of WiFi public hotspots worldwide is expected to increase dramatically. To face this huge increase in the number of densely deployed WiFi networks, and the massive amount of data to be supported by these networks in indoor and outdoor environments, it is necessary to improve the current WiFi standard and define specifications for high-efficiency wireless local area networks (HEWs). In this chapter, the emerging HEW technology is introduced and discussed, including typical use cases, environments, and potential techniques that can be applied for HEWs. We first give the typical HEW use cases, and analyze the main requirements from these use cases and environments. Then, potential techniques, including enhanced medium access, and spatial frequency reuse, are presented and discussed.

## 1 Introduction

### *1.1 High-Efficiency WLAN*

As one of the most successful wireless access technologies, IEEE 802.11 wireless local area networks (WLANs) provide cost-effective Internet access that can satisfy the communication requirements of most current wireless services and applications. Due to the high data rate, low deployment complexity, and low price, WLANs have been widely deployed around the world and experienced tremendous growth

N. Cheng (✉) • X. (Sherman) Shen
Department of Electrical and Computer Engineering, University of Waterloo,
Waterloo, ON, Canada N2L 3G1
e-mail: wmchengnan@gmail.com



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_24

in the recent years. Like many other wireless technologies, since the first standard IEEE 802.11a-1999, WLANs or WiFi technologies have evolved and integrated the latest technologies, in order to continue to improve the spectrum utilization and network performance. In IEEE 802.11n-2009, single-user multiple-input multiple-output antennas (SU-MIMO) is added to support a maximum data rate from 54 to 600 Mbps. In addition, frame aggregation is employed to reduce the protocol overhead. IEEE 802.11ac-2013 further extends the features, including multi-user MIMO (MU-MIMO), extending the channel bandwidth to 160 MHz, and high-density modulation (up to 256-QAM). With these feature, IEEE-802.11ac based WLANs can support maximum 866 Mbps data rate. IEEE 802.11af and 802.11ah extend the WLAN spectrum bands to the TV white spaces and sub 1G bands with cognitive radio technologies, so as to support long-range and power saving applications such as machine-to-machine (M2M) communications. In IEEE 802.11p is designed on 5.9 GHz to support Intelligent Transportation Systems (ITS) applications, such as toll collection, vehicle safety applications, and high-speed Internet access. The most popular IEEE 802.11 amendments are shown in Table 1.

The emerging 5G network is envisioned as a heterogeneous network, with macrocells, microcells, small cells and relays working cooperatively [36], as shown in Fig. 1. Different networks may have different features and satisfy different demands. For example, the macrocells provide the largest coverage, and offer high-rate data service to time-intensive applications, however with a relatively high price. On the other hand, small cells can provide local high-speed Internet access with low price, and can greatly enhance the spectrum efficiency. As shown in Fig. 1, in 5G heterogeneous networks, WiFi networks play an important role in delivering local mobile data in a cost-effective way. WiFi is recognized as one of the primary offloading technologies [1]. By delivering data originally targeted for cellular networks by WiFi, which is referred to as WiFi offloading, the congestion of cellular networks can be alleviated. WiFi offloading has been extensively studied for stationary or slow moving users [1, 6, 28, 35]. It is shown that around 65 % of the cellular traffic can be offloaded by merely using the most straightforward way of simply switching the IP connection from the cellular network to WiFi when the WiFi connectivity is available. In addition, significant amount (above 80 %) of data can be offloaded by delaying the data application [28].

In spite of the high data rates supported, the next-generation WLANs face three challenges. First, the popularity of WLANs has led and will continue to lead to very high WLAN deployment, which is a primary issue to address. Second, the emerging data-craving applications, such as real-time high-quality video streaming will significantly increase the users' throughput requirements. Third, the outdoor WLANs are getting more popular, which has a more complex communication environment. For the first challenge, legacy WLANs will find its difficulty to well behave in the scenario of very dense deployment of access points (APs). Based on the carrier sense multiple access with collision avoidance (CSMA/CA) mechanism WLANs employ, in the dense scenario, the contention problem will be severe and the channel utilization is poor. In addition, in the dense scenario, WLANs are not noise-limited system any more, and the signal from nearby co-channel basic

**Table 1** Popular IEEE 802.11 amendments

| Amendment | Frequency (GHz) | Bandwidth (MHz) | Maximum data rate (Mbps) | Responsibility |
|---|---|---|---|---|
| a | 5 | 20 | 54 | Enable up to 54 Mbps data transmission in 5 GHz unlicensed band by utilizing orthogonal frequency division multiplexing (OFDM) |
| b | 2.4 | 22 | 11 | Enable up to 11 Mbps data transmission in 2.4 GHz unlicensed band by utilizing Direct-sequence spread spectrum (DSSS) |
| g | 2.4 | 20 | 54 | Enable up to 54 Mbps data transmission in 2.4 GHz unlicensed band by utilizing OFDM |
| n | 2.4/5 | 20/40 | 150 | Utilize MIMO to support higher data rates |
| ac | 2.4/5 | 20/40/80/160 | 866.7 | Support MU-MIMO to further enhance data rate |

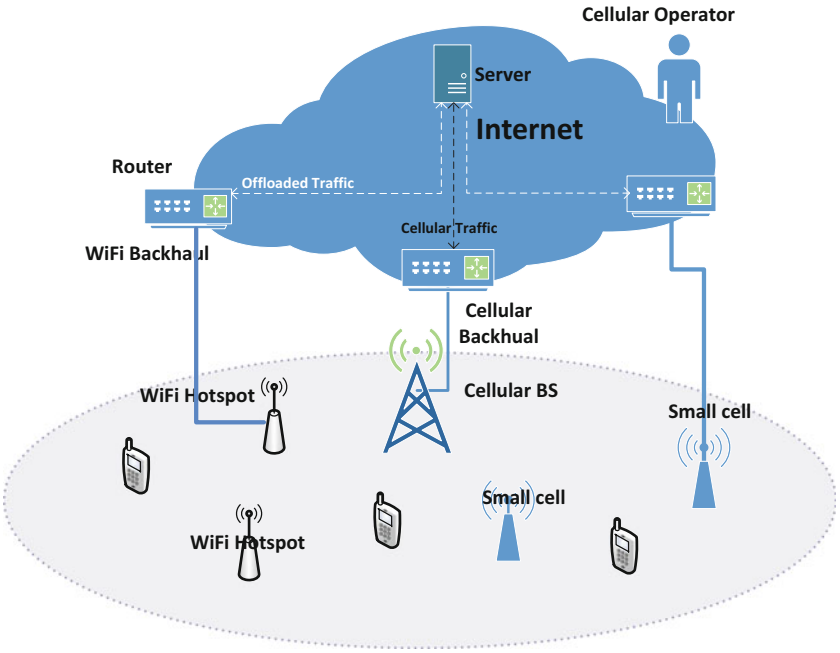


**Fig. 1** Heterogeneous 5G networks with small cells and WLANs

service sets (BSSs) may significantly interfere with the received signal, leading to low signal-to-interference-plus-noise (SINR), and thus degrade the performance. The overlapping BSS (OBSS) problem should be well addressed in next-generation WLAN. For the second challenge, since the current WiFi amendments are aimed at improving the network throughput performance, advanced mechanisms should be investigated to improve the single user's throughput. For the third challenge, the channel condition in outdoor environments is more unreliable, since the delay spread can be large due to the multi-path transmission. In addition, the non-LOS channel makes it difficult to maintain links, even with a good SINR. Due to the possible long transmission range, the power of user devices can be a limiting factor. Also, the interference can be severe, caused by unmanaged networks, and home gateways signal leaking outdoors.

To address these challenges, the High-Efficiency WLAN (HEW) Study Group [13] is working on a new 802.11 amendment called IEEE 802.11ax-2019. A typical dense deployment of HEW is depicted in Fig. 2, where there is a large number of access points (APs) and a large number of stations (STAs) associated with each AP. The scope of 802.11ax is stated in [15]:

> This amendment defines standardized modifications to both the IEEE 802.11 physical layers (PHY) and the IEEE 802.11 Medium Access Control layer (MAC) that enable at least one mode of operation capable of supporting at least four times improvement in the average throughput per station (measured at the MAC data service access point) in a dense

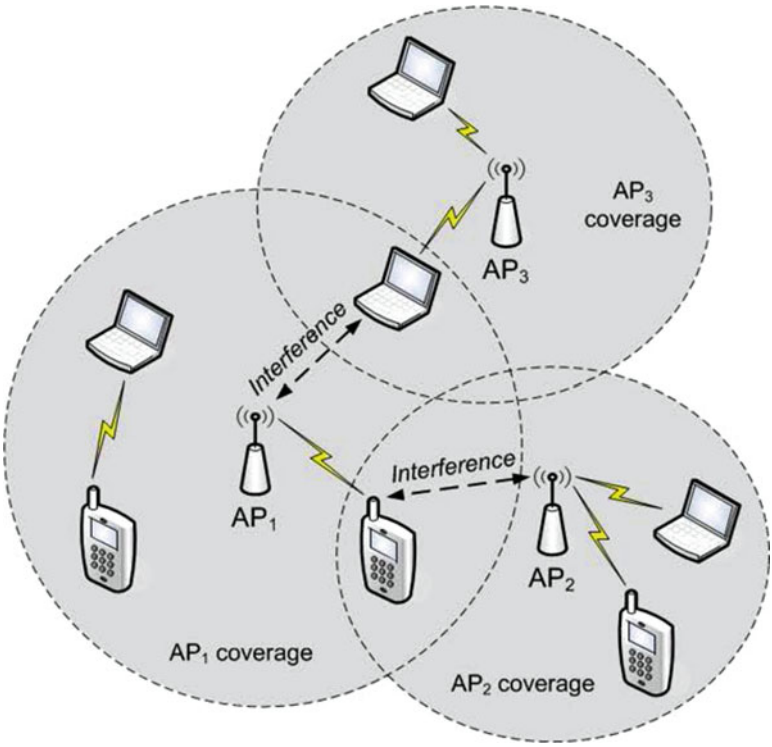


**Fig. 2** A dense HEW deployment

deployment scenario, while maintaining or improving the power efficiency per station. This amendment defines operations in frequency bands between 1 and 6 GHz. The new amendment shall enable backward compatibility and coexistence with legacy IEEE 802.11 devices operating in the same band.

In this chapter, we mainly present the features of the IEEE 802.11ax-2019 HEW, and discuss the potential techniques.

## *1.2 Requirements and Performance Metrics of HEW*

From the above statement, it can be seen that the new IEEE 802.11ax HEW, as one of the next-generation WiFi technologies, has the following requirements:

- *Throughput enhancement per station*. Due to the varied throughput requirements of individual users and emerging applications, the throughput performance should be enhanced from the perspective of not only the network, but also individual users. Average throughput per station is directly proportional to both

area throughput and aggregate BSS throughput. Therefore, the 5th percentile measure of the per station throughput can be used to evaluate the desired distribution of throughput among a number of stations in an area. These metrics, along with the satisfaction of the packet delay and the packet error ratio (PER) requirements of applications, will directly correspond to user experience in identified scenarios. Moreover, based on [15], the increase in average throughput per station is not limited to four times improvement, while a range of 5–10 times are targeted, depending on technology and scenario.

- *Efficient in dense scenario*. A dense HEW deployment will result in a high level of co-channel interference (CCI), due to adjacent BSS APs operating on the same frequency channel. Hence, it is necessary to mitigate the CCI among neighbouring BSSs, in order to maintain a stable HEW throughput and satisfy the quality of service (QoS) requirements of the supported applications. HEW will be evaluated with a set of typical deployment scenarios representative of the main expected usage models that are likely to suffer bottlenecks in the coming years: residential, enterprise, indoor and outdoor hotspots. In these scenarios, three categories of objectives are highlighted. First, Spectrum resources should be utilized in a more efficient way in scenarios with a high density of STAs per BSS. Second, spectral frequency reuse should be significantly enhanced, and interference between neighboring OBSS should be managed. Third, the robustness in outdoor propagation environments and uplink transmissions should be increased.
- *Power efficiency improvement*. Battery is one bottleneck of the mobile devices. Applications and WLAN infrastructure assume that the devices are constantly connected, which results in a large battery drain. In addition, current WLAN devices spend 90 % of time listening, which also consumes much energy. Several works in the literature have proposed mechanisms to enhance the STA power efficiency by using sleeping mechanisms, however, none has considered a dense deployment of WLANs in an outdoor environment. Power efficiency is intended to measure consumption of devices which can reasonably be assumed to be powered by batteries and will take into account average power consumption for a given scenario.
- *Backward compatibility*. The HEW should be able to provide backward compatibility, such that legacy devices can work in HEW, and in turn HEW devices can work well in legacy networks.

To evaluate the performance of HEW, performance metrics should be properly defined. In [15], it is stated that HEW aims to improve the metrics that reflect user experience, including per station throughput, the 5th percentile of per station throughput, and area throughput, in the scenarios of indoor/outdoor dense deployment. The detailed performance metrics of HEW are suggested and discussed in [12]. Among them, several important metrics are listed as follows.

- *Self-interference suppression*. This is a measurement of the ability of a device to receive signal when it is transmitting signal on the same frequency channel. Self-interference suppression is critical for full-duplex transmission, which is one of the key technologies of HEW.
- *MAC efficiency*. Defined by MAC throughput divided by PHY throughput, MAC efficiency is a proper metric to evaluate the performance of HEW MAC protocols.
- *Spectrum Efficiency*. Spectrum Efficiency is measured in bps/Hz, which should be maximized to enhance the network performance.
- *Throughput*. The 5 percentile and 95 percentile user throughput should be measured to evaluate the throughput at the center and edge of the cell. In dense deployment scenario, improving the throughput performance is important yet challenging, due to the presence of severe collision and interference.
- *Robustness*. Robustness evaluates the reliability and strategy of wireless connections in different scenarios. Association strategies include always connected, connected with minimum rate, and blocked when the admitted data rate is not met
- *Availability (or outage)*. Given the link received signal is good, the communication link can still be unavailable due to high congestion, overloading, or severe interference.
- *User experience*. User experience may include network selection and association latency, association successful ratio, data transmission latency, and handoff latency and successful ratio.

In the following of this chapter, we first study the HEW use cases, and then discuss several key enabling technologies, including enhanced 802.11 distributed coordination function (DCF) and spectrum spatial reuse. At last, the chapter is concluded.

## 2 HEW Use Cases

In this section, we list and discuss the usage models of IEEE 802.11ax HEW. Different from legacy WLAN, most of the use cases for HEW are characterized by a high density of STAs and/or BSSs. Users need WiFi access anytime and everywhere, which pushes its deployment/usage characterized by a high density of STAs and BSSs, such as hotspots in airport/train stations, malls, stadium, parks, streets, and campus. In addition, the increased usage of WiFi also leads to a higher density in "traditional" WiFi environment. Therefore, clearly defining the usage models can understand the specific requirements of each individual usage model, which can in turn facilitate the investigation of HEW key technologies.

## 2.1 Environment

Environment is the type of place in which the network of the use case is deployed. In HEW, seven environments are defined, which are enterprise, small office, hotspot in public places, outdoor hotspots, home, campus, and airplane/bus/train/ship, respectively.

- *Enterprise*. The enterprise environment can be on desk and in cubicle characterized by short range and line of sight (LOS) transmission, and conference room characterized by medium range and mostly line of sight transmission.
- *Small office*. The small office environment is characterized by single BSS with limited number of users and unmanageable interferences.
- *Hotspot in public places*. WLAN will be widely deployed in public areas, such as transportation hubs (airport, train station, and bus station), exhibition hall, and shopping malls. This environment is characterized by many STAs in a BSS.
- *Outdoor hotspots*. The typical examples include park, streets, stadium, and areas with special crowded events. The outdoor WLANs can be co-located with cellular base stations such as small cells, and user equipments such as private mobile APs.
- *Home*. The typical examples include dense apartment buildings and town houses, where WLANs can interfere each other.
- *Campus (Educational space, Hospital)*. The typical examples include auditorium/lecture halls/classrooms in the educational space, video conferencing/telepresence, and hospitals where remote medical assistance for operations is via wireless networks.
- *Airplane/bus/train/ship* WLANs can provide Internet access in large vehicles, such as in airplane and train. This environment is characterized by many STAs in a BSS.

Most environments are characterized by the overlap of multiple WLANs in the dense area that need to cohabit efficiently and fairly. There are different types of network overlap: (1) one or multiple cluster of APs (extended service set (ESS)), where each ESS is managed by a controller; (2) one or multiple stand-alone APs, each with its own management entity; (3) one or multiple single-link networks for P2P communications; and (4) the overlapping networks can be legacy networks (11a/b/g/n/ac).

## 2.2 Applications

In HEW, application is defined as a source and/or sink of wireless data that relates to a particular type of user activity. Several new/enhanced applications are considered, which include wireless docking, unified communications, display sharing, cloud computing, video distribution at home, progressive streaming, user

generated content (UGC) upload and sharing, interactive multimedia and gaming, real-time video analytics and augmented reality, wearable devices, geo-location services, etc.

## *2.3 Usage Models*

In the current discussion of HEW, five main usage models are defined, each of which include several different environments. A list of usage models, along with the environments, are given in Table 2, as listed in [16].

Comparing to the use cases defined in previous IEEE 802.11 amendments, such as 802.11ac, there are substantial differences with respect to the environments, applications, objectives, and metrics. In terms of environments, hotspots for cellular offload and public access in dense areas for indoor and outdoor are unique, and for general home and enterprise scenarios, it is more challenging towards high density deployment. For applications, HEW usage models consider a mix of traffic and clients, and focus on the quality of experience (QoE) of all clients in these more dense scenarios. For objectives, unlike previous IEEE 802.11 amendments which aim to improve the peak throughput, HEW targets to improve the WLAN efficiency by improving user experience and overall capacity per area. For metrics, the candidate metric that are under discussion include average per user throughput, 5th percentile of per user throughput cumulative density function (CDF), and area throughput. In the following, each use case is introduced.

### 2.3.1 Use Case 1

Use case 1 focuses on the scenarios with both high density of APs and high density of STAs per AP. This is the most important and challenging use case of HEW, where the high density of APs can cause severe interference, and the high density of STAs may congest the network, leading to packet collisions and retransmission. The typical environments in use case 1 are shown in Fig. 3.

(1a) Stadium. The stadium environment is an open area with few obstacles and multiple APs deployed by single/multiple operators. Most of the transmissions are LOS and the layout of APs are frequently changed. Interference can happen between APs belonging to the same managed ESS, and between unmanaged APs due to very high density deployment. In addition, interference may also caused by cellular networks and bluetooth devices. The traffic is characterized by bursty in time and is uneven due to users' participation in physical space. Example applications may include many users requesting high-rate best effort recreational content, receiving VHD video feed highly compressed, and following ESPN event or Twitter for supplemental event content.

**Table 2** Usage models classified by requirement characteristics

| Use case no. | Use case | Env. no. | Environment |
|---|---|---|---|
| 1 | High density of APs and high density of STAs per AP | a | Stadium |
| | | b | Airport/train stations |
| | | c | Exhibition hall |
| | | d | Shopping malls |
| | | e | E-Education |
| | | f | Multi-media Mesh backhaul |
| 2 | High density of STAs—indoor | a | Dense wireless office |
| | | b | Public transportation |
| | | c | Lecture hall |
| | | d | Manufacturing floor automation |
| 3 | High density of APs (low/medium density of STAs per AP)—indoor | a | Dense apartment building |
| | | b | Community WiFi |
| 4 | High density of APs and high density of STAs per AP—outdoor | a | Super dense urban street |
| | | b | Pico-cell street deployment |
| | | c | Macro-cell street deployment |
| 5 | Throughput-demanding applications | a | Surgery/health care |
| | | b | Production in stadium |
| | | c | Smart car |


**Fig. 3** Typical environments of use case 1. (**a**) Stadium. (**b**) Airport. (**c**) Exhibition hall. (**d**) Shopping mall

(1b) Airports and train stations. In airports and train stations, travelers are using the networks deployed by many service providers to check emails, watch movies, play online games, and access cloud services. The environment can be very complex and WLAN data transmissions can suffer severe interference, which may come from APs belonging to the same managed ESS, APs belonging to different managed ESS, and unmanaged networks.

(1c) Exhibition hall. Exhibition halls are an typical example of high density use cases. There may be many WLANs deployed by the exhibitors and event coordinator, through which people access the digital contents like demonstration and promotion videos of exhibitors. People also require Internet access to visit exhibitors' web sites, to send video and pictures to the companies, etc. In the indoor open spaces, there will be many APs and STAs and may suffer severe interference from APs belonging to the same managed ESS, stand-alone APs in the exhibition booths, and unmanaged networks.

(1d) Shopping mall. In shopping malls, shoppers can use the WLANs to localize themselves in the mall and receive coupons (potentially along with augmented

reality). For classical public access, users can also use the network to visit websites, watch videos from YouTube, play online games, and access cloud services.

### 2.3.2 Use Case 2

Use case 2 is characterized by high density of STAs in indoor environments. In this category, the density of APs may not be extremely high, but the density of STAs per AP is very high, resulting in the congested networks, and degraded performance. The typical environments in use case 2 are shown in Fig. 4.

(2a) Dense wireless office. In a dense office, devices are operating in close proximity, with the distance between two devices smaller than 50 meters. There are multiple APs per floor, and 20–30 STAs per AP. The computers of an office can wirelessly associate to an external server and display with user configuration. This requires a reliable WLAN transmission so that users do not feel a delay indicating the wireless communication. In addition, in


**Fig. 4** Typical environments of use case 2. (**a**) Dense wireless office. (**b**) Public transportation. (**c**) Lecture hall. (**d**) Manufacturing floor automation

office users may have their own fixed or mobile devices, which also requires Internet access with high data rates. The interference in this environment can be caused by peer-to-peer networks, between APs belonging to the same managed ESS, and between APs belonging to different managed ESS due to the presence of multiple enterprise in the same building.

(2b) Public transportation. In the public transportation vehicles, such as trains, buses, and planes, crowded users may have operational WLAN network for Internet access, especially during the peak hours, such as commuting time. The WLAN can provide contents such as video, audio, and interactive gaming through onboard entertainment system, and can also offer high-rate Internet access for in-vehicle users. Interference in this scenario can be caused by unmanaged networks, as well as by surrounding WLANs for public transportation systems.

(2c) Campus network—lecture hall. In a lecture hall, a professor can remotely deliver a lecture to hundreds of students. The lecture comprises of both real-time video of the professor and supplemental content. The professor can receive real-time video of the lecture hall, and can zoom the video to view the audience, while the students can use wireless devices to watch the video and access the supplemental content simultaneously.

(2d) Manufacturing floor automation. On a manufacturing floor, there are thousands of tasks happening each minute, and many of the tasks requirement communications. In such as scenario, a mix of applications may have very different QoS requirements. For example, streaming of live or CAD video requires high throughput, very low delay, and high reliability. Machine-machine communications and robotic material handling requires high reliability but less time sensitivity and lower data rate.

### *2.4 Key Technologies*

HEW considers MAC and PHY technologies that can significantly improve WLAN efficiency and robustness. Specifically, it should make more efficient use of spectrum resources in dense deployment scenarios, increase robustness in outdoor propagation environments, and improve the power efficiency. To achieve these goals, many new and efficient technologies should be employed to work together. We list the potential technologies in Table 3, and discuss two of them, i.e., enhanced DCF and spatial spectrum reuse, in detail in the rest of the chapter.

## 3 Enhanced 802.11 DCF

The distributed coordination function (DCF) is the fundamental access method of CSMA/CA technique, which has not been enhanced since the release of the legacy IEEE 802.11 standard. However, CSMA/CA techniques have poor performance

**Table 3** Potential technologies for HEW

| Category | Technology |
|---|---|
| Edge throughput enhancement | HARQ |
| | Larger contention period |
| MAC enhancement | Basic access mechanism enhancements |
| | Dynamic sensitivity control |
| | Traffic prioritization, QoE |
| | Multicast transmissions |
| MIMO/Beamforming | Massive MIMO, MIMO precoding |
| | DL/UL MU-MIMO |
| | Beamforming for OBSS |
| | Beamforming for interference handling |
| Multiplexing schemes | OFDMA, SDMA, OFDM-IDMA, FFR |
| | TD-uCSMA |
| | Channel bonding |
| Overlapping BSS handling | Interference management, antenna pattern nulling |
| | Efficient resource utilization |
| | Control frame transmission reduction |
| Full-duplex (FD) | MAC/PHY mechanisms for enabling In-band FD |
| | Enhancements for enabling out-band FD |

in very dense scenario due to collision and interference. IEEE 802.11ax HEW aims to enhance or change the underlying CSMA/CA scheme to minimize the collisions. There are two ways to do this: changing to a centralized MAC scheme, or enhancing the CSMA/CA based protocol. Centralized solutions, such as the hybrid coordination function controlled channel access (HCCA), have never been employed in WLAN. Therefore, to maintain backward compatibility, HEW sticks to the enhancement of CSMA/CA protocols. Potential development of the DCF operation is required in order to improve its performance in a highly dense scenario, i.e., when a BSS consists of a large number of STAs. To achieve this goal, two techniques can be employed. First, **full duplex technology** allows the simultaneous signal transmission and reception, which can increasing the number of simultaneous transmissions within a BSS and reduce transmission collision probability. Second, the **enhanced access category (AC)** can decrease the channel time for control information transmissions in order to improve the channel utilization.

## *3.1 Full Duplex Technology*

In IEEE 802.11 DCF, it is assumed that each STA is equipped with a half-duplex physical layer, i.e., either transmits or receives signal at one time. However, recently, the development of a full duplex technology has significantly advanced, which

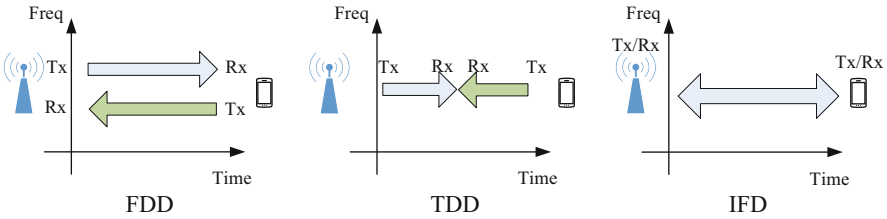


**Fig. 5** In-band full-duplex

allows a transceiver to simultaneously transmit and receive signals on the same frequency [2, 5, 24]. The concept of full duplex technique is shown in Fig. 5. With frequency division duplexing (FDD) or out-band full-duplxe (OFD), a device can transmit and receive at the same time, however, on different frequency bands, which wastes frequency resource. With time division duplexing (TDD) or in-band half-duplex (IHD), a device can transmit and receive on the same frequency band, but at different times, which wastes time resource. Differently, in-band full-duplex (IFD) radio can transmit and receive simultaneously on the same frequency channel, and thus does not waste frequency and time resources. The merit of IFD is that it can achieve high spectral efficiency. Given a link between two nodes a and b, we can calculate the theoretical ergodic capacities of ideal IFD as

$$C_{IFD} = C_{ab} + C_{ba}, \tag{1}$$

where $C_{xy} = \log_2(1 + SINR_{xy})$, while the theoretical ergodic capacities of OFD and IHD can be calculated by

$$C_{IFD} = \frac{1}{2}C_{ab} + \frac{1}{2}C_{ba}. \tag{2}$$

Therefore, theoretically, IFD can achieve twice spectra efficiency than OFD and IHD. However, there are two main challenges of implementing, self-interference (SI) and CCI. Figure 6 shows how SI is generated and influence the IFD operation. Since the average transmit power of WLAN is 20 dBm, and the noise floor is around −90 dBm, the SI should be canceled by 110 dB to a negligible level [2]. CCI may happen when an IFD capable AP supports the IHD capable users, where user A can transmit to the AP while AP is transmitting to user B. Thus, the transmission of user A causes interference at user B. Self-interference cancellation technologies are proposed to reduce the SI, and achieve more IFD gain. There are mainly three categories of SIC technologies, which are propagation SIC (PSIC) in antenna domain, analog SIC (ASIC) in circuit domain, and digital SIC (DSIC) in baseband domain. However, for PSIC, the merit of IFD in terms of spectral efficiency disappears since PSIC mainly employs different antennas for Tx and Rx, and cancel SI by physical isolation between Tx and Rx antennas. In addition, the

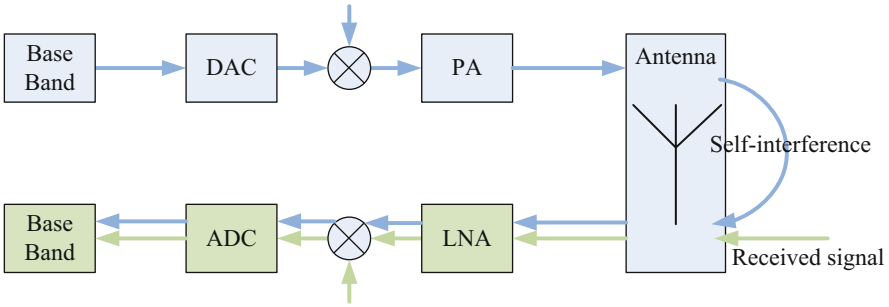


**Fig. 6** Self-interference

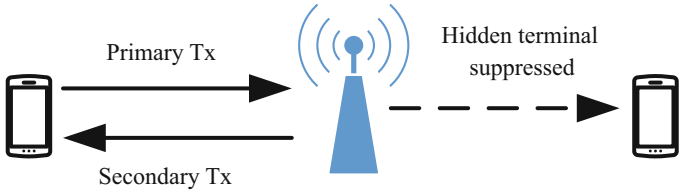


**Fig. 7** IFD solves hidden terminal problem

size of IFD transceiver gets larger. Therefore, SIC mainly rely on ASIC and DSIC, which operates with single antenna. With ASIC, The comparison of state-of-the-art SIC performance is shown in Table 4, which is summarized in [17].

Besides up to 2 times spectral efficiency, IFD can also facilitate advanced MAC protocols to solve many problems using half-duplex transceiver. It can reduce the DCF overhead by eliminating request-to-send (RTS)/clear-to-send (CTS) handshaking information exchange which precedes the actual data transmission [11]. This is because the receiver can immediately start to transmit, which suppresses nearby nodes, as shown in Fig. 7. In addition, when a WLAN is saturated, the bottleneck of the network is the AP, since it has the same chance to transmit as each of its serving clients. With IFD, the problem can be solved since the AP can transmit when STAs are sending packets to it. To date, a few studies have considered the development of MAC schemes based on a full duplex physical transceiver. Some of these studies propose techniques that are not based on the CSMA/CA mechanism [27]. In [27], AP behaves as a central node, scheduling the transmissions of all STAs under control. The AP first sends a probe packet, and all STAs registered under the AP will send the information of transmission length and current interference level to the AP. Based on the information, a schedule called Janus determines the packets to be transmitted concurrently, and the mode (half- or full-duplex) and data rate using a heuristic algorithm to achieve close to the best throughput performance. Simulation results show that Janus MAC can achieve nearly twice throughput comparing to the legacy MAC schemes. However, it may not be practical since centralized MAC is not favored by WLAN implementations.

**Table 4** Comparison of state-of-the-art SIC performance

| Institute | Year (Standard) | Freq. (MHz) | BW (MHz) | PSIC (dB) | ASIC (dB) | DSIC (dB) | Total SIC (dB) |
|---|---|---|---|---|---|---|---|
| NEC (Japan) | 2011 | 5 | 10 | 55 | None | 20 | 75 |
| Rice University (US) | 2011 | 2.4 | 10 | 57 | 24 | None | 81 |
| | 2012 | 2.4 | 20 | 65 | 20 | 85 | |
| | 2012 | 2.4 | 20 | 71 | 24 | 95 | |
| Stanford University (US) | 2010 | 2.48 | 5 | 30 | 25 | 15 | 70 |
| | 2011 | 2.4 | 10 | | 45 | 28 | 73 |
| | 2013 | 2.45 | 80 | None | 60 | 50 | 110 |
| DUPLO | 2014 | 2.45 | 6 | None | 50 | 50 | 110 |
| RF Window (Korea) | LTE WCDMA | 2 | 20 | 60 | None | 10 | 70 |
| WITHUS (Korea) | LTE WCDMA | 2 | 10 | 35 | None | 35 | 70 |
| AirPoint (Korea) | LTE TDD | 2.2 | | 55 | None | 35 | 90 |
| SOLiD (Korea) | | 2 | | 65 | None | 35 | 100 |

On the other hand, other research has developed full-duplex MAC schemes based on CSMA/CA [24, 33]. There is one case the hidden terminal cannot be addressed by IFD, where the primary and secondary packets may have different lengths. For example, in Fig. 7, if the secondary transmission is shorter than the primary transmission, the hidden terminal is no more suppressed. In [24], a busy tone mechanism is employed to address this problem. When a nodes finish transmitting before finish receiving, it will transmit a predefined busy tone, until it finishes receiving; when a node receives a packet but has no packet to transmit, it will also transmit the busy tone. Through the hardware implementation of a simple case where two full duplex nodes transmit packet to each other, it is shown that the proposed full-duplex MAC scheme can achieve better performance in packet reception ratio, throughput and fairness than half-duplex MAC schemes. In [33], a full-duplex physical layer (FD-PHY) is proposed and implemented, based on which a full-duplex MAC (FD-MAC) scheme is developed and implemented for infrastructure-based WLANs. To maximize full-duplex capability and guarantee fairness, queue information needs to be shared with neighboring nodes, and all nodes should be allowed to access the medium. To do this, three mechanisms are employed in FD-MAC. First mechanism is called shared random backoff, the backoff counters of two communicating nodes are synchronized so that they can transmit at the same time while other nodes are allowed to access channel during the backoff of the communicating nodes. The second mechanism is snooping on the header of all transmission within the radio range, even when the node is not transmitting during the network allocation vector (NAV). This enables the nodes to estimate their local topology, and discover the clique or hidden node with themselves. Lastly, the FD-MAC employs virtual contention resolution mechanisms to further balance the objectives between maximize full-duplex opportunities and fairness.

Current research on IFD-based MAC is still in its infant stage. The target MAC scheme should be able to maximize the benefit of full-duplex capabilities (i.e., the opportunity of simultaneous transmission and receiving by a STA) in infrastructure and independent BSSs, support half duplex operation (for backward compatibility), and maintain fairness among all half duplex and full duplex STAs. In addition, the employment of such technology will also facilitate developing more efficient transmit power control and interference mitigation schemes.

## *3.2 Enhanced Access Category*

The aim of HEWs is to satisfy the QoS requirements of various applications, in both dense indoor and dense outdoor environments, while reducing the power consumption. In HEWs, due to the DCF scheme, a STA's average throughput decreases with the increasing number of STAs that are being served by the AP, and the worse channel condition: the former is due to increase of transmission collisions, and the latter is due to that the STA with weak channel take a long time

for its data transmission [3, 29]. To satisfy STAs' QoS requirements, the pervious WLAN standards provide a basic AC scheme and an enhanced DCF scheme called Enhanced distributed channel access (EDCA) [9]. With the AC scheme, a STA is accepted or rejected based on medium time which is calculated from the minimum physical layer rate of the STA [8, 11, 26]. However, Such AC scheme is inefficient because calculating the medium time results in an underutilized channel. In addition, AC-based EDCA cannot guarantee the strict QoS requirements of real-time voice and video services [4]. Therefore, other AC schemes (e.g., [8, 11, 26, 30]) and further improvements for EDCA (e.g., [4]) have been proposed. However, none of these schemes has been designed for a dense WLAN deployment in an outdoor environment. In such deployment scenario, multiple APs coexist, wireless channels rapidly varys, and all the WLAN frequency channels are utilized by adjacent APs. Therefore, to achieve robust and optimal throughput and QoS performance, advanced AC and dynamic handover schemes based on these characteristics should be developed.

## 4 Spatial Frequency Reuse

OBSS is one of the main problems that HEW aims to solve, since high density deployment of WiFi networks is one targeting scenario of HEW. Intuitively, improving spatial reuse can efficiently mitigate OBSS. Spatial reuse has been extensively studied in the literature. In this section, we present a literature survey of two kinds of spatial reuse improvement methods, namely enhanced clear channel assessment (CCA) and transmit power control (TPC).

### *4.1 Enhanced Clear Channel Assessment*

CCA is a function for WiFi to determine the current usage state of the wireless medium, which is defined in IEEE 802.11-2007 standards [10]. There are two functions in CCA, namely carrier sense (CCA-CS) and energy detection (CCA-ED). The former is used for a receiver to detect and decode WiFi preamble, while the latter is used to detect non-WiFi energy. CCA functions in the way that if a WiFi signal is detected above the CCA-CS level (or CCA-CS threshold), or a non-WiFi signal is detected above the CCA-ED level (or CCA-ED threshold) at a STA, it indicates that the wireless medium is currently captured by other STAs or non-WiFi devices, and thus the STA has to stay off the air in order to avoid collisions. In [10], the CCA-CS and CCA-ED levels are defined, which are summarized in Table 5.

In IEEE 802.11-2007 standards, the CCA levels are defined relatively low values, which can guarantee a large coverage of the WLAN, minimize the interference (signals received below the CCA level is considered as noise), and mitigate the hidden terminal problem. This can work well in low density WiFi with proper AP

**Table 5** CCA levels defined in [10]

| Bandwidth (MHz) | CCA-CS level (dBm) | CCA-ED level (dBm) |
|---|---|---|
| 5 | −88 | −78 |
| 10 | −85 | −75 |
| 20 | −82 | −72 |

channel selection algorithms. However, in HEW, BSSs are deployed densely, and OBSSs are very common. Therefore, in HEW, with a low CCA level, an ongoing transmission can prevent many STAs of nearby BSSs from transmitting, which in turn degrades the overall performance of HEW. In order to improve the performance of HEW (mean and 5 % throughput), an intuitive solution is to allow concurrent transmissions with high data rates. To this end, enhanced CCA mechanisms are required to improve the spatial reuse of HEW.

### 4.1.1 Fixed CCA Enhancements

A straightforward solution is to increase the CCA threshold (CCAT) to a certain value such that HEW STAs can transmit in spite of the presence of signals from other OBSS STAs. The advantages of this kind of CCA schemes are easy-implementing, and in some cases, they can achieve high efficiency. In [14], the throughput performance of several CCATs in the scenario of indoor small BSSs are evaluated through simulations. It is shown that the average throughput increases with the CCAT, while the 5 % point in the throughput decreases with the CCAT. This is because with a higher CCAT, more concurrent transmissions in the same area are allowed, but the cell edge users suffer low SINR due to low signal from the AP and high interference from OBSS STAs. Theoretically, with a higher CCAT, the interference will become severer, and more collisions may happen since more concurrent transmissions are permitted. Therefore, the joint impact of concurrent transmissions, interference, and collision on the HEW performance should be investigated. In [25], the performance of varied CCATs is studied through a simulation of a dense cell-like topology of co-channel BSSs. It is demonstrated that the positive impact of spatial reuse dominates over the negative impacts of collision and interference in the dense scenario due to the short distances between APs and associated STAs, and the throughput performance is increased by 190 % when choosing optimal CCATs. In [38], the throughput performance of OBSSs with and without CCA adaption is evaluated. The simulation results show that with CCA adaption, the throughput of HEW STAs and legacy STAs can be increased to 59.6 and 12 Mbps from 14 Mbps and 2 kbps without CCA adaption, respectively. The fairness issue between HEW STAs and legacy STAs (no CCA adaption used) is one critical issue in CCA adaption, which will be discussed later. A simulation study of CCA performance is presented in [21], where the impact of different CCATs in HEW simulation scenario 1 (residential), 2 (wireless office), and 3 (indoor small

BSS) is investigated. It is shown that when the CCAT varies from −82 to −30 dBm, both mean and 5 % throughput increase with the CCAT in scenario 2, while in scenario 1 and 3, there are specific values of CCAT to achieve maximum throughput, and the optimal value for mean and 5 % throughput may be different. This is in accordance with the observations found in [22].

### 4.1.2 Dynamic CCA Enhancements

Fix CCA enhancement methods can be efficient in the scenarios where the network is deployed regularly, and the mobility of STAs is low. However, with a more complicated scenario where a huge number of BSSs are deployed densely and irregularly, and the STAs may move fast and frequently, fix and global CCA enhancement methods may not be suitable to achieve satisfying performance. Therefore, dynamic CCA schemes are required to adapt to the complicated scenarios and the STA mobility. Since the dynamic CCA is to adjust the radio sensitivity, they are usually referred to as dynamic sensitivity control (DSC).

In [32], a centralized joint CCA and channel bandwidth control scheme is proposed to adaptively control the CCAT of APs, and optimize the channel bandwidth allocated to STAs, in order to maximize the transmission opportunity of APs, and guarantee the required SINR of STAs. Simulation results show that the proposed joint CCA and channel bandwidth control scheme can achieve approximate twice in average throughput comparing to only CCA adaption is employed. In [31], a centralized CCA adaption scheme is proposed, in which CCATs of APs are set according to the current positions of STAs. A central controller uses a camera to obtain the positions of STAs, and CCATs for APs are then calculated based on the obtained positions in order to achieve a high SINR for STAs.

In [23], a simple distributed DSC scheme is proposed, where the CCATs of STAs are dynamically adapted to the received signal from the AP. STAs measure the received signal strength indication (RSSI) of the AP beacon, and adapts its CCAT $S$ by

$$S = \min(R, L) - M, \tag{3}$$

where $R$ denotes the RSSI of the AP beacon, $L$ sets the upper limit of $R$, and $M$ denotes a margin value. The upper limit $L$ of AP beacon RSSI is to prevent STAs close to AP having too high CCAT which may result in severe hidden terminal problem. The value of margin $M$ should be large enough to guarantee a large enough SNR and to deal with sudden changes in reception of beacon signals. With the proposed DSC scheme, the coverage of a BSS greatly reduces, and therefore the spatial reuse is improved. Through an analysis to a cell-like topology, it is shown that the network capacity can be improved by 7.58 times. In [7], a distributed physical carrier sensing tuning scheme is proposed for dense multiple access point architecture where a STA can associate with multiple APs to exploit the spatial diversity. To enhance the spatial reuse efficiency, the CCATs of STAs are adjusted

based on the packet loss rate and busy ratio. Simulation results show that the proposed scheme can greatly increase the system throughput comparing with the basic DCF scheme. In [18], a distributed CCA adaption scheme is proposed for each PLCP protocol data unit (PPDU). The transmitting device announces the CCAT it permits other devices to use for each PPDU, and dynamically adjusts the value. The aim of the adjustment is that if a STA successfully transmits packets successively, the surrounding STAs can use a higher CCAT to improve the spatial reuse; if a transmission fails, the STA will reset the CCAT of surrounding STAs to a low level in order for the successful transmission of following packets.

### 4.1.3 Fairness Among CCA STAs and Legacy STAs

HEW should be backward compatible, which means HEW should able to coexist with legacy networks and STAs with a certain level of performance preservation of the legacy network. However, with CCA enhancement implemented, HEW devices may have great advantages over legacy devices, which causes severe fairness issues. Consider a room where CCA-enhanced HEW and legacy WiFi network coexist densely, and HEW STAs have higher CCATs than legacy STAs. Due to the CSMA scheme, HEW STAs may have much higher transmission opportunity than legacy STAs since they can transmit when legacy STAs are transmitting, but legacy STAs cannot transmit when HEW STAs are transmitting since the wireless medium is sensed to be busy. In [38], it shows that the throughput of legacy STAs is about only 2 kbps while HEW STAs have 14 Mbps. In [19], through a simulation in residential scenario, it is shown that by changing CCAT from −82 to −62 dBm, the throughput of HEW STAs increases by 36 %, but the throughput of legacy STAs reduces by 48 %.

In [20], the reasons for the fairness issues are explained, and solutions are proposed accordingly. The first reason is that if a CCA-enabled transmission happens during the TXOP of the legacy device, the successive legacy transmission will be delayed until the end of the TXOP of CCA-enabled devices. To address this problem, the TXOP of CCA-enabled devices can be reduced such that the two TXOPs end at the same time. Another fairness issue is that if the CCA-enabled devices transmit in the middle of legacy devices' TXOP, collisions at legacy STAs may be caused, resulting higher power consumption due to retransmissions and reduced throughput. This issue can be addressed by carefully adjusting the transmit power of HEW devices. In [14], a joint TPC and DSC scheme is proposed to enhance the fairness between HEW and legacy devices. The TPC schemes will be discussed in detail in the following part.

### *4.2 CCA with TPC*

TPC schemes can also be used to enhance the spatial reuse in HEW. By properly adjusting the transmit power, the interference among nearby APs and STAs can be mitigated, collisions can be reduced, and more transmission opportunities will arise. In [37], a joint channel allocation and power control algorithm is proposed to adjust the transmit power of APs to minimize the coverage overlap among neighboring BSSs. First, channels are allocated to APs such that neighboring APs have large channel spacing. Then, the transmit power of APs is adapted to STAs' distance to the AP, in order to keep the SINR at STAs a constant. A minimum transmit power is set to keep the AP-STA link from breaking. Simulation results show that the proposed scheme can guarantee a 20 Mbps throughput even when the network density is high (200 APs in the simulated area), while with random channel allocation the throughput is only around 12 Mbps. In addition, interference can also be reduced by more than three times. In [34], a TPC scheme called distributed spatial reuse is devised for each STA to transmit at an exact power in order to minimize the interference and maximize the communication pair. A greedy maximum independent set algorithm is designed to select maximum number of non-interfering communication pair. Through evaluation, it is shown that the proposed scheme can improve the performance of both throughput and delay.

CCA schemes can reduce the coverage of BSSs, and increase concurrent transmission, but have the fairness issue with legacy networks. TPC schemes can decrease the interference among STAs, and control the SINR as required. However, TPC is not effective enough to reduce the coverage and increase the spatial reuse. Therefore, a combination of the two schemes is envisioned to increase the spatial reuse and guarantee the fairness to legacy networks at the same time. In [39], the relation between CCAT, transmit power, and data rate is theoretically analyzed, and based on the results, a distributed joint tuning scheme is proposed to enhance the spatial reuse and guarantee the fairness. The maximum achievable data rate is first estimated based on the RSSI of received AP frame. Then, the transmit power is reduced to save power and reduce interference, and CCAT is increased to improve spatial reuse accordingly. Simulation results show that the aggregate throughput increases from 27 to 55 Mbps by employing the proposed scheme. In [14], a joint CCA and TPC scheme is proposed in which the receiver requests the transmitter to transmit at power equal to CCAT plus margin. In this way, HEW STAs can have higher throughput according to higher CCAT, while legacy STAs have the RSSI of margin value, which guarantees the SINR (and data rate) to a certain level.

## 5 Conclusion

In summary, the next-generation WLAN can employ advanced PHY and MAC schemes to better utilize the spectrum resource, increase the network throughput,

while enhance the per-user throughput to satisfy the varying QoS requirements. WLAN will continue to be one of the most popular access technologies, providing high data rate, low price, and more reliable services to mobile users.

## References


1. A. Aijaz, H. Aghvami, M. Amani, A survey on mobile data offloading: technical and business perspectives. IEEE Wirel. Commun. **20**(2), 104–112 (2013)
2. D. Bharadia, E. McMilin, S. Katti, Full duplex radios. Proc. ACM SIGCOMM **43**(4), 375–386 (2013)
3. G. Bianchi, Performance analysis of the ieee 802.11 distributed coordination function. IEEE J. Sel. Areas Commun. **18**(3), 535–547 (2000)
4. X. Chen, H. Zhai, X. Tian, Y. Fang, Supporting QOS in IEEE 802.11 e wireless lans. IEEE Trans. Wirel. Commun. **5**(8), 2217–2227 (2006)
5. J.I. Choi, M. Jain, K. Srinivasan, P. Levis, S. Katti, Achieving single channel, full duplex wireless communication, in *Proceedings of ACM MobiCom* (2010)
6. S. Dimatteo, P. Hui, B. Han, V.O. Li, Cellular traffic offloading through WiFi networks, in *Proceedings of the IEEE MASS* (Spain, 2011)
7. Y. Hua, Q. Zhang, Z. Niu, Distributed physical carrier sensing adaptation scheme in cooperative map WLAN, in *Proceedings of IEEE GLOBECOM, Hawaii, USA* (2009)
8. M.L. Huang, S. Lee, S.-C. Park, An efficient admission control algorithm for ieee 802.11 e WLAN, in *Proceedings of IEEE VTC-Fall* (2008)
9. IEEE, Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications - amendment 8: medium access control (MAC) quality of service enhancements. Technical Report (2005)
10. IEEE, Especificacion 802.11-2007. Technical Report (2007)
11. IEEE, Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications. Technical Report (2012)
12. IEEE, 802.11-13/1054: HEW evaluation metrics suggestion, Technical Report (2013)
13. IEEE, IEEE 802.11 study group. Status of IEEE 802.11 HEW study group, Technical Report (2014) [Online]. Available: http://www.ieee802.org/11/Reports/hewupdate.htm
14. IEEE, 802.11-14/1207: OBSS reuse mechanism which preserves fairness. Technical Report (2014)
15. IEEE, 802.11-14/0165: 802.11 HEW SG proposed par, Technical Report (2014)
16. IEEE, 802.11-13/0657: usage models for IEEE 802.11 high efficiency WLAN study group (HEW SG) Liaison with WFA, Technical Report (2014)
17. IEEE, 802.11-15-0043-01-00ax: in-band full duplex radios and system performance, Technical Report (2014)
18. IEEE, 802.11-14/0872: a protocol framework for dynamic CCA. Technical Report (2014)
19. IEEE, 802.11-14/0372: HEW system level simulations on increased spatial reuse. Technical Report (2014)
20. IEEE, 802.11-14/0637: spatial reuse and coexistence with legacy devices. Technical Report (2014)
21. IEEE, 802.11-14/0889: performance gains from CCA optimisation. Technical Report (2014)
22. IEEE, 802.11-14/0846: CCA study in residential scenario. Technical Report (2014)
23. IEEE, 802.11-13/1012: dynamic sensitivity control. Technical Report (2014)
24. M. Jain, J.I. Choi, T. Kim, D. Bharadia, S. Seth, K. Srinivasan, P. Levis, S. Katti, P. Sinha, Practical, real-time, full duplex wireless, in *Proceedings of ACM MobiCom* (2011)
25. I. Jamil, L. Cariou, J.-F. Helard, Improving the capacity of future IEEE 802.11 high efficiency WLANs, in *Proceedings of IEEE ICT, Lisbon* (2014)

26. T. Kim, S. Lee, S.-C. Park, Call admission control based on adaptive physical rate for EDCA in IEEE 802.11 e WLAN system, in *Proceedings of IEEE CCNC* (2008), pp. 59–61
27. J.Y. Kim, O. Mashayekhi, H. Qu, M. Kazadiieva, P. Levis, Janus: a novel MAC protocol for full duplex radio. Stanford CS Tech reports in http://hci.stanford.edu/cstr/ CSTR **2**(7), 23 (2013)
28. K. Lee, J. Lee, Y. Yi, I. Rhee, S. Chong, Mobile data offloading: how much can WiFi deliver? IEEE/ACM Trans. Netw. **21**(2), 536–550 (2013)
29. P. Liu, Z. Tao, S. Narayanan, T. Korakis, S.S. Panwar, Coopmac: a cooperative MAC for wireless lans. IEEE J. Sel. Areas Commun. **25**(2), 340–354 (2007)
30. K. Medepalli, F. Tobagi, D. Famolari, T. Kodama et al., On optimization of csma/ca based wireless lans: Part ii-mitigating efficiency loss, in *Proceedings of IEEE ICC* (2006)
31. K. Murakami, T. Ito, S. Ishihara, Improving the spatial reuse of IEEE 802.11 WLAN by adaptive carrier sense threshold of access points based on node positions, in *Proceedings of IEEE ICMU, Hakodate City* (2015)
32. T. Nakahira, K. Ishihara, Y. Asai, Y. Takatori, R. Kudo, M. Mizoguchi, Centralized control of carrier sense threshold and channel bandwidth in high-density WLANs, in *Proceedings of IEEE APMC, Sendai, Japan* (2014)
33. A. Sahai, G. Patel, A. Sabharwal, Pushing the limits of full-duplex: design and real-time implementation (2011). arXiv:1107.0607
34. K.-P. Shih, C.-M. Chou, M.-Y. Lu, S.-M. Chen, A distributed spatial reuse (DSR) MAC protocol for IEEE 802.11 ad-hoc wireless LANs, in *Proceedings of IEEE ISCC, Cartagena* (2005)
35. S. Singh, H. Dhillon, J. Andrews, Offloading in heterogeneous networks: modeling, analysis, and design insights. IEEE Trans. Wirel. Commun. **12**(5), 2484–2497 (2012)
36. C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, E. Hepsaydir, Cellular architecture and key technologies for 5g wireless communication networks. IEEE Commun. Mag. **52**(2), 122–130 (2014)
37. Y. Wu, Y. Sun, Y. Ji, J. Mao, Y. Liu, A joint channel allocation and power control scheme for interference mitigation in high-density WLANs, in *Proceedings of IEEE ICCT, Guilin* (2013)
38. A.E. Xhafa, A. Batra, A. Zaks, On the coexistence of overlapping BSSs in WLANs, in *Proceedings of IEEE VTC, Baltimore* (2007)
39. Z. Zhou, Y. Zhu, Z. Niu, J. Zhu, Joint tuning of physical carrier sensing, power and rate in high-density WLAN, in *Proceedings of IEEE APCC, Bangkok* (2007)

# Shaping 5G for the Tactile Internet

**Adnan Aijaz, Meryem Simsek, Mischa Dohler, and Gerhard Fettweis**

**Abstract** The Tactile Internet is envisioned to transport touch and actuation in real-time. It will enable unprecedented applications and revolutionize almost every segment of the society. It is expected that the next generation 5G mobile communications networks will enable the Tactile Internet at the wireless edge. The tactile internet creates daunting new requirements for 5G network design. This chapter focuses on the key technological concepts which lay at the intersection of 5G and the Tactile Internet. The chapter outlines the key application areas of the Tactile Internet. This is followed by an end-to-end architecture of the Tactile Internet. The chapter also presents the key technical requirements of the Tactile Internet along with some potential solutions to meet these requirements. Such solutions revolve around protocol-level and system-level innovations.

## 1 Introduction

The Internet is unarguably the most important invention of the twentieth century. What originally started as a network of a handful of nodes, went on to define the economies of the late twentieth century. The wireless revolution in general and the unprecedented development of mobile/cellular technologies in particular, has totally transformed the way we perceive the Internet today by creating what could be termed as the Mobile Internet. The focus of mobile communications is now moving towards providing ubiquitous connectivity for machines and devices which will enable the Internet-of-Things. In a natural evolution to these Internet embodiments, the notion of Tactile Internet [1, 2] is emerging, in which

A. Aijaz (✉)
Toshiba Research Europe Ltd., Cambridge, UK
e-mail: adnan.aijaz@toshiba-trel.com

M. Simsek • G. Fettweis
Technical University Dresden, Dresden, Germany
e-mail: meryem.simsek@tu-dresden.de; fettweis@ifn.et.tu-dresden.de

M. Dohler
King's College London, London, UK
e-mail: mischa.dohler@kcl.ac.uk



W. Xiang et al. (eds.), *5G Mobile Communications*,
DOI 10.1007/978-3-319-34208-5_25

ultra-reliable, ultra-responsible, and sufficiently intelligent network connectivity will enable the delivery of real-time control and physical haptic experience, remotely. Unlike the conventional Internet, which provides the medium for voice and data communications, the Tactile Internet will provide the ability to transport touch and actuation through the Internet in real-time. Currently, the conventional Internet and its different embodiments are widely used for delivery content services (voice telephony, text messaging, file sharing, etc.). The Tactile Internet will provide a true paradigm shift from content delivery to remote skill-set delivery, and thereby revolutionize almost every segment of the society. Recently, standardization efforts for the Tactile Internet have started to emerge within IEEE and ETSI. State-of-the-art fourth generation (4G) mobile communications networks do not largely fulfil the technical requirements of the Tactile Internet. Therefore, it is expected that the fifth generation (5G) mobile communications networks will underpin the Tactile Internet at the wireless edge.

5G networks are expected to be commercially deployed by 2020. An early assessment of 5G scenarios and requirements has been carried out in the flagship EU METIS project [17] and more recently by the telecommunications industry alliance NGMN [18]. With Tactile Internet, daunting new requirements for 5G arise. Meeting such requirements will require revolutionary approaches to 5G network design.

Research efforts for both 5G and the Tactile Internet are in infancy. The main focus of this chapter is thus the intersection of 5G and Tactile Internet. To this end, this chapter is organized as follows. In Sect. 2 we provide a brief overview of 5G networks. In Sect. 3 we outline some of the key application areas of the Tactile Internet. Section 4 covers the end-to-end architecture for the Tactile Internet. In Sect. 5 we outline the key technical requirements for the Tactile Internet from a networking perspective. Section 6 discusses some potential solutions for 5G networks in realizing the Tactile Internet. Finally, the chapter is concluded in Sect. 7.

## 2 5G: A Brief Overview

The fifth generation (5G) of mobile communication networks has attracted a lot of attention from industry, research centers, and academia. 5G is expected to enable a fully mobile and connected society and to empower socio-economic transformations and to fulfill the wireless communication requirements for 2020 and beyond. The demands of a fully mobile and connected society are characterized by the tremendous growth in connectivity and density/volume of traffic, the required multi-layer densification in enabling this, and the broad range of use cases and business models expected. Overall there is a common understanding that 5G should not only support an evolution of traditional mobile communication services, such as personal mobile multimedia communication or personal mobile broadband services; 5G should in addition address novel use cases and connect a massive number of devices to the Internet of Things. The addressed novel use cases for

5G include machine-type communication in fields like, e.g. smart grids, vehicular communication and intelligent transport systems. These use cases have stringent requirements which were not in the focus of former mobile network generations. Hence, 5G is expected to cope with massive and stringent requirements in terms of throughput, ultra-low latency, ultra-high reliability, higher connectivity density, security and privacy, and higher mobility range when and where needed [4]. Hence, 5G will operate in a highly heterogeneous environment with a mix of various requirements for different use cases. This leads to the need to rethink the overall system architecture for 5G to enable flexibility for optimized network usage. The 5G architecture should include modular network functions which can be deployed and scaled on demand. The broad range of targeted 5G capabilities, i.e. stringent requirements and flexible, highly reactive network architecture, will make 5G an important enabler for the Tactile Internet—when humans will wirelessly control real and virtual objects in real-time—but not before we overcome enormous system design challenges.

## 3 Key Tactile Internet Applications

The Tactile Internet opens whole new areas of technology to the Internet, based on very low end-to-end latency. It enables control of real and virtual objects and adds a new dimension to human-machine interaction in real-time through the conveyance of the perception of touch through communication. The Tactile Internet has been described as a communication infrastructure combining low latency, very short transit time, high availability and high reliability with a high level of security [9]. Associated with cloud computing proximity through e.g. mobile edge-clouds and combined with the virtual or augmented reality for sensory and haptic controls, the Tactile Internet addresses areas with reaction times in the order of a millisecond. This enables the Tactile Internet to encompass mission critical applications (e.g., manufacturing, transportation, and healthcare), as well as non-critical applications (e.g., edutainment and events).

### *3.1 Manufacturing*

Around the world, traditional manufacturing industry is over to be revolutionized by a digital transformation that is accelerated by exponentially growing technologies, e.g. intelligent robots, sensors etc. The widespread adoption by manufacturing industry around the framework of information and communication technology is now paving the way for distributive approaches to development, production, and the entire logistics chain.

Industrial countries have already witnessed three industrial revolutions, which could also be described as disruptive leaps in industrial processes resulting in

significantly higher productivity [10]. In the first industrial revolution (starting in the 1780s), mechanical production facilities with the help of water and steam power was introduced. The second industrial revolution started in the 1870s and enabled the introduction of division of labor and mass production with the help of electrical energy. The use of electronic and Information Technology (IT) systems further automated the production in the third industrial revolution together with the first programmable logic controller in 1969. The upcoming fourth industrial revolution (industry 4.0) is based on real-time enabled cyber-physical systems and comes with key changes in manufacturing, engineering, material usage and supply chain and life cycle management, i.e. leading to flexible and self-organized smart factories. With increased agility and flexibility in the production process, industry 4.0 will enable to individualize any product and allow to react faster on changing market requirements. Cyber-physical systems in industry 4.0 comprise smart machines, storage systems and productions facilities capable of autonomously exchanging (wirelessly) information, triggering actions and controlling each other independently.

Industry 4.0 will naturally come together will some challenges. The services and applications provided by cyber-physical system platforms will connect people, objects, and systems to each other. This possesses novel requirements on safety, security, and reliability for everything from sensors to user interfaces in real-time [11]. Among other issues, this will involve addressing the challenges posed by the wide range of different data sources and devices. Identified critical system parameters in smart factories are latency (and jitter), safety, and energy consumption. Latency requirements of machines are in the range of several milliseconds, the sensitivity of rapidly moving devices' control circuits is significantly below 1 ms per sensor, while subsystems rely on a latency of several micro seconds. Hence, smart factories together with automation in industry is a key application field in the Tactile Internet.

### *3.2 Autonomous Transportation*

Technological advancements are creating a continuum between conventional, fully human-driven vehicles and autonomous vehicles, which partially or fully drive themselves. Autonomous vehicle technology with self-driving features will revolutionize the driving experience, and consumers will need time to learn how to use and manage the new features. In partially self-driving vehicles, automobiles are set-up with certain features such as self-parking, emergency braking, and adaptive cruise control and are already reality in selected applications that feature controlled environments, such as mining and farming. One of the next steps in autonomous vehicles with partially self-driving capabilities will be most-likely on-highway trucks as the first vehicles to feature the autonomous driving technology on public roads.

Fully self-driving cars, on the other hand, are also no longer a futuristic idea. Various companies have already released self-driving features that give the car the ability to drive itself. Compared to partially self-driving cars, fully self-driving cars' steering wheel will no longer exist. The vehicles will drive by themselves by using the same system of sensors, radar and GPS mapping that partially self-driving vehicles employ. Besides self-evident benefits, such as easier parking, significantly reduced (zero in the optimal case) accident rates, one of the key profits of self-driving cars is time saving. Autonomous car drivers will be able to spend travelling time working, relaxing, or accessing entertainment/digital media on the road which will have a big economic impact. However, these benefits come at the cost of supplementary challenges requiring significant improvement in communication technologies and hardware for the sensor domain, to get a reliable environment model, and for the connectivity domain, to obtain information from other cars and the infrastructure.

To facilitate fully self-driving capabilities, recent advancement in 5G wireless communication technologies and automobiles have enabled the evolution of Intelligent Transport System (ITS) to wirelessly connected and communicating vehicles. ITS addresses various vehicular traffic issues like traffic congestion, information dissemination, accident avoidance etc. Vehicular Ad-hoc NETwork (VANET), in particular, which is a distinctive class of Mobile ad-hoc Network (MANET), is an integral component of ITS. Hereby, wireless communication technologies play a vital role in supporting both Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) communication in VANET. Typically, a VANET consists of four major components namely: Vehicles, (GPS enabled) Devices/Sensors, road-side info-stations and Traffic Management Centre (TMC) [14, 15]. These components communicate using wireless communication standards/protocols that regulate the various aspects of communication such as data transmission range and rate, latency and security.

Enabling real-time communication and enhanced security, a considerable and sustainable reduction of road accidents and traffic jams can be realized by fully autonomous driving. The time needed for collision avoidance in today's applications for vehicle safety is below 10 ms [12]. In case of a bidirectional data exchange for automatic driving manoeuvres, a latency in the order of a millisecond will likely be needed [13]. This can technically be realized by the Tactile Internet and its 1 ms end-to-end latency. Fully autonomous driving is expected to change the traffic behavior entirely. Especially small distances between automated vehicles, in particular in platoons, potentially safety critical situations need to be detected earlier than with human drivers. This requires ultra-high reliable and proactive/predictive behavior in future wireless communication systems.

### *3.3 e-Health*

Information Technology (IT) is a key in the field of e-health. Most of the technological advancement in wireless networking has been applied to advance healthcare services. E-health and health care services are information based and, hence, a

better utilization of information can make health services more incorporated to enhance the patient safety. Consequently, wireless standards and IT need to be further improved to provide efficient, reliable, and robust real-time health services in tele-diagnosis, tele-surgery and tele-rehabilitation (i.e. tele-medicine). In [16], the impact of latency on a sergeant's precision in a robotic-assisted remote telepresence surgery has been examined. It has been shown that a non-real-time system leads to errors in the sergeant's behavior.

Tele-medicine uses (wireless) communication and information technology to overcome geographic distances, and increase access to health care services. This is particularly beneficial for emergency situations, long distances between patient and sergeant, and for rural and underserved communities in developing countries suffering from absence of access to health care.

Using advanced tele-diagnostic tools, medical expertise could be available anywhere and anytime regardless of the physician's location. Hereby, a tele-robot at the patient's location will be controlled by the physician, so that not only audio and/or visual information but also haptic feedback is provided. The same technical principle is applied to tele-surgery applications.

E-health comes together with stringent requirements on the reliability of wireless connection. Especially in tele-surgery and tele-diagnostic, reliability is of particular importance. Unreliable connectivity can lead to delayed imaging, so that poor image resolution may limit the efficacy of the sergeant's remote handling. In addition, an accurate tele-medical treatment can only be realized with haptic feedback which in turn is possible if the (physician/) human-to-machine interaction can be facilitated in real-time. This requires a deterministic real-time behavior which is not supported by recent communication systems. An end-to-end latency of a few milliseconds together with ultra-high reliability in wireless link connection and data transmission is required in e-health which can be realized by the Tactile Internet.

### *3.4 Edutainment and Gaming*

Edutainment is a combined word of education and entertainment and describes educational content with entertainment value. Teaching of exacting manual operation demanding fine-motor skills in a verbal or classical way is challenging and require permanent repetitions and corrections by the teacher. The Tactile Internet promises a completely new way of educations based on the haptic interaction of teachers and students, i.e. while a student performs a manual operation (on a virtual or real object), the teacher can follow him and correct him in case of failures remotely. To realize this in a timely synchronous manner, an identical multi-modal, i.e. visual, auditory, and haptic human-machine-interface, which operates on an extremely low end-to-end latency, is essential. This can be applied to medical students during their first surgeries or to students learning to play a musical instrument. Furthermore, the Tactile Internet will enable the distributed interactive playing of music, e.g. joint practicing of band members. This is with today's technologies not possible due to the high end-to-end latency.

Serious gaming, on the other hand, are real-world simulations designed for the purpose of solving a problem. The end-to-end latency in the interaction between players and games is a key factor influencing the quality of players' experience and the game's usability, since the delay influences directly the perceived realism of the game. Other key requirements in gaming are the synchronicity and extremely low delays in multi-player games. If this is not achieved, each player will experience a different status of the game and a so-called lag, which leads in the worst case to unfairness among players.

## *3.5 Further Application Areas*

Other Tactile Internet applications are for example virtual and augmented reality. The virtual reality is a shared, haptic virtual environment in which several users are physically coupled via a simulation tool to jointly/collaboratively perform tasks by perceiving the objects not only audio-visually but also via the touch sense. In augmented reality, on the other hand, the combination of real and computer generated content is visualized in the user's field of view. The major goal of future augmented reality applications, compared to today's static information augmentation, is the visualization of dynamic content and up-to-date information. Haptic feedback in virtual reality is a prerequisite for high-fidelity interaction. Especially, the perception of objects in virtual reality via the sense of touch leads to various applications relying on high level of precision. This precision can only be realized if the latency between the users and the virtual reality is a few milliseconds. The augmentation of additional information into a user's field of view enables the development of many assistance systems, e.g. maintenance, driver-assistance systems, education. With the Tactile Internet the content in augmented reality can be moved from static to dynamic. This enables a real-time virtual extension of a user's field of view, so that possible dangerous events can be identified and avoided.

Unmanned autonomous or remotely controlled systems are increasingly used in a large number of contexts to support humans in dangerous and difficult-to-reach environments, remotely controlled by humans, or for tasks that are too tedious or repetitive for humans. The remote control of an unmanned aircraft, for example, can be realized with high precision and without any reaction delay with a reduced end-to-end latency as a Tactile Internet application.

Besides these applications, efficient and reliable smart grids together with electro mobility, smart metering, and the automation of energy distribution and cloud computing (especially mobile edge clouds) are some other applications that can significantly benefit from the availability of the Tactile Internet.

## 4 End-to-End Architecture for the Tactile Internet

Unlike the conventional Internet which provides the medium for audio and visual transport, the Tactile Internet is envisioned to provide the medium for transporting touch, actuation, and control in real-time. The Tactile Internet will enable haptic communications [3] as one of the primary applications. The exchange of haptic information generally closes a global control loop with strict latency constraints due to the presence of haptic feedback. In the absence of haptic feedback, there is conventional audio/visual feedback only and no notion of a control loop is involved. Another key application of the Tactile Internet will be networked control systems, wherein sensors and actuators are connected and highly dynamic processes are controlled. The control and feedback signals are exchanged in the form of information packets through the network, closing a global control loop and leading to strict latency constraints. Unlike haptic communications, the human element is not involved in case of networked control systems.

As shown in Fig. 1, the end-to-end architecture for the Tactile Internet can be split into three distinct domains: a master/controller domain, a network domain, and a slave/controlled domain. In case of haptic communications and remote operation, the master domain consists of a human (operator) and human system interface (HSI). The HSI (usually a haptic device) converts the human input into a tactile input through various coding techniques and allows a user to manipulate objects in real and virtual environments. The master domain primarily controls the operation of the remote domain as discussed later. The master domain also has provisioning for audio and visual feedbacks which not only enable non-haptic control and remote operation but also play a critical role in enhancing the perceptual performance. In case of networked control systems, the master domain comprises a controller which gives commands to the sensor and actuator system.

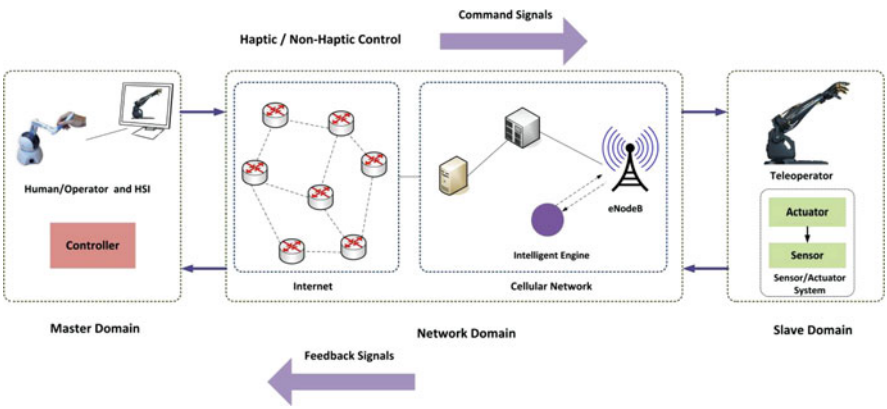


**Fig. 1** An illustration of the end-to-end Tactile Internet architecture

The slave domain consists of a teleoperator (controlled robot) in case of haptic communications, which is directly controlled by the master domain through various command signals. The teleoperator interacts with the remote environment and usually sends a haptic feedback on the reverse path. Typically, no *a priori* knowledge exists about the environment. In case of networked control systems, the slave domain consists of a sensor and actuator system. The sensors measure the state of the system whereas the actuators manipulate the system based on the commands received from the controller.

The network domain provides the medium for communication between master and slave domains. The Tactile Internet requires ultra-reliable and ultra-responsive network connectivity that would enable typical reliabilities and latencies for real-time control and remote operation. The underlying 5G-driven communication architecture is expected to meet the key requirements in realizing the vision of the Tactile Internet.

## 5 Networking Requirements for the Tactile Internet

The Tactile Internet will not be realized without overcoming an enormous set of cross-disciplinary design challenges, in order to meet the underlying requirements. Some of the most stringent design challenges for the Tactile Internet have been recently presented in [2]. In the following, we highlight the key technical requirements, from a networking perspective, for realizing the Tactile Internet.

### *5.1 Ultra-Reliable Connectivity*

The Tactile Internet is expected to service critical areas of the society, and therefore requires ultra-reliable network connectivity. The term ultra-reliability is quantified as seven nines reliability i.e., an outage probability of $10^{-7}$, which translates to milliseconds of outage per day. Ultra-reliability is critical for keeping packet losses to a minimum especially in error prone wireless environments. Packet losses results in various types of artifacts for haptic communications that directly affect the activity of the operator. Similarly, packet losses may destabilize a networked control systems.

### *5.2 Ultra-Responsive Connectivity*

Due to the real-time context, the Tactile Internet requires ultra-responsive network connectivity on the order of single-digit ms, to achieve truly immersive steering and control. For tactile application requiring no feedback, latency of up to 5 ms is tolerable. However, in case of feedback, the latency requirement becomes much more stringent; on the order of 1 ms.

The end-to-end latency in a communication system includes the time spent in processing and transmission of information at the transmitter end, propagation of information to the receiver end, and re-transmission from the receiver end (including processing and propagation time) in case of feedback. Considering the end-to-end latency in a communication systems, the latency budget for air-interface is reduced by an order of magnitude.

## *5.3 Edge-Intelligence*

Edge-Intelligence would be critical in overcoming two critical challenges in realizing the Tactile Internet. The first challenge is how to overcome the physical limitation arising due to the finite speed of light. This limits the range of tactile actuation to approximately 150 km (assuming no transmission, processing, etc.). The second challenge arises due to the fundamental nature of wireless channel. Even if a communication infrastructure providing latency of 1 ms is realized, deterministic delay guarantees cannot be provided as the wireless channel is time-varying in nature. Therefore, stability of control loop in case of haptic communications and networked control systems cannot be guaranteed. To overcome these challenges, the Tactile Internet must support a hybrid composition of machine and human actuation mixing real tactile actuation with intelligence-based predictive actuation. Such predictive actuation should be in close proximity of the tactile edge. Therefore, the edge of the network (mobile edge cloud) must be equipped with intelligence to facilitate predictive caching as well as interpolation/extrapolation of human actions. This necessitates the development of novel artificial intelligence techniques for edge-cloud architectures.

## *5.4 Security and Privacy*

Security and privacy are become particularly important requirements for the Tactile Internet as its application areas range from large-scale industrial systems to critical infrastructures and services. Security becomes equally important for wireless transmission and remotely operated entities. With stringent latency constraints, security must be embedded in the physical transmission and ideally be of low computational overhead. Novel coding techniques need to be developed for tactile applications that allow only the legitimate receivers to process a secure message. Absolute security will, hereby, be achieved if an illegitimate receiver cannot decode the date even with infinite computational power. This rises a challenge, especially in massive connectivity applications. Identification of legitimate receivers requires novel, reliable and low-delay methods. One such method could be the usage of hardware specific attributes such as biometric fingerprints.

### 5.5 Codecs Family

The Tactile Internet must handle the tactile information in the same way as the conventional audio/visual information. Hence, a standard codecs family, similar to the state-of-the-art audio (ITU-TH.264) and video (ISO/IEC MPEG-4) codecs is needed, which facilitate transmission of tactile data over packet-switched networks. Such a codec family would be a key enabler for scalability at the network edge and universal uptake.

## 6 Network Design for 5G-Enabled Tactile Internet

The Tactile Internet creates daunting new requirements for 5G networks. Such requirements can be met through advances in Physical layer techniques, novel approaches to legacy wireless access and radio resource management protocols, and innovative architectural designs. In the following, we discuss some potential solutions to meet the technical requirements for realizing the Tactile Internet.

### 6.1 Network Slicing

5G networks will support a range of use cases, spanning different vertical industries, not all of which share the same service requirements as the Tactile Internet. Therefore, one of the most important challenges lying at the intersection of 5G and the Tactile Internet is how one network, based on a common physical infrastructure, can be efficiently shared among different vertical applications. The research community has a general consensus that 5G networks must be designed in a flexible manner to cater for the service requirements of different vertical applications. Such flexibility is possible through network slicing [5], which will be indispensable for 5G network design. A network slice can be defined as a connectivity service, based on various customizable logical network (and associated device) functions, supporting the requirements of a particular use case. The basic idea behind network slicing is to create multiple logical networks over a common physical infrastructure, with each network tailored to the specific needs of a use case. Such envisioned network slicing would be possible through network function virtualization (NFV) [6] and software-defined networking (SDN) [7] paradigms. NFV provides the separation of network functions from the hardware infrastructure which can managed as software module in any standard cloud-computing infrastructure. SDN decouples control and data planes and enables direct programmability of network control through software-based controllers. Therefore, the end-to-end 5G network architecture must be designed to facilitate flexible network slicing.

## 6.2 Physical Layer

Given its attractive features, orthogonal frequency division multiple access (OFDM) is the primary Physical layer candidate for 5G networks. To overcome some of the weaknesses of OFDM, several variants have been proposed recently such as filter bank multi-carrier (FBMC) [FBMC], generalized frequency division multiplexing (GFDM) [GFDM], bi-orthogonal OFDM (BFDM) [BFDM], etc. It should be noted that OFDM, with its current adaptation for legacy 4G networks is not suitable for tactile applications. An end-to-end latency constraint of 1 ms pushes one-way air-interface transmission budget to approximately 100 μs. However, OFDM symbol duration in LTE/LTE-A is approximately 70 μs. One way to meeting this transmission budget is to change the underlying OFDM numerology in terms of symbol duration, sub-carrier spacing, FFT block size, cyclic prefix etc. Given the increasingly software nature of radios, it can be easily expected that OFDM would be implemented in a *tunable* manner to cater for requirements of different use cases.

In addition to waveform design issues, novel enhancements are required at the Physical layer for meeting the stringent Tactile Internet requirements. These include enabling high levels of diversity (for reliable signal detection and decoding), implementing fast channel estimation/equalization techniques, and using faster decoding mechanisms.

## 6.3 Protocol Stack Optimizations

The stringent requirements of the Tactile Internet require a revisit of legacy wireless access protocols. In order to reduce air-interface latency, optimizations are needed for both control and data planes. For example, to ensure the stability of closed-loop tactile applications, the base station must support a fast connection re-establishment feature at the radio link layer so that any ongoing tactile session is transparent to radio link failures. From a medium access control (MAC) layer perspective, the transmission time interval (TTI) needs to be reduced in order to achieve end-to-end latency on the order of 1 ms. Such reduction of TTI would be possible through spectrum aggregation techniques. To achieve the required levels of reliability, diversity becomes critically important. However, diversity must be embedded in frequency and spatial domains considering its enhanced latency if applied in the time domain. Hence, link-level techniques like hybrid ARQ (HARQ) must be replaced with suitable equivalents in frequency or spatial domains. Improving wireless connectivity to the required level for tactile applications is not an invincible task. By exploiting frequency diversity over multiple uncorrelated links, required reliability level can be approached.

## 6.4 Radio Resource Management

Radio resource management is a key functionality of cellular networks. It has a direct impact on QoS of a user and performance of higher layers. With the introduction of tactile applications into the 5G ecosystem, resource management becomes particularly challenging as available resources are shared with other human-to-human (H2H) and machine-to-machine (M2M) applications having different and often conflicting service requirements. Due to stringent service requirements radio resource must be allocated on priority for tactile applications without any external competition. In this context, slicing of radio resources (within broader context of network slicing) becomes particularly attractive. A separate slice of radio resources can be allocated to tactile applications which remains dedicated for any ongoing operation. Such radio resource slicing would be achieved through virtualization of radio resources. In this context, a flexible radio resource slicing strategy is desirable for 5G networks which not only maximizes the utility of various applications, but also maintains high utilization of the overall radio resource pool. Another benefit of such radio resource slicing is the application specific customization of radio resources within the allocated slice. Such customization of radio resources for haptic communications has been recently investigated in [8].

## 6.5 Core Networking

The key functionalities of the 5G core network relevant to the Tactile Internet include dynamic application-aware QoS provisioning, edge-cloud access, and security. Overall, a *thin* core network is desirable with substantial decrease in the protocol overhead. Core network thinning would be achieved through its functional decomposition and moving some of the functionalities to the access network. This will reduce the number of nodes in the data path and hence reduce the end-to-end latency. Functional decomposition is a method to separate out tightly coupled sub-functions of a network entity and can be achieved in either horizontal or vertical manner. The software-defined networking (SDN) paradigm is particularly attractive for the 5G core networking. It will introduce the required programmability and hence the flexibility for such functional decomposition through which data flows can be tailored according to the requirements of different applications.

The existing IP security functionalities are sufficient for providing the required security in the Tactile Internet. However, the placement of IP security far from the tactile edges increases end-to-end latency. Hence, novel approaches are needed to provide adequate security for tactile applications with minimal delays.

Finally, latency in the core Internet must be reduced, which is currently variable and largely dictated by queuing delays and geographic routing policies.

## 7 Concluding Remarks

The Tactile Internet is expected to have a massive impact on business and society. It will revolutionize almost every segment of the society by enabling wireless control and remote operation in a range of scenarios. It will create new opportunities for vendors, operators, content providers, and other members of the service chain.

The next generation (5G) mobile communications networks will play an important role in realizing the Tactile Internet. Research for both 5G networks and Tactile Internet is at a nascent stage. This chapter investigated the interesting area of 5G and Tactile Internet intersection. After introducing exciting Tactile Internet applications, key requirements for the Tactile Internet, from a networking perspective, have been identified. The chapter covered several technical issues and challenges in shaping 5G networks for realizing the vision of the Tactile Internet. The most important challenge would be to ensure tight and scalable integration of various technological solutions into a single network.

## References

1. G. Fettweis, The tactile internet: applications and challenges. IEEE Veh. Technol. Mag. **9**(1), 64–70 (2014)
2. A. Aijaz, M. Dohler, A.H. Aghvami, V. Friderikos, M. Frodigh, Realizing the tactile internet: haptic communications over next generation 5G cellular networks. IEEE Wirel. Commun. (2015), [Online]. Available: http://arxiv.org/abs/1510.02826
3. E. Steinbach et al., Haptic communications. Proc. IEEE **100**(4), 937–956 (2012)
4. J. Andrews et al., What will 5G be? IEEE J. Sel. Areas Commun. **32**(6), 1065–1082 (2014)
5. Ericsson, 5G systems: enabling industry and society transformation. *White Paper*, Jan. 2015. [Online]. Available: http://www.ericsson.com/res/docs/whitepapers/what-is-a-5g-system.pdf
6. B. Han, V. Gopalakrishnan, L. Ji, S. Lee, Network function virtualization: challenges and opportunities for innovations. IEEE Commun. Mag. **53**(2), 90–97 (2015)
7. B. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, T. Turletti, A survey of software-defined networking: past, present, and future of programmable networks. IEEE Commun. Surv. Tutorials **16**(3), 1617–1634 (2014)
8. A. Aijaz, Towards 5G-enabled tactile internet: radio resource allocation for haptic communications, in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC) Workshops*, April 2016
9. ITU-T, The Tactile Internet. Technology Watch Report, 2014. [Online]. https://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000230001PDFE.pdf
10. Deloitte, Industry 4.0 – challenges and solutions for the digital transformation and use of exponential technologies. *White Paper*, Oct. 2014. [Online]. http://www2.deloitte.com/content/dam/Deloitte/ch/Documents/manufacturing/ch-en-manufacturing-industry-4-0-24102014.pdf
11. N. Nikaein, S. Krea, Latency for real-time machine-to-machine communication in LTE-based system architecture, in *European Wireless (EW) Conference*, Vienna, Austria, April 2011
12. J.D. Lee, D.V. McGehee, T.L. Brown, M.L. Reyes, Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator. Hum. Factors Ergon. Soc. **44**(2), 314–334 (2002)

13. S.B. Mer, Smart vehicle-to-vehicle communication with 5G technology. Int. J. Recent Innov. Trends Comput. Commun. **3**(5), 3241–3244 (2015)
14. M. Faezipour, M. Nourani, A. Saeed, S. Addepalli, Progress and challenges in intelligent vehicle area networks. Commun. ACM **55**(2), 90–100 (2012)
15. T. Kosch, I. Kulp, M. Bechler, M. Strassberger, B. Weyl, R. Lasowski, Communication architecture for cooperative systems in Europe. IEEE Commun. Mag. **47**(5), 116–125 (2009)
16. M. Anvari et al., The impact of latency on surgical precision and task completion during robotic-assisted remote telepresence surgery. Comput. Aided Surg. **10**(2), 93–99 (2005)
17. A. Osseiran et al., Scenarios for 5G mobile and wireless communications: the vision of the METIS project. IEEE Commun. Mag. **52**(5), 26–35 (2014)
18. NGMN, 5G White Paper. Feb. 2015. [Online]. https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf