www.zte.com.cn/magazine/English

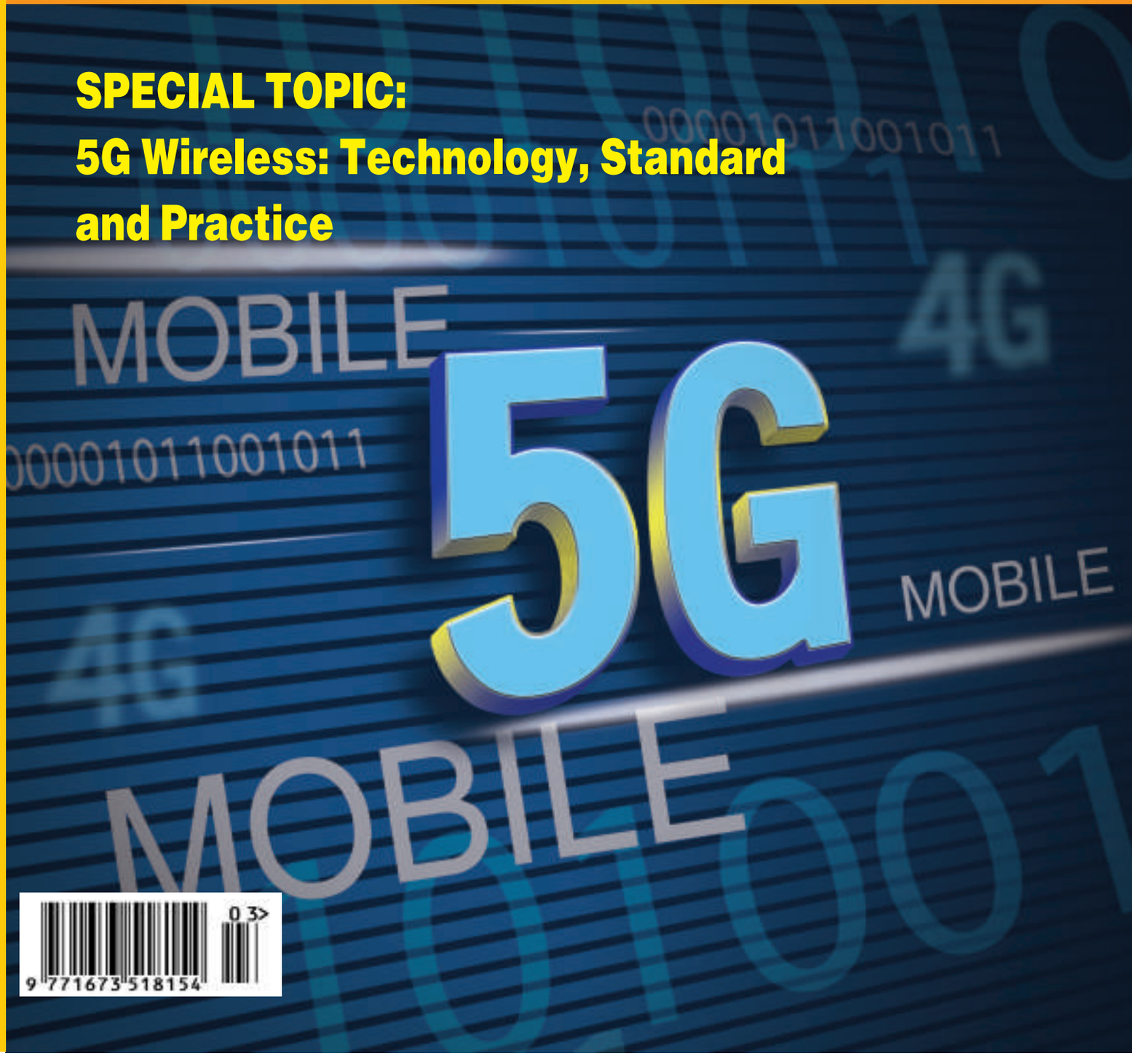
ISSN 1673-5188
CODEN ZCTOAK

ZTE COMMUNICATIONS

An International ICT R&D Journal Sponsored by ZTE Corporation

March 2015, Vol. 13 No. 1

SPECIAL TOPIC:
5G Wireless: Technology, Standard and Practice

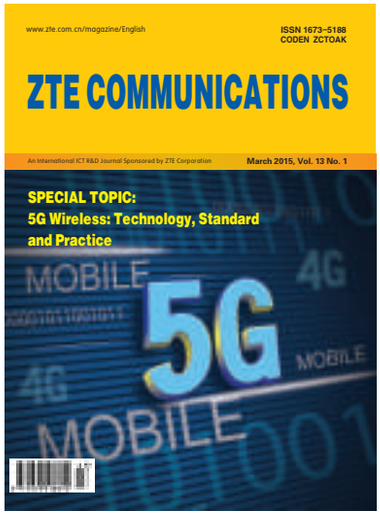


9 771673 518154 03>

ZTE中兴 ZTE COMMUNICATIONS
VOLUME 13 NUMBER 1 MARCH 2015

ZTE Communications Editorial Board

CONTENTS



Special Topic: 5G Wireless: Technology, Standard and Practice

CONTENTS

ZTE COMMUNICATIONS
Vol. 13 No. 1 (Issue 45)
Quarterly
First English Issue Published in 2003

Supervised by:
Anhui Science and Technology Department

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

Staff Members:
Editor-in-Chief: Sun Zhenge
Executive Associate Editor-in-Chief: Huang Xinming
Editor-in-Charge: Zhu Li
Editors: Paul Sleswick, Xu Ye, Yang Qinyi, Lu Dan
Producer: Yu Gang
Circulation Executive: Wang Pingping
Assistant: Wang Kun

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

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

Printed by:
Hefei Zhongjian Color Printing Company

Publication Date:
March 25, 2015

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

Advertising License:
皖合工商广字0058号

Annual Subscription:
RMB 80

5G Wireless: Technology, Standard and Practice

▶ Fa-Long Luo

Dr. Fa-Long Luo is chief scientist at two leading international companies, headquartered in Silicon Valley, dealing with software-defined radio and wireless multimedia. From 2007 to 2011, he was the founding editor-in-chief of *International Journal of Digital Multimedia Broadcasting*. From 2011 to 2012, he was the chairman of IEEE Industry DSP Standing Committee and technical board member of IEEE Signal Processing Society. He is now associate editor of *IEEE Access and IEEE Internet of Things Journal*. He has 31 years of research and industry experience in multimedia, communication and broadcasting with real-time implementation, applications and standardization and has gained international recognition. He has authored or edited four books, more than 100 technical papers, and 18 patents in these and closely related fields.

▶ Alexander S. Korotkov

Alexander S. Korotkov is a professor and head of the Integrated Electronics Department, St. Petersburg Polytechnic University, Russia. Professor Korotkov is a dedicated university lecturer, offering courses such as Circuits of Wireless Communications. On several occasions, he has been ranked one of the best professionals in evaluations carried out by St. Petersburg State Polytechnic University. He is a supervisor of the international master's degree program in microelectronics of telecommunications systems. His research interests include areas of circuits and systems for wireless communications, integrated circuit theory and design, and integrated circuit computer simulation. From 2003 to 2005, Professor Korotkov was an associate editor of *IEEE Transactions on Circuits and Systems, Pt II*. Currently, he is an associate editor of *Radio Electronics and Communications Systems (Ukraine) and St. Petersburg State Polytechnic University Journal, Journal of Computer Science, and Telecommunication and Control Systems* (Russia). Professor Korotkov has authored two books and published about 170 papers and reports.

5G wireless technology is developing at an explosive rate and is one of the biggest areas of research within academia and industry. With 2G, 3G and 4G, the peak service rate is the dominant metric that distinguishes these three generations. 5G will significantly increase the peak service rate but will also dramatically increase energy efficiency, frequency efficiency, spectral efficiency, and efficiency of other resources. It will dramatically increase flexibility, capacity, coverage, compatibility and convergence. In this way, it will satisfy the increasing demands of emerging big-data, cloud, machine-to-machine, and other applications. The successful development and deployment of 5G technologies will be challenging and will require huge effort from industry, academia, standardization organizations, and regulatory authorities.

This special issue deals with the application, technology, and standardization of 5G and aims to stimulate research and development of 5G by providing a unique forum for scientists, engineers, broadcasters, manufacturers, software developers, and other related professionals. The topics addressed in this special issue include system architecture, protocols, physical layer (downlink and uplink), air interface, cell acquisition, scheduling and rate adaption, access procedures, relaying, and spectrum allocation. The call-for-papers for this special issue attracted a number of excellent submissions. After two-round reviews, six papers were selected for publication. These papers are organized in two groups. The first group comprises three overview papers that outline technical aspects of 5G. The second group comprises three papers that provide new algorithms and theoretical analyses that can be used in the development of 5G wireless systems.

The first paper, "5G: Vision, Scenarios and Enabling Technologies," presents an excellent vision for 5G wireless communications systems, which are expected to be standardized around 2020. The paper states that service ubiquity is the key requirement of 5G from the end-user's prospective and is necessary to support a vast mesh of connections for human-to-human, human-to-machine, and machine-to-machine communications in an energy-efficient way. This paper discusses various technologies designed to improve radio link efficiency, expand operating bandwidth, and increase cell density. With these technologies, 5G systems can accommodate a massive volume of traffic, and this is fundamental for service ubiquity and supporting a massive number of connections, as outlined in the 5G vision of this paper. This paper also discusses the transition to intelligent cloud, in particular, cloud coordination of network access, which enables a flatter architecture.

The second paper, "Towards 5th Generation Wireless Communication Systems," discusses the targeted 5G system, including its driver, requirements, and candidate technologies that might help achieve its intended goals. Drawing on recent results obtained by the author's research team, the author discusses detection of and access to free spectrum over bands of a heterogeneous nature, extreme densification of

networks (mass base station deployments), extreme increase in the number of antennas in transmitter arrays and their interaction with a novel waveform, integration of both wireless and optical sides of telecom networks, and design of wireless networks from the perspective of complex systems science.

The third paper, "Signal Processing Techniques for 5G: An Overview," gives an overview of the main signal-processing techniques being developed for 5G wireless communications. At the beginning of this paper, the author reviews six orthogonal and non-orthogonal waveform-generation and modulation schemes: generalized frequency-division multiplexing (GFDM), filter bank multi-carrier (FBMC) transmission, universal filtered multicarrier (UFMC) transmission, bi-orthogonal frequency division multiplexing (BFDM), sparse code multiple access (SCMA) and non-orthogonal multiple access (NOMA). Then, the author discusses spatial signal processing algorithms and implementations of massive multiple-input multiple-output (massive MIMO), 3D beamforming and diversity, and multiplexing based on orbital angular momentum (OAM). The author also briefly reviews aspects of signal processing for other emerging techniques in 5G, such as millimeter wave, cloud radio access networks, full duplex mode, and digital RF processing.

The fourth paper, "Energy-Efficient Large-Scale Antenna Systems with Hybrid Digital-Analog Beamforming Structure," provides both theoretical analysis and simulations on the design of a large-scale antenna system (LSAS) with beamforming (BF), which is believed to significantly increase energy efficiency (EE) and spectral efficiency (EE) in a 5G wireless system. The paper investigates the optimal antenna configuration in an $N \times M$ hybrid BF structure, where N is the number of transceivers, and M is the number of antennas per transceiver. In such a structure, analog BF is introduced for each transceiver, and digital BF is introduced across N transceivers. The emphasis of this paper is EE-SE optimization when NM is fixed and when N and M are independent. The EE-SE relationship at "green" points is first investigated, then the effect of M on EE at a given SE is analyzed. In both cases, the authors show that there is an optimal M that provides the best EE for a given SE. The authors also discuss the optimal M when there is severe inter-user interference. These proposed analyses and results will be very useful in designing and deploying such LSAS for 5G.

The fifth paper, "An Optimal Lifetime Utility Routing for 5G and Energy-Harvesting Wireless Networks," looks at wireless sensor networking as a representative of all the different kinds of links involved in 5G. This paper also addresses energy efficiency. The authors propose an energy-harvest-aware route-selection method that incorporates harvest availability and energy storage capacity into routing decisions. In other words, the harvest-aware routing problem is formulated as a linear programming problem with a utility-based objective function that balances two conflicting routing objectives so that the proposed algorithm extends network lifetime. In addition, the authors investigate the effects of various network factors, such as topology, energy consumption rates, and prediction error, on energy savings.

The sixth paper, "Interference-Cancellation Scheme for Multilayer Cellular Systems," discusses interference cancellation, which is a challenging problem in a heterogeneous network that has coexisting multilayer cells, multiple standards and multiple application systems. First, an interference signal model that takes into account channel effect as well as time and frequency error is presented. An interference-cancellation scheme based on this model is then investigated. Following that, a method for compensating the timing and carrier frequency offset of an interference signal is presented. In the last step of processing, interference is mitigated by subtracting the estimation of interference signal. Computer simulation shows that the proposed interference-cancellation algorithm significantly improves performance in different channel conditions.

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

DOI: 10.3969/j. issn. 1673-5188. 2015. 01. 001
http://www.cnki.net/kcms/detail/34.1294.TN.20150203.1708.002.html, published online February 3, 2015

5G: Vision, Scenarios and Enabling Technologies

Yifei Yuan and Xiaowu Zhao
(ZTE Corporation, Shenzhen 518057, China)

Abstract

This paper presents the authors' vision for 5G wireless systems, which are expected to be standardized around 2020 (IMT-2020). In the future, ubiquitous service will be the key requirement from an end-user's prospective, and 5G networks will need to support a vast mesh of human-to-human, human-to-machine, and machine-to-machine connections. Moreover, 5G will need to support these connections in an energy-efficient manner. Various 5G enabling technologies have been extensively discussed. These technologies aim to increase radio link efficiency, expand operating bandwidths, and increase cell density. With these technologies, 5G systems can accommodate a massive volume of traffic and a massive number of connections, which is fundamental to providing ubiquitous services. Another aspect of 5G technology is the transition to an intelligent cloud that coordinates network access and enables flatter architecture.

Keywords

5G; IMT-2020; ultra-dense networks; massive MIMO; service ubiquity

1 Introduction

Cellular communications have gone through four decades of development, from 1G analog systems to 2G GSM and IS-95; 3G CDMA2000, UMTS and HSPA; and finally, 4G LTE. Worldwide, penetration of mobile phones is now more than 60%, even when counting under-developed countries where basic living conditions are still not guaranteed. The deployment of 3G and 4G has facilitated the proliferation of smart devices, which enable much easier access to electronic information and encourage interaction with remote computing systems, regardless of whether the user is stationary or on the move. This trend will continue with future 5G systems [1]. Human beings will rely more and more on cellular networks to acquire, disseminate, exchange, and manage information. 5G cellular services will be oriented towards user experience and satisfaction and will be supported by high-performance systems with capacity three orders of magnitude greater than 4G. This rich, vivid content will be instantly available anytime and anywhere.

User experience and satisfaction is the driving force of 5G. Wireless researchers and operators will come up with more innovative ways of converging devices, networks and services. As well as individuals, businesses, organizations and governments will also benefit greatly from 5G networks, which will be versatile, intelligent, and able to support a myriad of applications. 5G networks combine the advantages of cellular systems and wireless LANs. These two families of wireless technology have evolved along quite different paths since 3G, each being used for particular scenarios and unable to replace the other. With converged technologies, 5G networks will be comprehensive and able to penetrate more aspects of human life.

The so-called "mobile ICT era" implies ubiquitous mobility. Innovations in 5G will significantly increase efficiency in fields as diverse as education, healthcare, manufacturing, government, transportation, and finance. The boundary between the physical and digital worlds will be further blurred with 5G.

The rest of this paper is organized as follows. In section 2, we discuss 5G in terms of service ubiquity, vast interconnectedness, and energy efficiency. In section 3, we discuss some technical issues related to 5G, emphasizing the massive volume of data traffic and transition to intelligent cloud. In section 4, we draw some conclusions.

2 5G Vision

The story of high-end smart phones reveals that the need to provide better user experience is the main impetus for increasing network capacity. In 2007, there were no "killer applications" in the US. As a result, 3G cellular systems in the US were loaded to less than 20% of capacity, and operators began to question why they had outlaid such huge amounts of capital for 3G. However, in June 2007, Apple debuted its high-end phones, and this completely changed the situation [2]. The resulting jump in demand for wireless data throttled networks and forced operators to deploy more 3G equipment. Operators

were also pressed to accelerate 4G standardization. Increased system capacity fulfills the needs of users; the resultant appetite for fast data pushes the development of technology; and this further increases capacity. This cycle will continue with 5G, which will support a wide range of services and applications that will be vastly more interconnected and have greater effect on business and social life.

2.1 Service Ubiquity

The previous four generations of wireless systems were all designed to improve the peak rate and average throughput of a cell. The peak rate is the maximum data rate that can be achieved for a single user in the best propagation environment. Key performance indicators (KPIs) related to a technology give a good indication of the full potential of the technology but are often only significant for marketing. In reality, the user rarely fully experiences what the technology is touted to provide, even if the user is close to base stations in a lightly loaded network. In addition, the peak rate is usually only obtainable using the highest category of mobile terminal for the particular technology. Terminals that are not top-of-the-range will not be capable of delivering data at the peak rate advertised. Average cell throughput more accurately reflects what a user can typically expect from their services (**Fig. 1**). However, this is also not guaranteed for the majority of users. In many cases, a user at the cell edge experiences low data rate and a high rate of call dropping. This is a cause of customer complaints or even lawsuits.

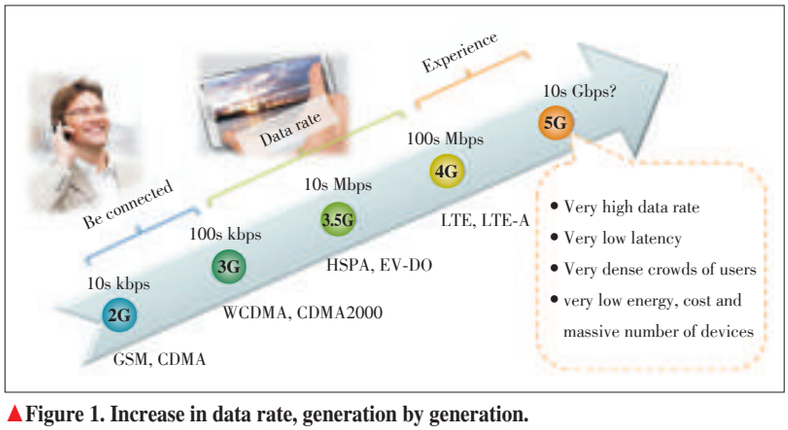
5G can significantly increase network capacity, peak data rate, number of connections, and traffic density within an area. It can also significantly reduce latency and provide highly accurate indoor positioning. Service ubiquity is a high priority for a system designed to be user-centric and tailored to different applications (**Table 1**). Ubiquity can be better measured by taking into account the resource usage patterns and traffic characteristics of wireless services to be provided. 5G will provide diverse services in areas such as office, social networking, and e-commerce, and online financial services. Peak data rate and average cell throughput are certainly not sufficient indicators of service ubiquity—KPIs such as data rate of worst 5% of users are more revealing in this respect. User experience depends on deployment scenario. In an office or dense residential area, the user is usually in a hotspot and can expect a much better experience than if they were in a wide-coverage scenario. The rate desired by someone participating in a big outdoor event in a regional area may not be possible in that area. 5G systems need to create an equally good experience regardless of location or scenario. This is what truly ubiquitous service implies.

▼**Table 1. User experience in different scenarios**

Applications	Very fast	Good service in a crowd	Ubiquitous	Around you	Super real-time and reliable
Virtual reality office	X				
Dense urban info. society	X	X	X	X	
Shopping mall		X	X		
Stadium		X			
Open air festival		X	X		
Real-time remote computing for mobile terminal				X	X
Tele-protection in smart grid			X		X
Traffic efficiency and safety				X	X
Massive deployment of sensors and actuators		X	X	X	

2.2 A Mesh of Connections

Cell phone penetration in developed countries is already 100%. As more and more people are expected to replace their old phones with powerful new terminals, such as tablets, the sheer number of cell phones is not expected to increase significantly. Nowadays, a smartphone is so omnipotent that you can talk, message, watch videos, listen to music, play games, chat, and surf the web all from one terminal. Many phones now support multiple standards for a single radio access, which make roaming much easier in many parts of the world. If we consider that cellphones are only used for human-to-human communication, cell phone penetration will remain flat or increase only slightly. However, 5G goes beyond traditional cellular services for personal use. A large chunk of traffic will derive from human-to-machine and machine-to-machine communication. The total number of devices that will need to be wirelessly connected in a few industries, e.g., retail, healthcare, manufacturing, transportation and agriculture, will be much higher than the human population. The total number of machine-to-machine connections will easily be counted in the hundreds of billions [3]. The requirements for the machine-to-machine communication in each industry will be drastically different: some industries will require a very high data rate; some require very short latency; some will require extremely high reli-



▲**Figure 1. Increase in data rate, generation by generation.**

ability. This creates great challenges for 5G networks. How can we serve vast meshes of human-to-human, human-to-machine, and machine-to-machine connections that far exceed those of 4G networks? Researchers need to come up with intelligent designs the ensure networks are robust and operate smoothly with massive numbers of connections.

Vast meshes of connections calls for 5G to be a conceptual change. In addition to interactive services such as conversation, gaming, video-conferencing and web surfing, 5G will have to provide many more automated services for machine-to-machine communication. Network design principles need to be rethought for 5G so that M2M communication is efficient and cost-effective. Compared with other wireless techniques and systems that used to be vertically integrated in each industry, 5G is better in terms of required performance and total cost of development and deployment. This would unleash the potential of economy-of-scale, which is often observed in the cellular industry, and would increase efficiency in many parts of society.

2.3 More Energy-Efficient Future

With the thousand-fold increase in capacity engendered by 5G networks, energy efficiency becomes a top priority. Capacity should not just be increased on average; throughput also needs to be significantly increased at the cell edge so that a user is guaranteed a superior wireless experience wherever they are. For service ubiquity and to support a massive number of connections, a 5G network infrastructure has to be very densely deployed. If efficiency remains at current levels, energy consumption will shoot up.

The idea of "going green" has taken root in many industries worldwide. The cellular industry is a major contributor of global CO_2 emission [4]. In the future, it will not be socially acceptable to chase ultra-fast network speed and excellent user experience at the expense of the environment. Therefore, researchers need to devise smarter ways to present energy information to users, reduce harmful interference caused by aimless transmission, and conform to Moore's Law by further shrinking circuits and reducing power consumption. Increasing energy efficiency involves more cost-effective site planning, construction, and maintenance. These traditionally account for a large proportion of energy consumption within a cellular system.

5G terminals also need to be energy-efficient, and this requires cooperation between system designers and device manufacturers. Moore's Law also applies to terminals as well as networks— the size of circuits in a terminal will continue to reduce, and more complicated data processing will occur with less power and smaller die area.

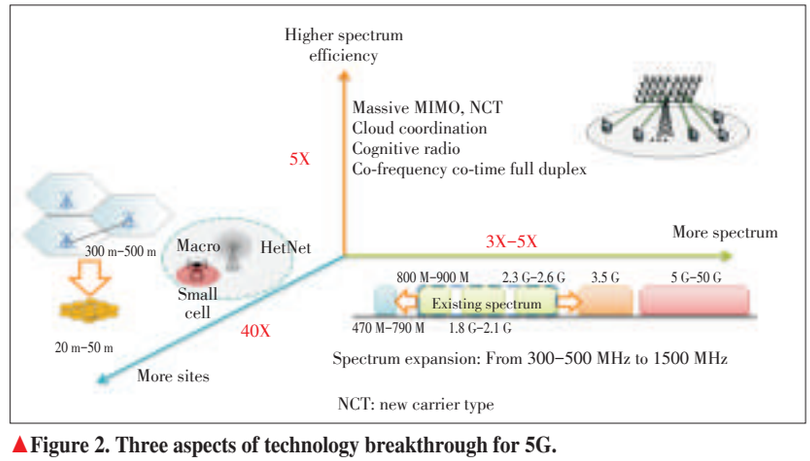
3 Enabling Technologies

From 1G to 4G, each generation is defined by a standout technology that represents the most important advancement in that generation. Traditionally, each cellular generation is distinguished by its unique multi-access method. For example, the distinguishing feature of 1G was frequency-division multiple access (FDMA) and the reliance on analog-domain signal processing and communication. The distinguishing feature of 2G was time-division multiple access (TDMA), and GSM is a good example of where digital signal processing and communications were first widely used in the cellular world. The distinguishing feature of 3G is code-division multiple access (CDMA), which enables radio resources to be shared between multiple users. The introduction of capacity-approaching Turbo codes significantly increases spectral efficiency. The distinguishing feature of 4G is orthogonal frequency-division multiple access (OFDMA) which, when combined with spatial multiplexing schemes such as MIMO, drastically increase system capacity without unreasonably increasing complexity at receivers.

5G will have diverse requirements and applications and is unlikely to be defined by a single radio-access scheme or dominant technology. However, there are a few noticeable technological trends that could underscore 5G networks. These technological trends will intersect with each other to construct the 5G system as a whole and fulfill the vision outlined earlier.

As with the previous four generations of cellular networks, capacity will continue to be very important technical goal of 5G. Capacity is a "hard" performance criteria and the basis of service ubiquity. There are three issues related to traffic volume that need to be considered in order to achieve a thousand-fold increase in capacity (**Fig. 2**). These aspects issues are: improving radio link efficiency, expanding and managing spectrum, and making cell sites denser.

Another important trend in future wireless networking is the transition to intelligent cloud. Cloud-like concepts are seen in various branches of information technology, and 5G network architecture is no exception. Many kinds of network intelligence will be re-formulated and integrated into the cloud structure. With the cloud, human-to-human, human-to-machine, and ma-



▲**Figure 2. Three aspects of technology breakthrough for 5G.**

chine-to-machine communication will all be seamless. The huge digital ecosystem created by 5G will allow any person or machine to be a content generator or digital library that can potentially accelerate the growth of data traffic in future networks.

3.1 Improved Efficiency of Radio Links

Since the advent of 3G, more technologies related to multiple antennas have been adopted to increase system capacity. In a 3G system, multiple transmitting antennas are primarily used for beamforming, which can improve the signal-to-noise ratio (SNR) of a link. Orthogonal-frequency-division multiplex (OFDM) greatly reduces receiver complexity in a 4G MIMO system, especially in a multipath propagation environment. OFDM has given rise to a wave of spatial multiplexing schemes that have now been standardized. With spatial multiplexing, a radio link can support multilayer transmission, which helps double or even triple the peak data rate of a 4G system compared to a 3G system. The advancement of MIMO in LTE [5] allows more flexible multiuser MIMO, a technology that increases the sum capacity of users who are simultaneously scheduled.

With 5G, there will be order of magnitude increase in the number of antennas at the base station. Dozens or even hundreds of antennas are possible and will form a so-called massive MIMO system [6]. The freedom offered by massive MIMO in the spatial domain is so huge that the effect of thermal noise will become negligible, and system performance will, in theory, only be limited by the pilot pollution. This huge degree of freedom can support high-order multiuser communication, where a large number of users share the same time and frequency resources and do not significantly interfere with each other. Improved SNR and high-order multiuser transmission can provide a several-fold increase in spectral efficiency in the system, although the gain may not be exactly proportional to the number of antennas.

With massive MIMO, radiated energy is directed more towards intended users rather than being radiated in all directions. Link budget can also be dramatically improved. In other words, a transmitter uses less power while providing higher system throughput and wider coverage. In addition, transmit power can be distributed over many antennas in massive MIMO, and the power amplifier of each antenna operates at linear region. This facilitates high-order modulation and coding without using the need for expensive power amplifiers and pre-distortion algorithms. Therefore, massive MIMO will be critical in reducing energy consumption of 5G systems and achieving the goal of green communication.

Turbo codes in 3G and 4G have already pushed the spectral efficiency of a single-antenna channel very close to the Shannon limit; however, there is still some room for more research on new coding and modulation schemes. Channel coding in 3G and 4G is mainly aimed at approaching the Shannon limit for traffic channels on the condition that code blocks are enough long and the channel is only affected by additive white Gaussian noise (AWGN). In 5G, more flavors will be added to the design of the channel codes and modulation schemes. For vast meshes of connections, the number of links that access to the network is huge, yet each link carries only small amount of data. Because short blocks of data prevail, the channel-coding community is searching for powerful codes that can approach the channel capacity even when the code block is short, e.g., less than 100 bits [7]. Cellular communication is often corrupted by fast fading, either at frequency domain or time domain. Link adaptation is effective for handling fast fading and helps transmitters choose modulation/coding schemes to match the current channel condition. The link adaptation schemes in 3G and 4G can work but often fail to keep up with the channel variation when users are moving fast. New link-adaptation methods in 5G enable the transmission scheme to be quickly adjusted to suit the channel characteristics. This reduces resource waste and improves link efficiency in fading environment.

Faster than Nyquist (FTN) is another link-level coding scheme that has recently gained attention in 5G. Instead of using traditional QAM to map coded bits to complex symbols, coded bits in FTN are shifted and overlapped in time. This superimposition forms a real-valued convolutional encoder. Relying on the sequence detection at the receiver, coded bits can be detected with low bit-error rate. The amplitude distribution of FTN signals is closer to that of Gaussian signals than that of QAM signals. In an environment with high signal-to-noise-ratio (SNR), FTN more closely approaches the Shannon limit. An FTN signal has lower peak-to-average power ratio (PAPR) than a QAM signal, which allows power amplifiers to operate more efficiently.

Multiuser communication has been around since 3G. Standards specifications define some physical control signaling and reference signal structure (format) to better support multiuser features; however, many multiuser schemes are implementation-oriented. One example of this is linear superposition of signals of different users and then using receiver-side interference cancellation to pick up each user's signal (data). Interference cancellation at receivers is usually implementation-specific. In 5G, more sophisticated superposition is expected for multiuser support. Such superposition includes code-domain superposition, which more closely approaches the sum-rate of the broadcast channel/multi-access channel or reduces the need for resource scheduling and is less dependent on interference cancellation at receivers. These superposition codes form new family of new coding and modulation techniques that includes bit-division multiplexing [9].

At the system level, network coding should increase the total throughput of a system when it has multihop transmissions. Relay node deployment is an example of multi-hop systems, especially when the backhaul link (base station to relay node) and access link (relay node to user terminal) share the same

frequency. So far, network coding has never gone beyond academia. This situation will change with 5G, and we may see the widespread adoption of network coding in various standards.

With an increase in processing power at the receivers, interference cancellation can be brought to a higher level so that transmission and reception can occur at the same time in the same frequency [10] with little spatial isolation in-between. With analog- and digital-domain interference cancellation, full duplex communication is possible in 5G. In theory, link efficiency will be doubled if all co-existence issues can be solved.

In applications such as machine-to-machine communication for manufacturing or vehicle-to-vehicle communication for intelligent transport systems, short latency and high reliability are critical requirements. 3G and 4G systems were designed primarily for human-to-human communication, and their physical-layer structures and higher-layer protocols do not meet these stringent latency and reliability requirements. Physical structures with shorter transmission time interval (TTI) are needed in 5G so that the number of residual errors is made extremely low within a very short timeframe. New coding and modulation schemes will help achieve this goal as well.

OFDM, the widely used waveform in 4G, enables simple receiver implementation in MIMO systems. However, OFDM signals tend to have significant out-of-band emissions, and this requires precise synchronization and orthogonal resources. Filter-bank multi-carrier (FBMC) is a promising waveform technology that reduces out-of-band emissions and lowers the requirement for synchronization. The basic idea of FBMC is to replace the rectangular window in OFDM with a bank of filters. In MIMO systems, FBMC requires some extra signal processing in channel estimation and filter algorithms in MIMO systems, FBMC is very suitable for dynamic spectrum allocation scenarios where the low out-of-band emission makes the systems more compatible with various band combinations.

Radio link efficiency can also be improved by using a software-defined air interface so that the system can support multiple radio access technologies (multi-RAT) in a multi-spectrum deployment. Such operational flexibility can improve the utilization of radio resources and result in better 5G performance.

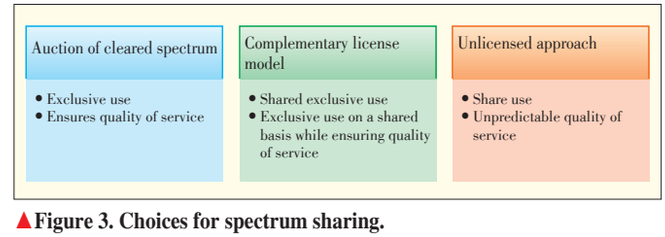
3.2 Spectrum Expansion and Management

Data rate is the product of spectral efficiency and occupied bandwidth. As cellular communication evolves from 1G to 4G, system bandwidth continuously increases, as does the bandwidth each user can occupy. Over the past few decades, link-level spectral efficiency has increased significantly; however, it is wideband operation that really provides order-of-magnitude gain in system capacity and peak rate. Each user now has access to wide bandwidth because of the fast development of radio frequency components and digital communication processors. Nowadays, terminal chipsets are so powerful that they support bandwidth well beyond 20 MHz with 64 QAM and multilayer MIMO transmission. User experience is greatly improved by this.

Cellular bands are traditionally allocated below 3 GHz, where scattering, reflection and diffraction help the electromagnetic (EM) wave uniformly shine on the targeted terminals. This can fill coverage holes when shadow fading is present. Below 3 GHz, an EM wave also usually has less penetration loss when passing through walls and windows. So EM signals can penetrate deeper into buildings and reach indoor users. Such superb link budget in this lower spectrum enables an operator to deploy fewer base stations to cover an area and reduce capex and opex. This is why in many countries 700 MHz, which was formerly used for TV broadcasting, is now being re-farmed for cellular use. The propagation characteristics at 700 MHz are very favorable for coverage.

Wireless spectrum is a scarce resource and is regulated for satellite, military, scientific, and radio and TV uses, not just for cellular communication. Over the past four generations of cellular development, the spectrum below 3 GHz regulated for cellular services has been used for terrestrial communication. Unless 2G is phased out soon and its occupied bands are re-farmed, there is little room for spectrum expansion below 3 GHz. This problem is compounded by the fact that spectrum allocations are very fragmented in lower bands. It is extremely difficult to find a wide and contiguous band at low frequencies, which means a high data rate is very unlikely. Propagation characteristics between 3 GHz and 6 GHz are slightly different from those below 3 GHz. Between 3 GHz and 6 GHz, there is less scattering and path loss tends to be a little higher, although the fundamental propagation mechanism remains the same as that in the lower bands. A lot of un-licensed spectrum is regulated in the 3–6 GHz range, e.g., Wi-Fi which provides cost-effective solutions for local wireless access. 5G will not only operate in the licensed spectrum; it will also be compatible with technologies that are tailored to unlicensed spectrum. In this way, spectrum can be shared, and network potential can be further unleashed. 3GPP [11] has started to look into this opportunity in their recently proposed study item [11]. **Fig. 3** shows more choices for spectrum sharing.

Spectrum shortage can be largely solved by using high bands, e.g., above 6 GHz or even millimeter wavelength. Traditionally, high-frequency communication has been limited to point-to-point communication, e.g., between macro base stations. There are several reasons for this. First, base station antennas are usually well above the building's roofline, which ensures line of sight (LOS) and compensates for the heavier path



▲**Figure 3. Choices for spectrum sharing.**

loss of free-space propagation at higher bands. Second, the transmitter and receiver are all stationary so that antennas with sharp directivity can be used to further improve the channel SNR. Third, the cost of implementing high-frequency RF and chipsets is not a major concern given the small number of wireless backhauls needed and the cost of devices compared to the total cost of a macro base station.

There is a strong motivation to use high-frequency bands for cellular access in 5G. The relatively abundant spectrum above 6 GHz makes it possible to allocate a contiguous band with huge bandwidth, e.g., 500 MHz, so that wireless subscribers have a super-fast experience. Indoor hotspots, where access points are deployed indoors, are especially suitable for high-frequency communication because short distance does not require high-powered transmission. Greater penetration loss helps isolate each access point; therefore, less inter-access-point interference is expected.

High-frequency bands might also be used for wide-area communication. There are many challenges: heavy path loss needs to be addressed so that users deep within the building or far from the base station are guaranteed coverage. In a wide-area scenario, the user is usually moving. Because of the short wavelength, moderate traveling speed can result in very high Doppler Effect. Mobility management therefore becomes more difficult.

One benefit of using a shorter wavelength is the increased effective aperture of both the transmitter and receiver antennas when the antenna size is kept the same. Massive MIMO operating in high frequency bands can be very compact and highly integrated for lower cost. This significantly increases the opportunities to deploy massive MIMO, which may be deployed for access points.

For high-frequency communication, the most challenging issue may be devices. Despite their low cost and ease of integration, traditional silicon-based chips may not be capable of providing the required processing speed, noise level, or energy efficiency. New materials such as Gallium-silicon are being studied with the prospect of delivering good performance for lower cost.

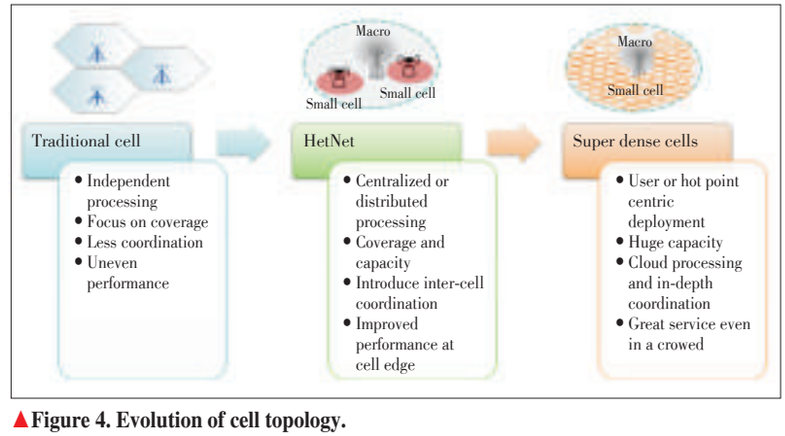
Spectrum management is not only a technological issue; it is also a political issue. Regulating spectrum involves balancing the interests of various parties, some of which may have historical rivalries. That is why the ITU is cautious about discussing and allocating spectrum. Right now, the ITU is still working on WRC-15 and is focused on bands under 6 GHz. Discussion on spectrum allocation of bands above 6 GHz for cellular use will not start until 2018. Consequently, work on the full standardization of 5G high-frequency communication will not kick off until after 2018.

3.3 Cell Site Densification

For hotspot scenarios, such as offices and dense urban apartments, cell density needs to be drastically increased to accommodate huge volumes of traffic within a small geographical area. In the first three generations of cellular networks, cell layouts had rather homogenous topologies, e.g, system base stations had the same configurations for transmit power and antenna gain. Base stations were more or less equally distanced. In 4G, heterogeneous networks (HetNets) came into the picture. Low-power nodes (LPN), such as pico, femto and relay [12] nodes, combined with macro base stations to provide high throughput via cell-splitting gain (**Fig. 4**). In a HetNets scenario, cell splitting can be viewed as offloading traffic from macro base stations to low-power nodes, which help fill the coverage holes in a homogeneous layout so coverage approaches 100%. Users who are not in hotspots are still attached to macro base stations. The effect on mobility management is minimal because users in hotspots are not expected to be fast-moving.

In 5G, many more LPNs are deployed within a macro area, resulting in less than 20 meters inter-node distance. With such closeness, even the building penetration may not be enough to reduce interference from neighboring LPNs. With an outdoor gathering, more inter-site interference is expected. Advanced interference coordination and cooperative transmission schemes are crucial to ensure reasonably good spectral efficiency per LPN and good user experience, regardless of whether the user is at the center or edge of the cell.

As the cell site density is increased, it is more difficult to choose sites that have wired backhaul connected to the networks. In cities with historical sites, preserving old streets and structures is a high priority, and it is almost impossible to break new ground to lay down telecommunications cables. For systems with a large number of LPNs, the cost of installing and maintaining the cable backhaul is daunting. Therefore, there is strong motivation for wireless backhaul. Conventional wireless backhaul uses proprietary technologies and often operates in non-cellular bands at high frequency. In 5G, wireless backhaul will be based on standard air interface and work for cellular bands that may not necessarily be high-frequency. Standardization can help achieve economy of scale, and low-to-medium frequency bands enable wireless backhaul to work in diverse en-



▲**Figure 4. Evolution of cell topology.**

vironments, including non-line-of-sight (NLOS) environments. Consequently, capex and opex can be minimized, even with super-dense cell deployment.

Device-to-device (D2D) communication is a special case of network densification [13]. A D2D-capable device can act as a low-power node for unicasting, multicasting or broadcasting traffic directly to the user without being routed through the network. D2D is especially useful for proximity services where users in the vicinity share and exchange local information. D2D communication helps increase the density of low-power nodes with wireless backhaul. In 5G networks, D2D is expected to increase system capacity, particularly in dense urban environments and for big outdoor events.

3.4 Cloud Coordination of Network Access

By the time 5G systems are deployed, many 3G and 4G networks (and even some 2G networks) will still be in use. Within each generation of network, the allocated spectra may be different, depending on operator, country, or year of deployment. Radio resources may also include lightly licensed and unlicensed spectrum where technologies of wireless LANs dominate. In this sense, 5G networks will be a mix of new network components and existing systems and assets as well as radio access technologies (RATs) of a non-cellular kind. Cloud architecture will coordinate different types of radio and network resources and manage inter-RAT and inter-frequency radio access in a seamless and transparent manner. Multi-RAT convergence is possible in 5G with a unified cloud architecture (**Fig. 5**).

Cell densification worsens interference between cells. Potential solutions include advanced interference coordination or cooperative transmission. The integrated nature of cloud means that more dynamic interference coordination is necessary to significantly improve network performance at the cell edge. In addition, a significant amount of capex and opex can be saved by using cloud because a lot of the DSP can be centralized and implemented in common-purpose digital processors. Less effort in terms of customization and higher utilization of digital processors saves cost. Energy can also be saved because many processors can share the same air conditioner.

3.5 Flatter Architecture

To increase commercial viability and versatility, flatter architecture has been recommended for 5G. Virtualization and software-defined networking (SDN) are promising techniques for reducing the complexity of 5G networks and optimizing the system performance. In 5G networks, the control plane will be centralized and the user plane will be streamlined. With a flatter architecture, an operator can focus on value-added features or services. Flatter architecture is helps reduce latency; low latency is critical in machine-to-machine and vehicle-to-vehicle communication. With flat architecture, services and networks can be deeply converged.

SDN encourages innovation in network protocols and rapid deployment of networks by operators. With SDN, traffic flow can be controlled in more flexible way. Centralized control also enables more coordinated management of traffic flow and network resources. The OpenFlow protocol currently being discussed still has issues in mobile scenarios.

An indispensable aspect of flat architecture is network function virtualization (NFV). The key idea behind NFV is decoupling node functions from node hardware. Standard high-performance hardware replaces special equipment that has been cus-

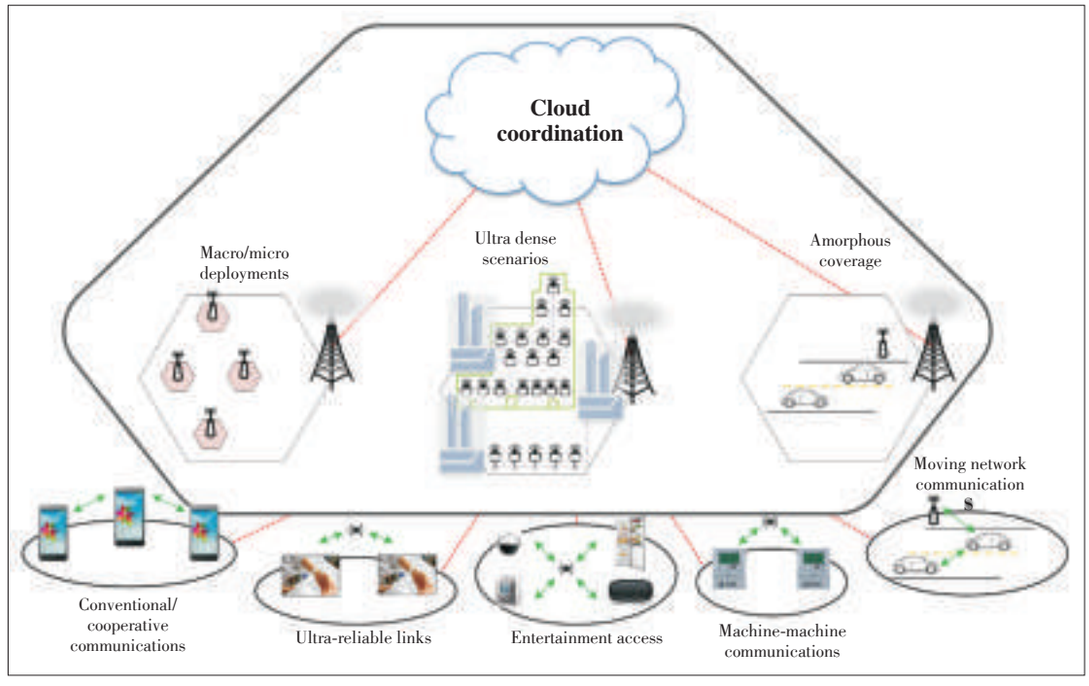


Figure 5. Cloud coordination for network access.

tomized for each node. This can simplify the design of the hardware platform and reduce network cost. Not only is the deployment more flexible, but NFV also encourages openness and innovation in equipment manufacturing. Nevertheless, issues such as software portability, interoperability, and stability need to be addressed before NFV can be widely used.

With flatter architecture, a content-delivery network can have advanced technology such as a smart content router, which redirects user requests to the nearest web cache. This significantly improves responsiveness and user experience, especially for data-heavy applications such as high-definition video.

4 Conclusions

Service ubiquity, a massive number of connections, and energy efficiency are the three key requirements for 5G networks. 5G technologies need to support massive traffic volume by helping increase radio link efficiency, expand spectrum to high-frequency bands, and densify cells. Networks should also be transitioned to cloud architecture, which is flatter and can coordinate radio resources of multi-RATs and intercell interference arising from dense network deployment.

Acknowledgement

The authors would like to thank Xinhui Wang and Longming Zhu at ZTE for the contribution to an earlier white paper which provides some of the foundations of this paper.

References

[1] ITU. (2012). IMT for 2020 and beyond [Online]. Available: http://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020
[2] F. Vogelstein, "The untold story: how the iPhone blew up the wireless industry," *Wired Magazine*, issue 16.02, January 2008.
[3] P. Popovski, V. Braun, H.-P. Mayer, *et al.*, "Scenarios, requirements and KPIs for 5G mobile and wireless system," METIS, ICT-317669-METIS/D1.1, April 2013.
[4] The Climate Group, "Smart 2020: Enabling the low-carbon economy in the information age," 2008.
[5] Y. Yuan, *LTE/LTE-Advanced Key Technologies and System Performance*, Beijing, China: Posts & Telecom Press, Jun. 2013.
[6] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010. doi: 10.1109/TWC.2010.092810.091092.
[7] M. Dohler, R. W. Heath, A. Lozano, *et al.*, "Is the PHY layer dead," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 159–161, Apr. 2011. doi: 10.1109/MCOM.2011.5741160.
[8] J. B. Anderson, F. Rusek, and V. Owall, "Fast-than-Nyquist signaling," *Proceedings of the IEEE*, vol. 101, no. 8, pp. 1817–1830, Aug. 2013. doi: 10.1109/JPROC.2012.2233451.
[9] H. Jin, K. Peng, and J. Song, "Bit division multiplexing for broadcasting," *IEEE Transactions on Broadcasting*, vol. 59, no. 3, pp. 539–547, Sept. 2013. doi: 10.1109/TBC.2013.2254269.
[10] J. Il Choi, S. Hong, M. Jain, *et al.*, "Beyond full duplex wireless," in *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, USA, pp. 40–44. doi: 10.1109/ACSSC.2012.6488954.
[11] Ericsson, "Study on Licensed-assisted access using LTE," 3GPP RP-141397 RAN#65, Sept. 2014.
[12] Y. Yuan, *LTE-Advanced Relay Technology and Standardization*, New York City, USA: Springer, Jul. 2012.
[13] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution in 5G," *IEEE Communications Magazine*, pp. 82–89, Feb. 2014. doi: 10.1109/MCOM.2014.6736747.

Manuscript received: 2014-09-06

Biographies

Yifei Yuan (yuan.yifei@zte.com.cn) received his BS and MS degrees from Tsinghua University, China. He received his PhD from Carnegie Mellon University, USA. From 2000 to 2008, he worked with Alcatel-Lucent on 3G and 4G key technologies. Since 2008, he has worked for ZTE researching 5G technologies and standards for LTE-Advanced physical layer. His research interests include MIMO, iterative codes, and resource scheduling. He was admitted to the Thousand Talent Plan Program of China in 2010. He has written a book on LTE-A relay and a book on LTE-Advanced key technologies and system performance. He has had more than 30 patents approved.

Xiaowu Zhao (zhao.xiaowu@zte.com.cn) received his PhD degree from the Institute of Software, Chinese Academy of Science, in 2001. He has participated in the Third Generation Partnership Project 2 (3GPP2) since 2001. He was the 3GPP2 Technical Specification Group-Access Network Interface (TSG-A) chair from 2009 to 2012 and System Aspect and Core Network (TSG-SX) chair from 2013 to the present. In recent year, he has also participated in the 3GPP RAN plenary meeting. He has submitted hundreds of contributions that have been accepted by 3GPP2. His current research interests include beyond LTE-Advanced and 5G technology and taking charge of the whole industry standardization of ZTE Corporation.

DOI: 10.3969/j. issn. 1673-5188. 2015. 01. 002
http://www.cnki.net/kcms/detail/34.1294.TN.20150210.1339.001.html, published online February 10, 2015

Towards 5th Generation Wireless Communication Systems

Nicola Marchetti
(CTVR/Telecommunications Research Centre, Trinity College Dublin, Ireland)

Abstract

This paper introduces the general landscape of next-generation wireless communication systems (5G), including the impetus and requirements of 5G and the candidate technologies that might help 5G achieve its goals. The following areas, which the author considers particularly relevant, are discussed: detection of and access to free spectrum over bands of a heterogeneous nature, extreme densification of networks (massive base station deployments), extreme increase in the number of antennas in base station arrays and their interaction with a novel waveform, integration of both wireless and optical sides of telecom networks, and study of wireless networks from the perspective of complex systems science. The author discusses recent research conducted by his team in each of these research areas.

Keywords

5G; spectrum; cell densification; efficiency; optical wireless integration

1 Introduction

Social development will lead to changes in the way communication systems are used. Increasingly, on-demand information and entertainment will be delivered over mobile and wireless communication systems. This is expected to lead to huge growth in global demand for mobile broadband data services, and several indications point to the fact that such growth will continue [1], [2]. Human-centric communication, predominant in today's communication scenarios, will be complemented by machine-to-machine (M2M) communication, which is expected to increase massively. Some analysts have estimated there will be 50 billion connected devices by 2020 [3]. The coexistence of human-centric and machine-type applications will lead to more diverse communication characteristics.

Different applications place diverse requirements on mobile and wireless communication systems. Fifth-generation (5G) technology will have to satisfy these requirements [4], which include stringent latency and reliability, a wide range of data rates, network scalability and flexibility, very low complexity, and very long battery life. One of the main challenges is satisfying these requirements while at the same time minimizing costs.

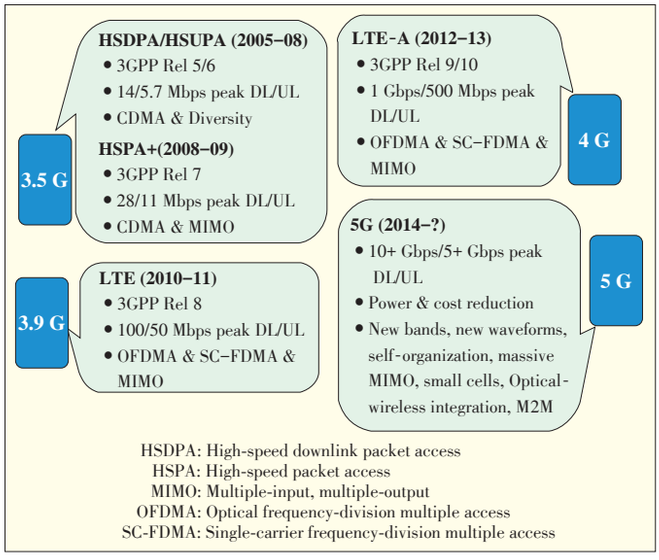
Fresh quantitative evidence shows that the wireless data explosion is real and will continue. A recent Visual Networking Index (VNI) report stated that an incremental approach to 5G will not be enough to meet the demands on networks in the coming years [5]. Indeed, it is likely that 5G will have to be a paradigm shift that includes (among other things) very high carrier frequencies with large bandwidths, extreme base station and device densities, and massive numbers of antennas. The motivation behind chasing spectrum in high frequencies is the scarcity of RF spectra allocated to cellular communication. Cellular frequencies use UHF bands for cellular phones, but spectra in these frequencies have been used heavily, and it is difficult for operators to acquire more. Another challenge is the high energy consumption of advanced wireless technologies. Cellular operators have reported that the energy consumption of base stations (BS) accounts for more than 70% of their electricity bills [6].

Unlike previous generations of technology, 5G will also be highly integrative and will tie any new air interface and spectrum to LTE and Wi-Fi. This will enable universal high-rate coverage and seamless user experience [7].

Academia is involved in large collaborative projects, such as METIS [4] and 5GNOW [8], and industry is driving preliminary 5G standardisation. To support these activities, the 5G In-

This work is supported in part by Science Foundation Ireland through CTVR CSET grant 10/CE/I1853, and in part by the European Commission's FP7 project ADEL, under grant agreement ICT-619647.

frastructure Public-Private Partnership (5G PPP) was recently established in Europe [7], [9]. **Fig. 1** shows the recent evolution of wireless communication standards. 5G includes elements that are disruptive compared to the past.



▲**Figure 1. Characteristics of recent generations of wireless systems.**

2 Drivers and Requirements

Although 5G will have diverse requirements, not all of these will need to be satisfied at once because different applications make different demands on system performance. For example, very-high-rate applications, such as streaming HD video, may have less stringent latency and reliability requirements compared to driverless cars or public safety applications, where latency and reliability are paramount but lower data rates can be tolerated [7]. 5G aims to connect the whole world and enable seamless, ubiquitous communication between anybody and anything, regardless of time, location, device, service, or network [10]. 5G will provide the foundational infrastructure for building smart cities, which will push mobile network performance and capabilities to their extremes [11]. The main drivers of 5G are Internet of Things (IoT), gigabit wireless connectivity, and tactile Internet. The main challenge for IoT is scalability— fitting more than 100,000 M2M nodes into a cell at low cost and ensuring long lifetime. In terms of gigabit connectivity, users might want to download streaming 3D content at a rate on the order of 100 Mbps. Such rates of connectivity are also expected at large gatherings where there are many (possibly interactively) connected devices. Tactile Internet is a part of many real-time applications that require extremely low latency. Inspired by the sense of touch, which has latency on the order of 1 ms, 5G will be used in steering and control scenarios and will be a disruptive change in communications [12].

As we move to 5G, cost and energy consumption should not increase on a per-link basis. Because the per-link data rates will increase about a hundred-fold, the Joules per bit and cost per bit will need to decrease by at least a hundred-fold. We should therefore advocate technological solutions that are low-cost and enable power scaling. Because of the increased BS densities and bandwidth in 5G, cost is a major consideration when backhauling from the network edges to the core [7].

3 Candidate Technology Components

Many believe that the required increase in data rate will be achieved, for the most part, through [7]:

1) extreme network densification, to improve the area spectral efficiency (more nodes per unit area and Hertz)
2) increased bandwidth, mainly by moving towards millimeter-wave spectrum and making better use of unlicensed spectrum at 5 GHz (more Hz)
3) increased spectral efficiency, mainly through advances in multiple-input multiple-output (MIMO) techniques (more bits/s/Hz per node).

The motivation for network densification is that making cells smaller and denser is a straightforward and effective way of increasing network capacity. Networks are now rapidly evolving to include nested small cells, such as picocells, which have a range of less than 100 m; femtocells, which have a range similar to that of Wi-Fi; and distributed antenna systems [13]–[15].

Challenges that need to be addressed in 5G systems are fragmented spectrum and spectrum agility. It is unlikely that these challenges can be overcome using orthogonal frequency-division multiplexing (OFDM), and new waveforms that are more flexible and robust are required. In the 5GNOW project, several alternative candidate waveforms have been proposed [16]. These waveforms include filter bank multicarrier (FBMC) [17]. Recent studies suggest that millimeter-wave frequencies could be used to augment the current saturated radio spectrum bands for wireless communication [18]. By increasing the RF channel bandwidth for mobile radio channels, data capacity can be significantly increased, and latency for digital traffic can be significantly decreased. This enables much better Internet-based access and applications that require minimal latency. Because of their much shorter wavelength, millimeter-wave frequencies may exploit new spatial processing techniques, such as massive MIMO [19], which increase spectral efficiency.

M2M communication in 5G involves satisfying three fundamentally different requirements associated with different classes of low-data-rate services. These requirements are: 1) support for a massive number of low-rate devices, 2) minimal data rate in almost all circumstances, and 3) very-low-latency data transfer [20]. The questions for industry are: Should we design the same network for both human and machine communication? Should we design a new network for machines? Or should we design a hybrid network [21]?

It is unlikely that one standard and one model of network de-

ployment will be fit for all future scenarios. Mobile networks and equipment need to be flexible and able to be optimized for different scenarios, which may be dynamic in terms of space and time. The requisite for flexibility will significantly affect the design of new network architectures. In [22], the authors provide this flexibility by leveraging cloud technology to operate a radio access network (RAN). Radio access infrastructures with cloud architecture will provide on-demand processing, storage, and network capacity wherever needed. Software-defined air interface technologies will be seamlessly integrated into 5G wireless access network architectures. This will enable RAN sites to move towards a "hyper transceiver" approach to mobile access, which will help with joint-layer optimization of radio resources [11].

The ultimate goal of communication networks is to provide access to information whenever, wherever, and in whatever format we need it. Wireless and optical technologies and access networks are key to achieving this goal and can be thought of as complementary. Optical fiber cannot be used everywhere, but where it is used, it provides a huge amount of bandwidth. Wireless access networks, on the other hand, can be used almost everywhere but are highly constrained in terms of transmission bandwidth and impairments. Future broadband access networks must leverage both wireless and optical technologies and converge them seamlessly [23].

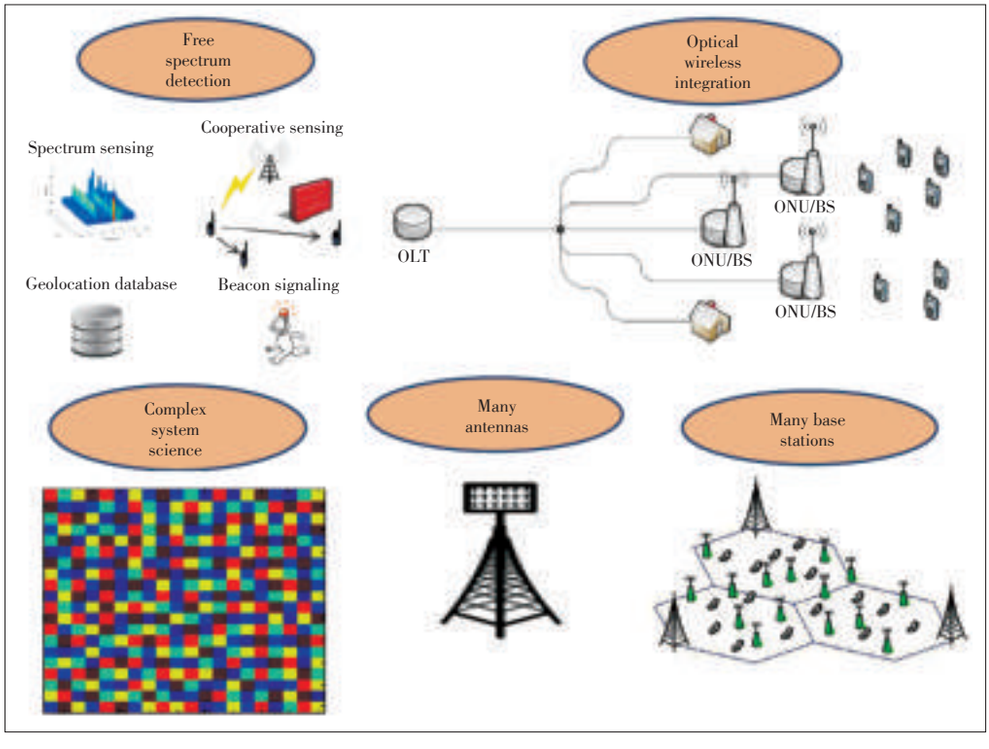
Fig. 2 shows some of the components that may be prominent in building up and consolidating 5G. The following sections of this paper will present some of the highlights of work done in these areas by the CTVR/Telecommunications Research Centre, Trinity College Dublin.

4 Spectrum

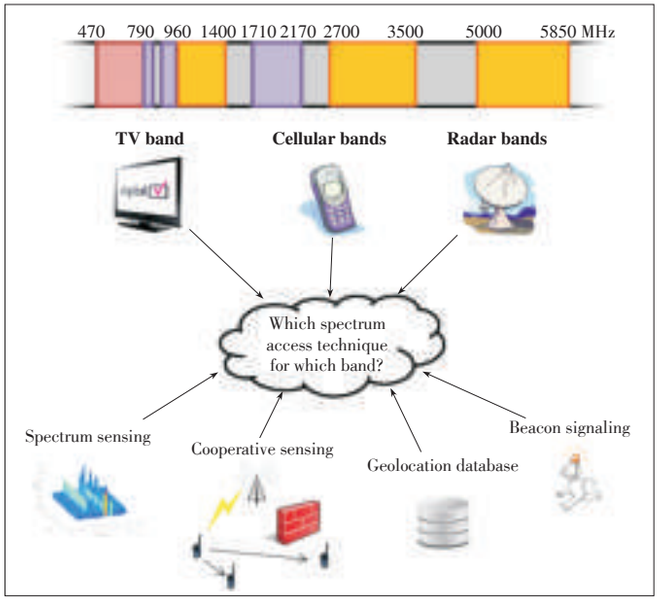
5G systems are expected to provide gigabit-per-second data rates anytime, anywhere. This can only be realized with much more spectrum than is currently available to International Mobile Telecommunications (IMT) systems through the International Telecommunication Union (ITU) process. All spectra currently available to cellular mobile systems, including IMT systems, are concentrated in the bands below 6 GHz because of favourable propagation in such bands. As a result, these bands have become extremely crowded, and there is little prospect of large chunks of new spectra below 6 GHz becoming available for IMT systems [24]. As a result, regulators have considered opportunistic spectrum access (OSA) for a number of different spectrum bands. In [25], we discuss and qualitatively evaluate techniques used to discover spectrum opportunities (white spaces) in radar, TV, and cellular bands (**Fig. 3**). These techniques include spectrum sensing, cooperative spectrum sensing, geolocation databases, and the use of beacons.

Each of the three bands in Fig. 3 calls for a different set of spectrum access techniques. Geolocational databases are well suited to TV bands, and a database-assisted spectrum-sensing mechanism may be the best solution to exploit spectrum holes in radar bands. The unpredictability of cellular systems calls for beacon signalling, which is a more coordinated spectrum access approach that can be implemented using already-established cellular infrastructure and spare bits of its logical channels [25].

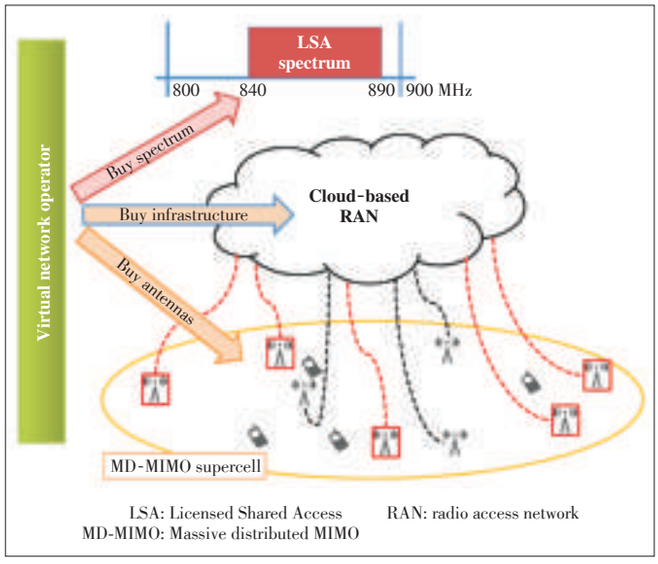
Another potential way of making large amounts of spectrum available to future IMT systems is licensed shared access (LSA). With LSA, underutilized non-IMT spectrum can be integrated with other IMT spectrum in a licensed, pre-determined way upon mutual agreement between licensees [24]. In [26], we propose a cloud-RAN Massive Distributed MIMO (MD MIMO) platform as architecture to take advantage of LSA. This architecture is worth exploring in the context of LSA because our architecture has similar principles in terms of resource use. LSA involves sharing spectrum from a pool of virtual spectrum resources, and cloud RAN provides a way of managing the pool of vir-



▲**Figure 2. 5G candidate technology components.**



▲Figure 3. Which access technique for which band?



▲Figure 4. Cloud-based MD-MIMO RAN for M_{max} = 8 and W_{max} = 100 Mhz. The virtual network operator uses a subset of the available spectrum (W = 50 MHz) and antennas (M = 5) to maximise the number of bits transmitted to the K = 4 users per cost unit. The red lines indicate the antennas that are being used by the virtual network operator.

tual network resources.

The antenna resources in a cloud-based RAN and spectrum resources in LSA are ideally infinite. In practice, they are much more than what a single virtual operator requires, but there is cost associated with their use. Therefore, resource allocation involves taking into account the unconstrained pool of resources and their usage cost. In [26], we discuss the problem of choosing the optimal set of spectrum and antenna resources to maximize resource efficiency (defined as the number of bits transmitted per cost unit, or resource utilisation cost). We assume that K users demand a wireless service from a virtual network operator who rents antennas from the cloud RAN and who rents LSA spectrum for the time needed to transmit the information. Using the cloud RAN infrastructure (processing, backhaul, antennas, etc.) and LSA spectrum involves cost, which is measured in currency units per second. The aim of the network operator is to choose the optimal number of antennas M and bandwidth W so that the number of transmitted bits per currency unit is maximised (**Fig. 4**).

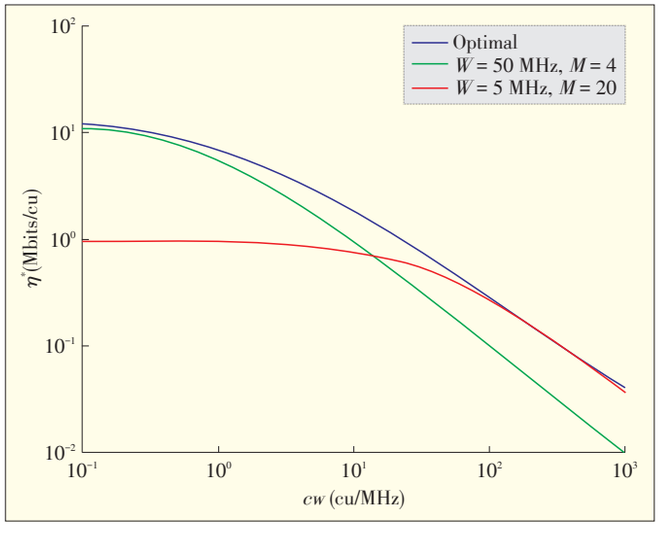
To serve K users, the service provider chooses M antennas and W MHz from the pool of resources offered by the cloud-based RAN and LSA. The spectrum cost c_w is the cost of using 1 MHz of bandwidth from the LSA for 1 s. The antenna cost c_m is the cost of using one distributed antenna for 1 s. The operative cost c_o is the cost of using the cloud infrastructure, e.g., backhaul and processing, for 1 s.

The cost efficiency is the number of transmitted bits per cost unit (bits/cu) and is given as the ratio of the total rate to costs:

$$\eta(M,W)=\frac{W\sum_{k=1}^{K}\log\left(1+\frac{r_k}{N_0W}\right)}{c_mM+c_wW+c_o} \quad (1)$$

where r_k is the power received by the kth user and N_0 is the noise power spectral density.

Fig. 5 shows the optimal cost efficiency and the cost efficiency when an arbitrary strategy is used that either maximises the number of antennas or bandwidth. We assume that W_{max} = 50 MHz and M_{max} = 20. The results show that the optimal solution transmits up to an order of magnitude more information for the same cost. Fig. 5 also shows that when the bandwidth cost is very low, maximising the bandwidth is near-optimal. Similarly, if the bandwidth is expensive, the number of active anten-

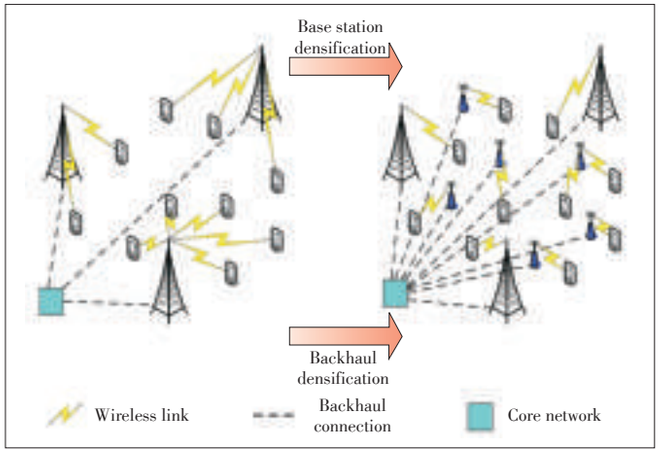


▲Figure 5. Cost efficiency for optimal spectrum and number of antennas chosen vs. cost efficiency for maximized spectrum or number of antennas.

nas should be maximised.

5 Small Cells

Base stations will be smaller and more numerous so that more users will be accommodated within the same spectrum. However, base station densification needs to be supported by a widely spread backhaul network. Hence, the number of backhaul links will increase along with the number of base stations. Backhaul links can be either wired or wireless. **Fig. 6** shows network densification.



▲**Figure 6. Network and backhaul densification.**

As part of the emerging interest in small cells, researchers have been investigating the performance gained by splitting cells. For instance, in dense networks, when the path loss attenuates proportional to a power of the distance, cell-splitting provides linear area spectral efficiency (ASE) gain with the density of nodes [27]. However, to date, researchers of small cells have not focused on how the total transmit power in the network changes as cells are split and what transmit power levels are needed to maintain linear gain. To address this, we first provide the expression for the minimum transmit power needed to guarantee linear ASE gain while splitting cells [28]. Then, we apply this expression, showing that the total transmit power of the network, i.e., the sum of the transmit power of all the base stations within a portion of the network, needed for linear ASE gain while splitting cells is a decreasing function of node density. This means that total transmit power can be significantly reduced by shrinking the cells and increasing node density.

The ASE is given by

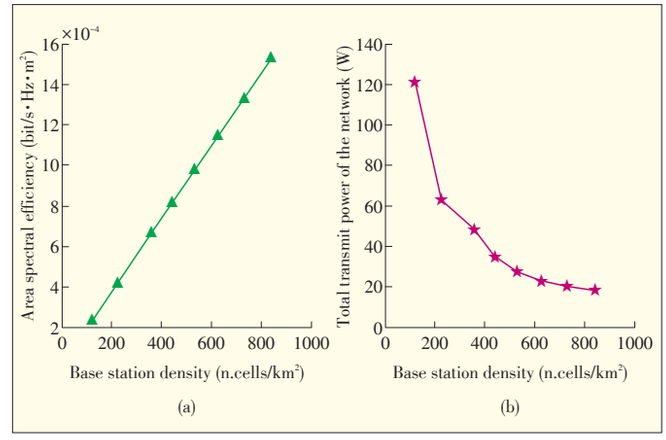
$$ASE = d \cdot \overline{C_{cell}} \tag{2}$$

where $d = D^{-2}$ is the cell density, D is the scaling factor, and $\overline{C_{cell}}$ is the average cell capacity. The total transmit power of the network, obtained by setting the nodes' power at the minimum value that still guarantees linear ASE during network scaling, is given by

$$P_{TX,tot} = P_0 D^{\beta-2} \bar{\alpha} L^2 \tag{3}$$

where L^2 is a square area of the network and contains N base stations; $P_0 > 0$ is an arbitrary power; $\beta \in \Re$ is the path loss exponent; $\bar{\alpha} = \frac{1}{N}\sum_{k=0}^{N-1}\alpha_k$, and α_k is related to the transmit power of the nth base station by $P_{TX,k} = P_0 D^{\beta}\alpha_k$.

Fig. 7a shows the linear gain in ASE as base station density increases, and **Fig. 7b** shows the reduction in total transmit power as base station density increases. Therefore, the total transmit power of the network can be decreased while maintaining linear ASE gain. Figs. 7a and 7b show two advantages of increasing the base station density: it enables higher throughput and reduces the overall power radiated by the base station antennas. The overall transmit power reduction achievable by setting the transmit power as specified in [28] may have positive implications for reducing the aggregate interference experienced by an incumbent willing to share spectrum with a secondary system of small cells. This may be particularly useful in future scenarios involving LSA or Authorised Shared Access (ASA) schemes in which small-cell networks exploit new spectrum-sharing opportunities.



▲**Figure 7. a) ASE gain vs. base station density. b) transmit power reduction vs. base station density.**

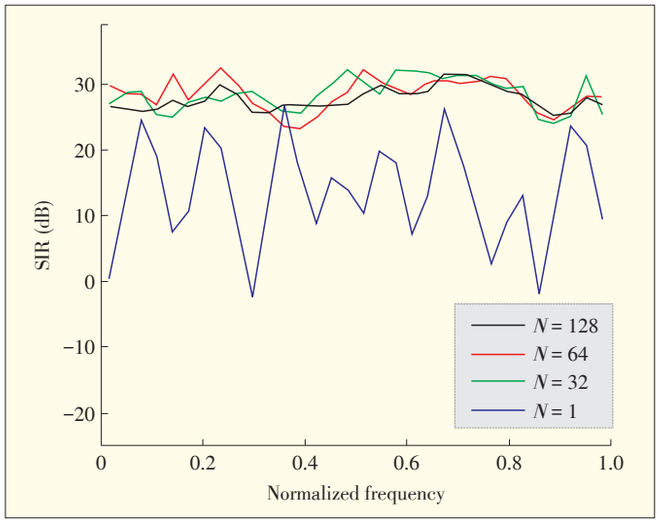
6 Combination of Massive MIMO and Filter Bank Multicarrier

In recent years, massive MIMO has gained momentum as a candidate to increase the capacity of multiuser networks. By increasing the number of antennas at the base station, the processing gain can be increased as much as necessary, and in theory, network capacity can be boundlessly increased [29]. An assumption made in [29] and followed by other researchers is that OFDM may be used to convert the frequency-selective channels between each mobile terminal and the multiple anten-

nas at the base station into a set of flat fading channels. Accordingly, the flat gains associated with subcarriers constitute the spreading gain vector that is used to de-spreading the respective data stream.

In [30], we introduced FBMC to massive MIMO communication. We found that in the case of massive MIMO systems, linear combining of the signal components from different channels smooths channel distortion. Therefore, we can relax the requirement of approximately flat gain for the subcarriers and significantly reduce the number of subcarriers in an FBMC system. In this way, we can reduce both system complexity and latency caused by the synthesis filter bank at the transmitter and analysis filter bank at the receiver. Also, constellation sizes can be bigger, which further increases bandwidth efficiency in the system. Moreover, increasing subcarrier spacing reduces the sensitivity to carrier frequency offset (CFO) and peak-to-average power ratio (PAPR). An additional benefit of FBMC is that carrier/spectral aggregation (i.e., using non-contiguous bands of spectrum for transmission) becomes simpler because each subcarrier band is confined to an assigned range, and the interference with other bands is negligible. This is not the case with OFDM [31].

Fig. 8 shows the effect of increasing the number of antennas at the receiver on the signal-to-interference ratio (SIR) for different numbers of subcarriers in a single-user case. This has implications for the system's ability to achieve a flat channel response over each subcarrier band. For the channel model used here, the total bandwidth is 2.8 MHz, which is equivalent to the normalised frequency one in in Fig. 8. The subcarrier spacing is 2800/L kHz, where L is the number of subcarrier bands. For example, when L = 32, the subcarrier spacing is 87.5 kHz. Compared with the subcarrier spacing in OFDM-based standards such as IEEE 802.16 and LTE, this is relatively broad (87.5/15 ≈ 6 times larger). Reducing the number of subcarriers or, similarly, increasing the symbol rate in each subcarrier band reduces transmission latency, increases bandwidth efficiency, and reduces sensitivity to CFO and PAPR.



▲**Figure 8. SIR of matched filter linear combining technique for N receive antennas and 32 subcarriers.**

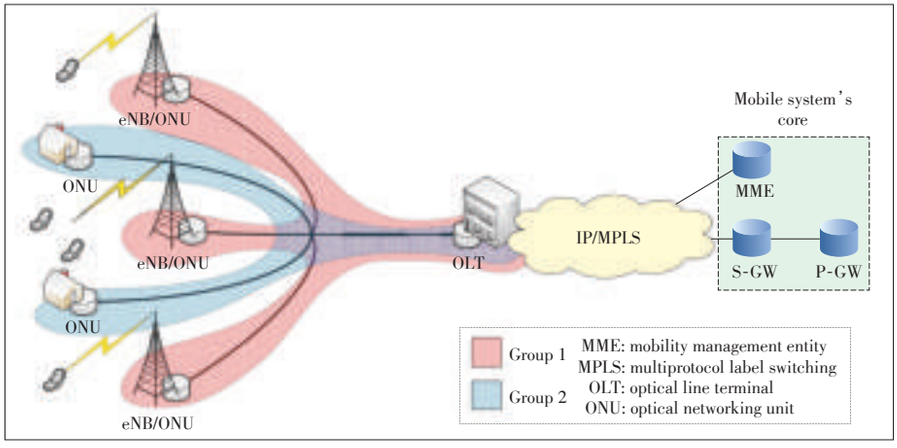
7 Optical-Wireless Integration

The main reason it has taken many years for fibre-to-the-home (FTTH) access systems to be deployed is financial viability. Ultimately, the main technology that enabled fibre access to become widespread was passive optical networking (PON) because it enabled the costs of fiber installation and deployment of electronic terminations to be shared between users [32]. As mobile network operators start looking into deploying many small cells to offer high capacity per user, shared, low-cost fibre backhaul based on existing PON systems is attractive. PON systems seem to be an ideal technology for mobile backhaul of small and large cells because such systems have the potential to provide ubiquitous access points. Indeed, the latest ITU-T PON standard, 10-gigabit-capable passive optical network (XG-PON) already encompasses fibre-to-the-cell (FTTCell) scenarios [33]. **Fig. 9** shows an FTTCell scenario where an XG-PON network is used to backhaul an LTE system by connecting the base stations to the core LTE components. In this architecture, the core components are connected to the root of the PON, also called the optical line terminal (OLT), and the base stations are connected to the optical network units (ONU) at the leaves of the tree-shaped optical distribution network.

Typically, a PON and its Dynamic Bandwidth Assignment (DBA) algorithms are designed for ONUs that are independent of each other and often represent individually billed entities. However, there are scenarios (e.g., FTTCell) where wireless operators may require more than one ONU per PON to provide service in different locations. These entities may wish to have a single service level agreement (SLA) for their group of ONUs and share the contracted capacity within the group of ONUs. In [34], we discuss hierarchical DBA algorithms that enable bandwidth to be scheduled to a group of base stations rather than scheduled individually. By assuring bandwidth to a group of base stations, a mobile operator can ensure that whenever one base station is not using its bandwidth, the bandwidth can be assigned first to a base station of the same operator rather than to anyone else in the PON.

By doing this, an operator can leverage the properties of statistical multiplexing and the heterogeneity of traffic from the base stations. With careful dimensioning, it is possible for base stations to transmit at their peak rate without guaranteeing the peak rate to each base station. It could be argued that the same effect could be achieved with best-effort bandwidth and without assuring bandwidth to the group, but in this case backhaul performance would depend on other users of the PON, possibly including competing mobile operators.

In [34], we propose group-GIANT (gGIANT), an algorithm



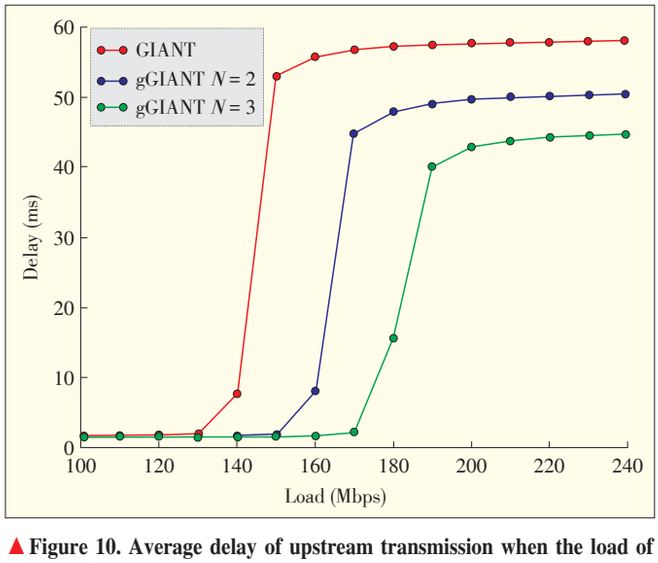
▲Figure 9. Fibre-to-the-cell architecture.

that enables group-assured bandwidth to be assigned, and we evaluate its performance through simulations in homogeneous and heterogeneous traffic scenarios. In the heterogeneous simulation, we changed the load of only one ONU, and kept the load of all the other ONUs constant. The average delay is shown in **Fig. 10**, where N is the size of the group, i.e., the number of ONUs belonging to the group.

By adding ONUs to the group, extra capacity is available to ONUs that need it. Comparing the more homogeneous scenarios with the more heterogeneous scenarios, we found that the gGIANT algorithm provides a larger performance increase when the traffic load is unbalanced. This supports the idea that the more heterogeneous and bursty the traffic is, the bigger the gains from group-assured bandwidth.

8 A Complex Systems Science View of 5G

Modern ICT systems are becoming increasingly larger because they encompass more and more components. At the same time, there is an ever-growing flow of information within the systems. As the technological and social trends in communications shift from systems based on closed hierarchical or semi-hierarchical structures to open, distributed networked organisations, new paradigms are needed to model, design, monitor, and control these new kinds of systems. Communication engineers are faced with the task of designing networks capable of self-organisation, self-adaptation and self-optimisation and that can satisfy user demand without disruption. In this regard, help comes from recent studies of complex systems in nature, society, and engineering. These studies give ideas on how to design and control modern communication systems [35].



▲Figure 10. Average delay of upstream transmission when the load of only one ONU is increased.

In [36], we move a step towards a comprehensive, rigorous study of communication systems by drawing on understanding and tools from complex systems science. We can apply complex system science to the problem of self-organizing frequency allocation in wireless systems by using local information and adaptations to achieve global network-wide behaviour. The system we propose is complex, both in terms of entropy and complexity metrics and shows that simple agents, such as cellular automata cells, are capable of nontrivial interference-free frequency allocation.

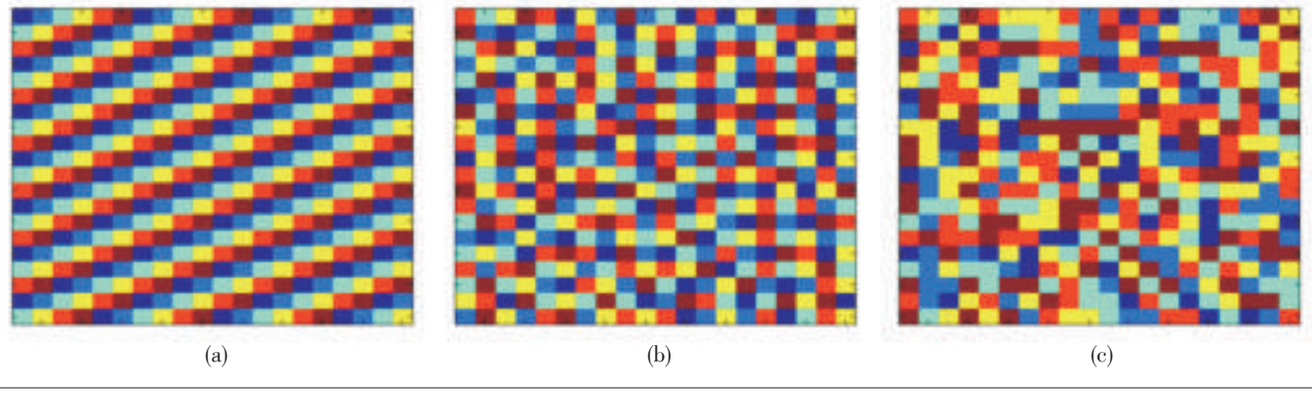
From the literature on complex systems science, complexity and entropy are two distinct quantities. The relationship between complexity and entropy is discussed in [37] and [38]. In [36], we use excess entropy to measure complexity. Excess entropy can be expressed in different ways, but we express it as the convergence excess entropy E_C, which is obtained by considering how the entropy density estimates converge to their asymptotic value h. In two dimensions, the entropy density h can be expressed as [39]

$$h = \lim_{M \to \infty} h(M) \tag{4}$$

where $h(M)$ is the entropy of a target cell X conditioned on the cells in a certain neighbourhood of X. Then, the excess entropy E_C is

$$E_C = \sum_{M=1}^{\infty} \left(h(M) - h \right) \tag{5}$$

We then study the global network-wide behaviour with respect to the complexity of the channel allocation matrix. **Fig. 11a** shows the channel-allocation matrix resulting from the regularly spaced assignment of $N = 6$ channels. This is a typical example of the frequency allocation resulting from centralized frequency planning. **Fig. 11b** shows the channel-allocation matrix resulting from the self-organising algorithm described in



▲**Figure 11. a) Regular channel assignment; b) Channel assignment resulting from the self-organising algorithm described in [36]; c) Random channel assignment.**

[36]. **Fig. 11c** shows a random allocation of $N = 6$ channels.

We estimate E_c and h for the three types of channel assignments in Fig. 11 by using $10^4 \times 10^4$ matrices. For the channel assignment in Fig. 11a, the entropy estimates are $h(M) = 0, \forall M = \{1,2, \ldots ,6\}$. Hence, $E_c = 0$ and $h = 0$. This is consistent with the crystal-like, completely ordered structure of the channel allocation matrix. For the random channel assignment matrix, the entropy estimates are $h(M) = 2.58, \forall M = \{1,2, \ldots ,6\}$. Hence, $E_c = 0$ and $h = 2.58$. Because the channel assignment matrix is completely disordered, the entropy is the maximum possible for $N = 6$ channels, i.e., $\log_2(6)$. In the case of the channel assignment matrix in Fig. 11b, the entropy estimates are $h(M) = 1.29, \forall M = \{4,5,6\}$. Hence, $h = 1.29$ and E_c is 2.04. Therefore, the channel assignment that emerges from self-organisation is highly structured, and neither centralised nor random channel allocation are capable of such a degree of structure.

With networks that allocate frequencies in a self-organised way, we can talk in terms of complex systems. With networks that manage frequencies in a centralized way, the resulting allocation has a regular crystal-like configuration, and there is no point in studying such networks using complex systems science. We consider [36] as a step towards a comprehensive study of complex communication systems, drawing on the philosophy and results from the multi-disciplinary field of complex systems science to design and analyze communication networks.

9 Conclusion and Outlook

It is a time of unprecedented change: traffic on telecommunication networks is growing exponentially, and many new services and applications are emerging. We cannot predict exactly what lies ahead, and the best we can do is to extrapolate some trends and make educated guesses. A possible way forward is to design networks with change in mind so that they will be more robust to disruption caused by growing demands, changing user patterns, and yet-unimagined applications. In this way, the risks associated with investing in these kinds of networks will be lower because they will be more durable and scalable. Networks that are designed with change in mind will also make effective use of spectrum, bandwidth, power, processing capabilities, and backhaul and will ensure a sustainable future.

5G is the next generation of wireless communication systems and is part of this picture. In this paper, the general landscape of 5G systems, including likely requirements and candidate technologies for satisfying these requirements, have been introduced. A few relevant areas of 5G have been discussed, and recent research in these areas by the author has been presented. In the last part of this paper, future communication networks were dicsussed from the perspective of complex systems science. One of the most widely accepted definitions of a complex system is "a system in which large networks of components with no central control and simple rules of operation give rise to complex collective behaviour, sophisticated information processing, and adaptation via learning or evolution" [40]. This view resonates with the author's research team when talking about future wireless networks. Indeed, networks are becoming increasingly distributed and formed by an ever-growing number of nodes that must make local decisions (because of limited signalling exchange capacity) when reacting to the environment. Yet these nodes also have to achieve a global level of good user experience and network performance in general. Therefore, there is ground to believe that telecommunications systems are evolving from being simple monolithic structures to being complex structures and that complex systems science might benefit the analysis and design of such structures.

References

[1] Qualcomm, "Rising to Meet the 1000x Mobile Data Challenge," White Paper, 2012.
[2] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2013–2018," White Paper, June 2014.

[3] Ericsson, "More than 50 billion connected devices," White Paper, 2011.
[4] METIS. (2012). FP7 European Project 317669 METIS (Mobile and wireless communications Enablers for the Twenty - twenty Information Society). [Online]. Available: https://www.metis2020.com/
[5] Cisco. (2014, Feb.). Visual networking index (NVI) white papers. [Online]. Available: http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/white-paper-listing.html
[6] C. Han, T. Harrold, S. Armour, *et al.*, "Green radio: radio techniques to enable energy efficient wireless networks," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 46–54, Jun. 2011. doi: 10.1109/MCOM.2011.5783984.
[7] J. G. Andrews, S. Buzzi, W. Choi, *et al.*, "What will 5G be?,", *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014. doi: 10.1109/JSAC.2014.2328098.
[8] 5GNOW. (2012). FP7 European Project 318555 5G NOW (5th Generation Non-Orthogonal Waveforms for asynchronous signalling). [Online]. Available: http://www.5gnow.eu/
[9] 5G PPP. (2013). 5G-Infrastructure Public-Private Partnership. [Online]. Available: http://5g-ppp.eu/
[10] C.-X. Wang, F. Haider, X. Gao, *et al.*, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 97–105, Feb. 2014. doi: 10.1109/MCOM.2014.6736752.
[11] Huawei, "5G: a technology vision," White Paper, 2013.
[12] G. P. Fettweis, "A 5G Wireless Communications Vision," *Microwave Journal*, Dec. 2012.
[13] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, Mar. 2013. doi: 10.1109/MCOM.2013.6476878.
[14] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: past, present, and future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, Apr. 2012. doi: 10.1109/JSAC.2012.120401.
[15] R. W. Heath, S. Peters, Y. Wang, and J. Zhang, "A current perspective on distributed antenna systems for the downlink of cellular systems," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 161–167, Apr. 2013. doi: 10.1109/MCOM.2013.6495775.
[16] G. Wunder, P. Jung, M. Kasparick, *et al.*, "5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 97–105, Feb. 2014. doi: 10.1109/MCOM.2014.6736749.
[17] G. Wunder, M. Kasparick, S. ten Brink, *et al.*, "System-level interfaces and performance evaluation methodology for 5G physical layer based on non-orthogonal waveforms," in *Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, USA, Nov. 2013, pp. 1659–1663. doi: 10.1109/ACSSC.2013.6810581.
[18] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 101–107, Jun. 2011. doi: 10.1109/MCOM.2011.5783993.
[19] F. Rusek, D. Persson, Buon Kiong Lau, *et al.*, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan. 2013. doi: 10.1109/MSP.2011.2178495.
[20] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, Feb. 2014. doi: 10.1109/MCOM.2014.6736746.
[21] G. P. Fettweis and S. Alamouti, "5G: personal mobile internet beyond what cellular did to telephony," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 140–145, Feb. 2014. doi: 10.1109/MCOM.2014.6736754.
[22] P. Rost, C. J. Bernardos, A. D. Domenico, "Cloud technologies for flexible 5G radio access networks," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, May 2014. doi: 10.1109/MCOM.2014.6898939.
[23] N. Ghazisaidi, M. Maier, and C. M. Assi, "Fiber-wireless (FiWi) access networks: a survey," *IEEE Communications Magazine*, vol. 47, no. 2, pp. 160–167, Feb. 2009. doi: 10.1109/MCOM.2009.4785396.
[24] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5G era," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 90–96, Feb. 2014. doi: 10.1109/MCOM.2014.6736748.
[25] F. Paisana, N. Marchetti, and L. A. DaSilva, "Radar, TV and cellular bands: which spectrum access techniques for which bands?" *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1193–1120, Jul. 2014. doi: 10.1109/SURV.2014.031914.00078.
[26] I. Gomez-Miguelez, E. Avdic, N. Marchetti, I. Macaluso, and L. Doyle, "Cloud-RAN platform for LSA in 5G networks tradeoff within the infrastructure," in *6th International Symposium on Communications, Control, and Signal Processing*, Athens, Greece, May 2014, pp. 522–525. doi: 10.1109/ISCCSP.2014.6877927.
[27] J. Ling and D. Chizhik, "Capacity scaling of indoor pico-cellular networks via reuse," *IEEE Communications Letters*, vol. 16, no. 2, pp. 231–233, Feb. 2012. doi: 10.1109/LCOMM.2011.121311.111971.
[28] C. Galiotto, N. Marchetti, and L. Doyle, "The role of the total transmit power on the linear area spectral efficiency gain of cell-splitting," *IEEE Communications Letters*, vol. 17, no. 12, pp. 2256–2259, Dec. 2013. doi: 10.1109/LCOMM.2013.101413.131620.
[29] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010. doi: 10.1109/TWC.2010.092810.091092.
[30] A. Farhang, N. Marchetti, L. Doyle, and B. Farhang-Boroujeny, "Filter bank multicarrier for massive MIMO," in *IEEE Vehicular Technology Conference*, Vancouver, Canada, Sep. 2014.
[31] B. Farhang-Boroujeny, "OFDM versus filter bank multicarrier," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 92–112, May 2011. doi: 10.1109/MSP.2011.940267.
[32] M. Ruffini, L. Wosinska, M. Achouche, *et al.*, "DISCUS: an end-to-end solution for ubiquitous broadband optical access," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 24–32, Feb. 2014. doi: 10.1109/MCOM.2014.6736741.
[33] ITU. (2010 Mar.) *10-Gigabit-Capable Passive Optical Networks (XGPON) Series of Recommendations*, ITU Standard G.987.x. [Online]. Available: http://www.itu.int/rec/T-REC-G/e
[34] P. Alvarez, N. Marchetti, D. Payne, and M. Ruffini, "Backhauling mobile systems with XG-PON using grouped assured bandwidth," in *19th European Conference on Networks and Optical Communications*, Milano, Italy, Jun. 2014, pp. 91–96. doi: 10.1109/NOC.2014.6996834.
[35] P. Antoniou and A. Pitsillides, "Understanding complex systems: a communication networks perspective," University of Cyprus, Nicosia, Cyprus, Tech. Rep. TR-07-01, 2007.
[36] I. Macaluso, H. Cornean, N. Marchetti, and L. Doyle, "Complex communication systems achieving interference-free frequency allocation," in *IEEE International Conference on Communications*, Sydney, Australia, Jun. 2014, pp. 1447–1452. doi: 10.1109/ICC.2014.6883525.
[37] R. López-Ruiza, H. L. Mancinib, and X. Calbet, "A statistical measure of complexity," *Physics Letters A*, vol. 209, no. 5–6, pp. 312–326, Dec. 1995. doi: 10.1016/0375-9601(95)00867-5.
[38] H. V. Ribeiro, L. Zunino, E. K. Lenzi, P. A. Santoro, and R. S. Mendes, "Complexity-entropy causality plane as a complexity measure for two-dimensional patterns," *Plos One*, vol. 7, no. 8, Aug. 2012. doi: 10.1371/journal.pone.0040689.
[39] D. P. Feldman and J. P. Crutchfield, "Structural information in two-dimensional patterns: entropy convergence and excess entropy," *Physical Review E*, vol. 67, no. 5, May 2003. doi: http://dx.doi.org/10.1103/PhysRevE.67.051104.
[40] M. Mitchell, *Complexity: A Guided Tour*, Oxford, UK: Oxford University Press, 2011.

Manuscript received: 2014-07-11

Biography

Nicola Marchetti (marchetn@tcd.ie) is an assistant professor at Trinity College Dublin and a member of the National Telecommunications Research Centre (CTVR). He received his PhD degree in wireless communications from Aalborg University, Denmark, and he also holds MSc degrees in electronics engineering and mathematics. He has collaborated on projects with Samsung, Nokia Siemens Networks, Huawei, and Intel Mobile Communications. His research interests include 5G systems, complex systems science, integrated optical-wireless networks, multiple antenna systems, radio resource management, small cells and HetNet, and waveforms. He has authored papers that have appeared in more than 50 refereed journals and conference proceedings. He has also authored and edited two books on telecommunications.

DOI: 10.3969/j. issn. 1673-5188. 2015. 01. 003
http://www.cnki.net/kcms/detail/34.1294.TN.20150227.1110.001.html, published online February 27, 2015

Signal Processing Techniques for 5G: An Overview

Fa-Long Luo
(Element CXI, San Jose, California 95131, USA)

Abstract

This paper gives an outline of the algorithms and implementation of the main signal processing techniques being developed for 5G wireless communication. The first part contains a review and comparison of six orthogonal and non-orthogonal waveform-generation and modulation schemes: generalized frequency-division multiplexing (GFDM), filter-bank multicarrier (FBMC), universal filtered multicarrier (UFMC), bi-orthogonal frequency-division multiplexing (BFDM), sparse-code multiple-access (SCMA), and non-orthogonal multiple access (NOMA). The second part discusses spatial signal processing algorithms and implementations for massive multiple-input multiple-output (massive-MIMO), 3D beamforming and diversity, and orbital angular momentum (OAM) based multiplexing. The last part gives an overview of signal processing aspects of other emerging techniques in 5G, such as millimeter-wave, cloud radio access networks, full duplex mode, and digital radio-frequency processing.

Keywords

3D beamforming 5G; massive MIMO; GFDM and spatial multiplexing

1 Introduction

Signal processing techniques have had the most important role in wireless communications since the second generation of cellular systems. In 2G, 3G and 4G, the peak service rate has been the dominant metric for performance. Each generation has a development cycle of about ten years and is defined by a standout signal processing technology that represents the most important advancement in that generation. In 2G, this technology was time-division multiple access (TDMA); in 3G, it was, code-division multiple access (CDMA); and in 4G, it is orthogonal frequency-division multiple access (OFDMA). However, this will not be the case for 5G systems— there will be no dominant performance metric that defines requirements for 5G technologies. Instead, a number of new signal processing techniques will be used to continuously increase peak service rates, and there will be new emphasis on greatly increasing capacity, coverage, efficiency (power, spectrum, and other resources), flexibility, compatibility, and convergence. In this way, 5G systems will be able to handle the explosion in demands arising from emerging applications such as big data, cloud services, and machine-to-machine (M2M) communication.

A number of new signal processing techniques for 5G systems have been proposed and are being considered for international standardization and deployment. This article gives an overview of promising signal processing techniques, both from a practical and standardization point of view. In particular, it emphasizes orthogonal and non-orthogonal modulation techniques as well as spatial processing techniques such as massive multiple-input multiple-output (massive-MIMO), three-dimensional beamforming and diversity, and multiplexing based on orbital angular momentum (OAM).

The rest of this paper is organized as follows. In section 2, we present six modulation schemes that offer better data transmission and higher peak rates than existing modulation schemes. The six modulation schemes we present are: generalized frequency-division multiplexing (GFDM) [1], filter bank multi-carrier (FBMC) [2], universal filtered multi-carrier (UFMC) [3], bi-orthogonal frequency division multiplexing (BFDM) [4], sparse-code multiple access (SCMA) [5], and non-orthogonal multiple access (NOMA) [6]. In section 3, we discuss spatial signal processing for 5G, focusing on massive-MIMO, adaptive 3D beamforming, and OAM-based multiplexing. In section 4, we give an overview of signal processing-aspects of emerging 5G techniques, such as millimeter wave, cloud radio access, full duplex access, and digital radio-frequency processing. In section 5, we offer some conclusions.

2 Signal Processing Algorithms for Modulation and Waveform Generation

Modulation processing involves using data and information to be transmitted to change the signal waveforms in specific al-

gorithms. Such processing determines many factors in a wireless system, including transmission speed, spectral efficiency, power consumption, signal-to-noise ratio, and implementation complexity. OFDM and OFDMA are used in 4G systems and have the following advantages:

- They eliminate inter-cell interference by ensuring orthogonality between each subcarrier.
- They use fast Fourier transform (FFT), which means they can be easily implemented and integrated with MIMO and multiple antennas.
- They can average the interferences within a cell by using allocation with cyclic permutation.
- They adapt transmission power according to the bit rate of the user.
- They ensure frequency diversity by spreading the carriers across the used spectrum.
- They are robust to inter-symbol interference (ISI) and multipath distortion.

OFDM and OFDMA have the following main disadvantages:

- a relatively high peak-to-average power ratio (PAPR) due to the fact that modulated symbols are transmitted in parallel and each contains part of the transmission,
- limited spectral efficiency due to the need for a cyclic prefix (CP) and null guard tones at the spectral edges,
- high sensitivity to frequency offsets and phase noise and the need for strict synchronization.

These disadvantages prevent OFDMA schemes from being immediately used in 5G systems, and more advanced modulation schemes, such as GFDM, FBMC, UFMC, BFDM, SCMA and NOMA, have been investigated.

2.1 GFDM

Equations (1) and (2) are the simplified equations for modulating a time-domain symbol in the OFDM scheme and GFDM scheme, respectively. Modulation of the time-domain signal occurs before CP processing.

$$x(n)=\sum_{k=0}^{K-1}d(k,m)e^{-j2\pi\frac{kn}{N}} \tag{1}$$

$$x(n)=\sum_{m=0}^{M-1}\sum_{k=0}^{K-1}d(k,m)g[n]\times\delta[n-mN]e^{-j2\pi\frac{kn}{N}} \tag{2}$$

where n is the time index, k is the sub-carrier index, m is the time-symbol index, M is the number of symbols per sub-carrier, K is the number of active sub-carriers, N is the total number of sub-carriers (the length of Discrete Fourier Transform), $x(n)$ is the transmitted samples, $g[n]$ is the coefficients of the shaping filter, and $d(k,m)$ is the coded data related to mth symbol. With OFDM, only one symbol can be modulated over active subcarriers (frequency bins); however, with GFDM, multiple symbols can be modulated over active subcarriers. This means that GFDM allocates the data in a two-dimensional time-frequency block structure by introducing flexible pulse shaping for the individual subcarriers. Because GFDM modulates multiple symbols per subcarrier through a two-dimensional time-frequency structure, it has a number of benefits in terms of efficiency, performance, and complexity:

- GFDM can control out-of-band (OOB) emission and reduce PAPR much more than OFDM by adjusting the shaping filters. OOB emission and PAPR are two of the major drawbacks of OFDM for advanced wireless systems. In addition, GFDM allows fragmented spectrum and dynamic spectrum allocation without these severely interfering with existing services or other users.
- Orthogonality between the subcarriers is dismissed in GFDM, and variable pulse-shaping filters reduce the effect of frequency offset and phase noise without the need for strict synchronization processing of symbols.
- A short CP is a simple way of reducing the multipath distortion. A matched filter receiver with iterative interference cancellation can also reduce ISI and intercarrier interference (ICI) caused by subcarrier filtering. More importantly, by adding a single CP for an entire block that contains multiple symbols, GFDM can be used to improve the spectral efficiency of the system.

From (2), the filter processing, i.e., the circular convolution in the time domain or multiplication in the frequency domain, significantly increases computational complexity in a GFDM system. On the other hand, GFDM has advantages, such as reduced PAPR, that result in the reduction or even elimination of other processing units, such as digital pre-distortion (DPD) or crest-factor reduction (CFR), both of which are essential in current wireless and broadcasting systems [7].

Equation (1) is a special case of (2). In (1), M and coefficients are all unity, i.e., only one symbol and rectangular-form filtering shape. In other words, a GFDM system can made compliant and easily integrates with existing OFDM systems.

2.2 FBMC

Equation (1) can be generalized to the following, which is used in FBMC:

$$x(n)=\sum_{k=0}^{K-1}d(k,m)\times g_k(n) \tag{3}$$

where $g_k(n)$ is the impulse response of the kth filter. The relationship between the indexes m and n is not shown in (3). FBMC can be described as a synthesis-analysis filter-bank-based modulation scheme, where synthesis filtering is introduced using FFT/IFFT and a poly-phase filter structure.

In FBMC, the modulated data at each subcarrier is shaped by a well-designed prototype filter that is different from the rectangular pulse filter in OFDM. This prototype filter has many degrees of freedom to use different waveform shapes and can greatly suppress a signal's side-lobes, making them strictly band-limited. By filtering on a per-subcarrier basis, inter-carrier interference can be greatly reduced when there is fre-

quency jitter and offset due to the Doppler Effect or misaligned oscillators. This strict band-limiting requirement makes the transmit filter impulse response long. Typically, the filter has three or four times the length of the symbols. As a result, FBMC can only provide good spectral efficiency if the number of transmit symbols is large. Unlike GFDM, FBMC is not suitable in low-latency scenarios, where efficiency must be high for bursty transmission.

Because there is no CP in an FBMC system, a more complex equalization system is required at the receiver side of the system to resolve multipath issues and further reduce ISI. Meanwhile, FBMC requires much more additional processing than OFDM to insert the CP in the transmitter and remove the CP at the receiver side.

Equation (3) shows that the OFDM scheme is a special case of FBMC, where a prototype filter with a rectangular impulse response is applied, and the overlap factor is the unity. This suggests that FBMC is compatible with OFDM-based systems and also has a large PAPR, which is similar to OFDM. This makes FBMC systems especially vulnerable to nonlinearity in the transceiver chain, which includes power amplifier, digital-to-analog converter, and analog-to-digital converter. More importantly, existing techniques for reducing PAPR in OFDM cannot be immediately used in an FBMC system [8]. These techniques include amplitude clipping, coding, interleaving, partial transmit sequence, selected mapping, tone reservation, tone injection, and active constellation extension. It is highly desirable to develop more efficient algorithms to reduce PAPR in an FBMC system.

2.3 UFMC

FBMC introduces filtering on a per-subcarrier basis; therefore, the lengths of the filter are comparatively long. To overcome this problem, filtered OFDM can be used. With filtered OFDM, filtering is introduced over the whole band. As a result, the filter bandwidth is much higher, and the filter length is much shorter than that in FBMC. For more flexibility and greater generalization, UFMC introduces filtering to subsets of the whole band instead of to a single subcarrier or the whole band. Modulation processing in UFMC can be simplified as

$$x(n)=\sum_{b=0}^{B-1}\sum_{k=0}^{K_b-1}d(k,m)\times g_k(b,n) \tag{4}$$

where B is the number of sub-bands, K_b is the number of subcarriers in the bth sub-band, and $g_k(b,n)$ is the impulse response of the corresponding kth filter in bth sub-band.

UFMC becomes filtered OFDM if filtering is introduced across the whole band (K_b is the unity) and becomes FBMC if filtering is introduced for a single subcarrier (B is the unity). As a result, UFMC has the advantages of both FBMC and filtered OFDM but does not have any new drawbacks. On the other hand, extensive design trade-offs in terms of performance, complexity, latency and spectrum are needed in UFMC to determine the number of sub-bands and subcarriers in each sub-band. For example, if UFMC is applied when there is fragmented spectrum, B should be selected according to the number of available spectral sub-bands. Furthermore, B may even vary with time if some of the spectral sub-bands are only occasionally populated by other wireless services. There is an optimal combination of B and K_b for a given performance index, such as bit error rate (BER), packet error rate (PER), signal-to-interference noise ratio (SINR), or PAPR. Moreover, the number of subcarriers in one sub-band used in UFMC may be different from that in another sub-band. In other words, K_b in (4) could be a variable, which allows greater flexibility in designing a UFMC system. In addition, the single sub-band may be subdivided into smaller chunks of different sizes in every sub-band. This streamlines the overall system and enables more fine-grain control of spectral characteristics at the cost of increased implementation complexity.

The design of the filter response $g_k(b,n)$ is another important aspect of a UFMC system because side-lobe attenuation determines the reduction of OOB and filter length, both of which greatly contribute to implementation complexity. On the other hand, many well-designed filters, such as finite impulse response (FIR) filters (defined by Dolph-Chebyshev windows), could be used in UFMC to reach a good compromise between complexity and performance.

2.4 BFDM

BFDM changes the set of signals at the transmitter and receiver sides so that they are bi-orthogonal instead of orthogonal. This gives the time-frequency representations of these signals pairwise (not individual) orthogonality and enables greater flexibility in terms of side-lobe attenuation, filter response, and implementation complexity when designing a transmission prototype [9].

An important reason for introducing BFDM in a 5G network is efficiently support machine-type-communication (MTC), characterized by a dramatic increase in sporadic traffic. Bulky 4G random-access procedures cannot handle such traffic [4]. Unlike LTE, where data is only carried using the physical uplink shared channel (PUSCH), BFDM-based schemes enable small user data packets to be transmitted through the available physical layer random-access channel (PRACH). A new data section, called Data-PRACH, is introduced between synchronous PUSCH and PRACH to support efficient asynchronous data transmission and to significantly reduce signaling overhead. In the proposed Data-PRACH processing, a pulse sequence shapes the spectrum of the preamble signal by using of PRACH guard bands under acceptable interference.

Bi-orthogonality and relatively long PRACH symbols ensure the BFDM scheme is more robust than conventional OFDM to frequency offset and phase noises during transmission. In other words, the main advantage of BFDM over conventional OFDM is a better compromise on performance degradation caused by

time and frequency offset. However, the effect of frequency offset and phase noise on BFDM is still much higher than that on GFDM, FBMC, and UFMC.

All the other advantages of OFDM still remain in BFDM—ISI and multipath distortion are easily reduced by CP, BFDM is easy to implement because of FFT and IFFT processing. However, BFDM needs to handle long pulse tails, which reduces efficiency of bursty transmission. This efficiency is critical for low-latency and M2M applications. More importantly, the BFDM scheme discussed here cannot be easily integrated with massive MIMO unless modifications, such as generalization of the above concepts to UFMC or GFDM, are made [10], [11].

2.5 SCMA

Modulation processing involves changing the source binary sequences to a new binary sequence before being sent to the transmitter front-end. There are many ways to do the desired modulation processing. SCMA is, in effect, a modified version of multicarrier CDMA based on low-density signature (LDS), where mapped symbols, following forward error correction (FEC), are allocated according to a pre-designed low density spreading sequence. In this way, near maximum likelihood (ML) performance is achieved at the receiver side [5], [12].

Unlike the above two-step processing, SCMA merges the mapping of the bits that are coded by FEC into complex symbols with the spread of these mapped symbols. This results in one-step processing. In other words, SCMA directly maps the binary outputs of FEC to a complex code word that is in a multidimensional domain and should be selected from a predefined codebook called SCMA codebook. By generating multiple different codebooks that are predefined for different users or layers, SCMA supports multiple access. In fact, each user has a unique codebook in SCMA. Code words in the SCMA code books are sparse, so the iterative message-passing algorithm can still be used for near-optimal detection without significantly increasing the processing complexity. Any increase in such complexity can be compensated to some extent by the advantages (in terms of hardware implementation) resulting from one-step processing.

SCMA codebook based on multidimensional lattice constellation exploits shaping and coding gain, which helps SCMA increase spectral efficiency and makes link adaptation more reliable because of related interference averaging and management. Moreover, SCMA enables massive connectivity while having good features, such as overloaded signal superposition, low signaling overhead, low latency, and high flexibility in the link-adaptation mechanism [13].

Designing SCMA codebooks for multiple users or layers is very complicated in terms of optimization and programming. Practical solutions are highly desirable and still being developed. One-step processing to map the binary sequences that are output from the FEC to code word could be considered as complex nonlinear mapping. This mapping could be performed by universal mapping neural networks, such as multilayer-perceptron (MLP) neural networks and radial-basis-function (RBF) neural networks [14]. Furthermore, the related global optimized learning rules in these mapping neural networks could be applied to the design of an SCMA codebook.

2.6 NOMA

In addition to the time and frequency domains (used in the modulation schemes previously mentioned) and the spatial domain (used in MIMO, beamforming, and OAM multiplexing), NOMA makes use of the power domain for modulation processing and multiplexing according to the power difference and loss between users. Specifically, NOMA superposes multiple users in the power domain and forms superposition coding, where users are separated by successive-interference cancellation (SIC) and available capacity-achieving channel codes, such as Turbo code and low-density parity check (LDPC). A user with high channel gain is allocated less power, and a user with low channel gain is allocated more power [6]. In this way, all users with different channel gains have similarly high decoding probability and similarly large interference cancellation. This increases system capacity and coverage and supports mass connectivity. Moreover, NOMA promises robust performance in practical wide-area deployments despite mobility or channel-state information (CSI) feedback latency because user multiplexing in NOMA does not require fine feedback signaling from the user side, frequency selective channel quality indicator (CQI), or CSI.

The authors of [6] and other related publications have studied NOMA in terms of multiuser power allocation, signaling overhead, SIC error propagation, performance enhancement in high-mobility scenarios, and integration with MIMO and have shown that NOMA increases capacity and cell-edge throughput. The basis carrier waveforms in NOMA can still be generalized from OFDMA or FBMC, which means that NOMA retains the advantages of OFDMA and FBMC.

Table 1 shows a side-by-side comparison of the six algorithms discussed so far. From a standardization point of view, in-depth investigation and testing are required before any individual algorithm can be included in 5G specifications. It is perhaps more practical to develop one new algorithm that combines all the advantages of the six algorithms and minimizes the disadvantages. Moreover, crossover between and combination of FEC, modulation processing, and even source coding could be another direction towards achieving a better system.

Previous wireless standardization has occurred without enough consideration of hardware chips and real silicon (computing platforms and digital signal processors). Thanks to great advancements in computing and processing technology, in particular system-on-chip (SoC) and reconfigurable processing technology, a fully software-defined modulator and waveform generator is even possible in 5G standards. A system with these technologies could support multiple algorithms or even

▼Table 1. Six modulation algorithms

	GFDM	FBMC	UFMC	BFDM	SCMA	NOMA
PAPR	Low	High	Medium	High	High	High
OOB	Very Low	Low	Low	Medium	Medium	Medium
Spectral efficiency	Medium	High	High	Medium	Medium	Medium
Processing complexity	Medium	High	High	Low	Medium	High
CP	Yes	No	No	Yes	No	Yes
Orthogonality	No	Yes	Yes	Yes	No	No
ISI/ multipath distortion	Medium	High	High	Low	Medium	Low
Synchronization requirement	Medium	Low	Low	Medium	Low	Low
Effect of frequency offset and phase noise	Medium	Medium	Medium	Medium	Low	Low
Latency	Short	Long	Short	Long	Low	Long
Compatibility with OFDM	Yes	Yes	Yes	Yes	No	No
Ease of integration with MIMO	Yes	Yes	Yes	No	Yes	Yes

any algorithm, without any performance cost, by simply changing related software. In other words, the 5G standard only needs to define the related interfaces and control information and allow all other processing units, from FEC through modulation, to be open and software-programmable [15], [16].

3 Spatial Signal Processing for 5G

Spatial-domain signal processing techniques such as MIMO, beamforming and antenna diversity have primarily been used in 4G and digital broadcasting systems. In 5G, these spatial signal processing techniques will be further improved, and related new algorithms, such as massive MIMO [17]–[19] and three-dimensional beamforming [20]–[22], are being developed with an emphasis reaching a good compromise between processing complexity and performance. OAM-based spatial-processing techniques could be used improve a number of factors in 5G, such as capacity, efficiency and coverage [23]–[26]. Here, we discuss important practical aspects of these spatial signal processing techniques.

3.1 Massive MIMO

Strictly speaking, original MIMO is actually a kind of multi-channel time-domain processing, where processing is mainly done in the baseband alone and not much in the spatial domain. However, in 4G, MIMO uses multiple antennas at both the transmitter side and receiver side in order to multiply the capacity of a radio link by exploiting multipath propagation. That is, 4G MIMO exploits spatial-domain properties or, for example, spatial multiplexing, by allowing a base station to simultaneously serve multiple users who are using the same time-frequency resource.

Although the number of antennas is not strictly specified in current standards, four or eight antennas are most common. Massive MIMO expands on MIMO by dramatically increasing the number of antennas used at the base station (on the order of hundreds). This suggests that the number of antennas is significantly higher than the number of users being simultaneously served in the same time-frequency block. Hundreds of antennas serving dozens of users simultaneously increase spectral efficiency five- to ten-fold, and many degrees of freedom become available to increase SINR through transmission signal shaping, interference nullification, and formation of desired directional patterns.

Channel estimation, signal detection, pre-coding, and pilot contamination reduction are the main aspects of signal processing in massive MIMO. Channel estimation involves estimating the coefficients (parameter matrix) of channels according to the available samples and optimization criterion, such as minimum mean-square error (MMSE) or least square (LS). Signal detection in massive MIMO involves detecting the desired data streams from the samples, which are affected by interference and noise in the either passive or active form. Pre-coding could be considered for multiplying the original signal vector from all channels (antennas) at the transmitter side. In massive MIMO based on time-division duplexing (TDD), pilot sequences transmitted from users in the uplink become active interference sources and affect channel estimation processing. Eigenvalue-based filtering can be introduced to reduce the effect of pilot contamination.

Computation in massive MIMO mainly involves matrix multiplication, matrix inverse, eigenvalue decomposition (ED), or singular-value decomposition (SVD). The key to implementing massive MIMO in 5G is ensuring that the implementation of these very large dimensional matrix computations in real silicon is effective in terms of power, price and performance. One promising architecture being developed is reconfigurable computing array. With this architecture, the matrix computations required in massive MIMO can be performed in a manner as good as that of ASIC, as flexible as FPGA/DSP, and as easy as C language because the architecture has homogeneous interfaces and heterogeneous processing units [15], [16].

3.2 3D Beamforming and Diversity

Beamforming is a major spatial signal processing technique where using multiple antennas are used to change the beam pattern and steer the beam in a specific direction so that SINR is increased. With diversity technology, multiple antennas are used at the receiver side, and spatial filtering is introduced to optimize reception in noisy and mobile environments.

Traditionally, beamforming is only designed for the horizontal plane and is thus called two-dimensional beamforming. The spatial-domain information and properties in the vertical plane are unused. 3D beamforming, which encompasses both the elevation and azimuth planes, could open up more space to improve performance at the cost of increased processing complexity. In a 5G system, 3D beamforming can increase user capaci-

ty, coverage, and spectral and energy efficiency, and it can reduce inter-cell and inter-sector interference [20]–[23]. For example, different power levels can be allocated to the 3D beam patterns that serve the cell edge and cell center separately so that inter-cell interference is more effectively reduced. Both the vertical and horizontal beam patterns can be shaped and steered by adjusting the antenna tilt, the angle between the horizontal plane and boresight direction of the antenna.

At the receiver side, both the azimuth and elevation of arrivals can also be used, and additional degrees of freedom are available to improve the performance of antenna diversity. 3D techniques can also provide more flexibility in the design and configuration of an antenna array. Planar, circular, spherical, cylindrical, uniform, non-uniform, and end-fire topologies can be used. Also, both the co-polarized and cross-polarized antenna elements can be included in the antenna array.

Beamforming and diversity processing mainly involves matrix (vector) multiplication units, filtering units, and IFFT/FFT units. These main processing units have linear properties; thus, changing the order of these processing units does not affect system performance but greatly reduces processing complexity. **Fig. 1** shows a post-FFT diversity scheme where M FFT operations need to be performed and $N \times M$ unknown coefficients need to be estimated. The number of antennas is given by M, and the length of the FFT is given by N. As in **Fig. 2**, when the order changes, the number of unknown variables is given by $N + M$, which is a large reduction the $N \times M$ unknown variables in the original post-FFT scheme. Furthermore, only one FFT operation is needed when the order is changed instead of M FFT operations for the post-FFT scheme (Fig. 1). Using the cumulative and distributive properties of linear processing, the output is the same for these two schemes.

3.3 OAM

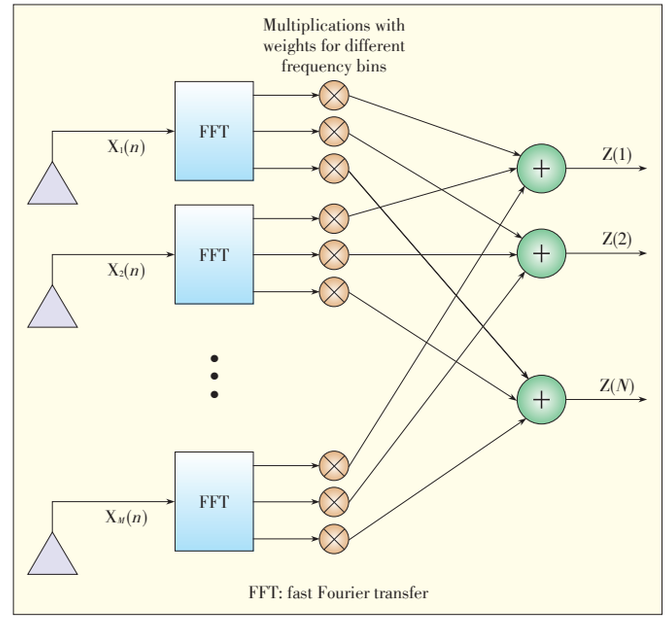
OAM-based spatial processing is a new tool for increasing capacity, spectral efficiency and scalability and decreasing channel interference in a 5G system.

The angular momentum carried in electromagnetic (EM) fields comprises spin angular momentum (SAM) and orbital angular momentum (OAM). These describe the polarization state and phase structure distribution, respectively. An EM wave carrying OAM has a helical transverse phase structure $\exp(j\ell\varphi)$, where φ is the transverse azimuthal angle, and ℓ is an unbounded integer (the state number of OAM). Each OAM beam at the same carrier frequency can carry an independent data stream; therefore, an OAM system can increase capacity and spectral efficiency by a factor equal to the values of state number ℓ. In addition, OAM beams with different ℓ values are mutually orthogonal, which implies low channel interference and crosstalk in transmitted and received data. Communicating over sub-channels given by OAM states is a subset of MIMO solutions; therefore, it does not provide any additional increase in system capacity if spatial multiplexing uses multiple spatially separated transmitter and receiver aperture pairs to transmit multiple data streams. In other words, multiplexed beams based on OAM should be completely coaxial throughout the transmission medium and use only one transmitter and receiver aperture (with certain minimum aperture sizes) to achieve OAM beam orthogonality and efficient de-multiplexing.

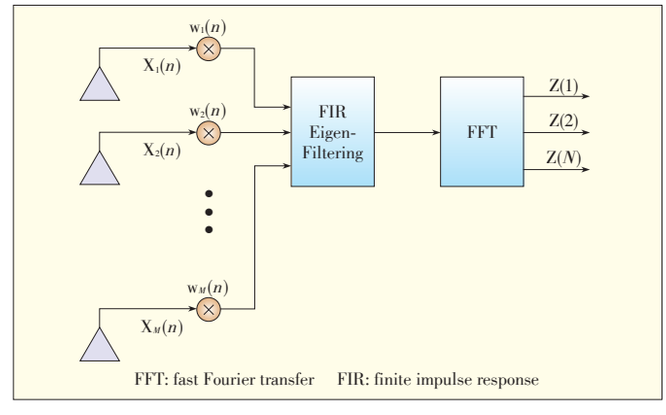
The benefits of OAM-based spatial multiplexing for wireless communication have been shown in millimeter-wave band, which 5G will encompass. However, there needs to be extensive R&D on OAM before OAM-based processing is included in 5G standards. A possible major use of OAM-based systems is to serve the link between wireless and optical channels.

4 Signal processing for Other Emerging 5G Techniques

In addition to advanced modulation algorithms and spatial



▲Figure 1. A post-FFT diversity scheme.



▲Figure 2. Version of the post-FFT scheme with order changed.

processing, a number of other new techniques will be used in 5G. These techniques relate to system architecture, protocols, physical-layer (downlink and uplink), air-interface, cell acquisition, scheduling and rate adaption, access procedures, relaying, and spectrum allocation. The techniques include centimeter- and millimeter-wave, smart spectrum sharing and access, simultaneous transmission and reception (full-duplex), device-to-device communication, advanced inter-node ordination, cloud radio access networking (C-RAN), software-defined networking, and digital radio frequency (RF) processing. Signal processing algorithms and implementations will be central to these emerging technologies, enabling them to make 5G a marketable reality.

Because of the much higher frequencies and wider bandwidths in millimeter-wave transmission, the physical propagation channels are more complicated, and advanced algorithms for channel modelling and estimation, signal detection, equalization, and error-correction coding are necessary. The main drawback of millimeter-wave is the high path loss, so new time-varying signal processing techniques, such as rapid beam adaptation, are necessary. In addition, millimeter-wave transmission only complements lower frequencies by providing high capacity and high data rates in dense urban areas instead of replacing lower frequencies, which should remain in the backbone to provide full wide-area coverage [27]. This implies a need for algorithms and implementations that support multiband and wideband.

Full-duplex mode enables simultaneous transmission and reception by sharing available resources between both directions of communication. In theory, this can double the link capacity. More importantly, full-duplex mode benefits the signaling and control layers because the uplink and downlink no longer need to be separated. Cancellation of the self-transmitted signal is the most important issue in full-duplex mode and needs to be done at all stages—from antenna, RF filtering, and digital front-end to baseband processing. The gain of the PA in full-duplex mode might be limited by the level of cancellation of the self-transmitted signal. In other words, a transceiver with a large PA gain becomes unstable in full-duplex mode if the self-transmitted signals are not sufficiently cancelled [28].

5G systems not only need new and higher frequency bands; they also aim to use the available spectrum as efficiently as possible through spectrum sharing, which can be implemented by authorized shared access (ASA) and co-primary shared access. Signal processing algorithms for spectrum sensing are central to smart spectrum sharing and co-existence. Spectrum sensing involves monitoring other frequency channels that can be used by the primary channel and deciding whether users served in one channel can be switched to the other without interrupting the link. Signal processing related to spectrum sharing include weak-signal detection, signal classification, estimation of location and direction, channel aggregation, and interference cancelation [29].

C-RAN is a promising architecture for expanding a wireless network. In C-RAN, the baseband processing unit (BBU) is moved from the base station (BS) to the control unit (CU), and the BS operates as the radio unit (RU). Baseband signals are transferred between the CU and RU through front haul links in complex in-phase and quadrature (IQ) samples. Because of the high bit rate and large bandwidth of the transferred data, effective signal compression prior to transmission on the front haul link is desirable. The signal sampling can be reduced, but this can result in significant performance loss. Nonlinear quantization is the second-simplest solution, but it negatively affects related interfaces. IQ-data-based compression algorithms are being closely investigated and are categorized as point-to-point, multi-terminal, multivariate, and structured coding. These algorithms reach the best compromise between compression rate (efficiency), system complexity, interface compatibility, and implementation cost [30].

Flexibility is one of the most important features of a 5G systems. Achieving the desired flexibility in a 5G system depends not only on advanced signal processing algorithms but also powerful hardware for processing. As mentioned in section 3, conventional processors, such as ASIC, FPGA and DSP, are not the best practical solution for 5G, and new kinds of processors that take into account the properties of new 5G signal processing algorithms are desirable. Also, new signal processing algorithms need to be designed with hardware architecture and programming model in mind— algorithm development should not be disconnected from hardware implementation. In addition, digital signal processing technologies for RF and front-end have advantages in terms of power efficiency, cost, time-to-market, and SDR networking. These technologies support multiple bands, multiple standards, and multimode applications in 5G. RF signal processing techniques encompass digital pre-distortion; digital up-conversion; digital down-conversion; DC-offset calibration; PAPR, CFR; pulse-shaping; delay, gain, and imbalance compensation; noise shaping; numerical controlled oscillator (NCO); full-duplex decoupling; and MIMO channel calibration [7].

For more detailed algorithms and implementations of emerging 5G techniques, refer to the publications in the reference list. Excellent representatives include [31] and [32], which describe full-dimensional MIMO; [33], which describes 3D channel modeling; and [34], [35], which describe millimeter-wave systems.

5 Conclusion

This paper outlines and compares six promising modulation algorithms for 5G in terms of PAPR, OOB, processing and implementation complexity, spectral efficiency, CP requirement and related ISI/ multipath distortion, orthogonality and related frequency offset and phase noise, synchronization in the time and frequency domains, latency, compatibility, and integration

with other processing units. These six algorithms are GFDM, FBMC, UFMC, BFDM, SCMA and NOMA. Spatial signal processing techniques—i.e., 3D beamforming, massive MIMO, and OAM-based multiplexing—have been discussed from both an algorithm and hardware implementation point of view. This paper also briefly discussed signal processing for other emerging technologies in 5G, such as millimeter-wave, C-RAN, full-duplex access, smart spectrum sharing, and digital RF processing. To bring the desired 5G to market, huge effort needs to be put into R&D on algorithms and silicon implementation.

References

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

[2] B. Farhang-Boroujeny, "OFDM versus filter bank multicarrier," *IEEE Signal Processing Magazine*, vol.28, no.3, pp.92– 112, May 2011. doi: 10.1109/MSP.2011.940267.

[3] F. Schaich and T. Wild, "Waveform contenders for 5G: OFDM vs. FBMC vs. UFMC," in *Proc. 6th International Symposium on Communications, Control and Signal Processing*, Athens, Greece, May 2014, pp. 457–460. doi: 10.1109/ISCCSP.2014.6877912.

[4] M. Kasparick, G. Wunder, P. Jung, *et al.*, "Bi-orthognal waveforms for 5G random access with short message support", in *Proc. 20th European Wireless Conference*, Barcelona, Spain, May 2014, pp. 1–6.

[5] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications*, London, United Kingdom, Sept. 2013, pp. 332– 336. doi: 10.1109/PIMRC.2013.6666156.

[6] Y. Saito, Y. Kishiyama, A. Benjebbour, *et al.*, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in Proc. IEEE 77th Vehicular Technology Conference, Dresden, Germany, Jun. 2013, pp.1–5. doi: 10.1109/VTCSpring.2013.6692652.

[7] F.-L. Luo, *Digital Front-End in Wireless Communications and Broadcasting: Circuits and Signal Processing*, Cambridge, England: Cambridge University Press, Nov. 2011.

[8] Z. Kollár and P. Horváth, "PAPR reduction of FBMC by clipping and its iterative compensation," *Journal of Computer Networks and Communications*, vol. 2012, article ID 382736.

[9]. R. Ayadi, M. Siala, and I. Kammoun, "Transmit/receive pulse-shaping design in BFDM systems over time-frequency dispersive AWGN channel," *Proc. of IEEE International Conference on Signal Processing and Communications*, Dubai, UAE, 2007, PP. 772–775. doi: 10.1155/2012/382736.

[10] C. Lélé, P. Siohan, and R. Legouable, "The alamouti scheme with CDMA-OFDM/OQAM," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, article no. 2. doi:10.1155/2010/703513.

[11] N. Michailow, M. Matthé, I. S. Gaspar, *et al.*, "Generalized frequency division multiplexing for 5th generation cellular networks," *IEEE Transactions on Communications*, vol. 62, no.9, pp. 3045– 3061, Sept. 2014. doi: 10.1109/TCOMM.2014.2345566.

[12] R. Hoshyar, R. Razavi, and M. AL-Imari, "LDS-OFDM an efficient multiple access technique," in *Proc. IEEE 71st Vehicular Technology Conference*, VTC-Spring, May 2010, pp. 1–5. doi: 10.1109/VETECS.2010.5493941.

[13] M.Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE 80th Vehicular Technology Conference*, Vancouver, Canada, Sept. 2014.

[14] F.-L. Luo and R. Unbehauen, *Applied Neural Networks for Signal Processing*, Cambridge, England: Cambridge University Press, 1997.

[15] F.-L. Luo, *Mobile Multimedia Broadcasting Standards: Technology and Practice*, Berlin, Germany: Spring Verlag, 2008.

[16] F.-L. Luo, W. Williams, M. R. Rao, *et al.*, "Trends in signal processing applications and industry technology," *IEEE Signal Processing Magazine*, vol. 29, no.1, pp.174–184, Jan. 2012. doi: 10.1109/MSP.2011.943129.

[17] L. Lu, G. Y. Li, A. L. Swindlehurst, and A. Ashikhmin, "An overview of massive MIMO: benefits and challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no.5, pp.742– 758, Oct. 2014. doi: 10.1109/JSTSP.2014.2317671.

[18] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, Feb. 2014. doi: 10.1109/MCOM.2014.6736761.

[19] F. Rusek, D. Persson, B. K. Lau, *et al.*, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no.1, pp. 40–60, Jan. 2013. doi: 10.1109/MSP.2011.2178495.

[20] S. Mohammad-Razavizadeh, M. Ahn, and I. Lee, "Three-dimensional beamforming: a new enabling technology for 5G wireless networks," *IEEE Signal Processing Magazine*, vol. 31, no.6, pp. 94– 101, Nov. 2014. doi: 10.1109/MSP.2014.2335236.

[21] Y. Li, X. Ji, and D. Liang, "Dynamic beamforming for three-dimensional MIMO technique in LTE-advanced networks," *International Journal of Antennas and Propagation*, vol. 2013, article ID 764507. doi:10.1155/2013/764507.

[22] M.-T. Dao, V.-A. Nguyen, Y.-T. Im, *et al.*, "3D polarized channel modeling and performance comparison of MIMO antenna configurations with different polarizations," *IEEE Transactions on Antennas and Propagation*, vol. 59, no. 7, pp. 2672–2682, Jul. 2011. doi: 10.1109/TAP.2011.2152319.

[23] W. Chen, M. M. Tentzeris, Y. Yao, *et al.*, "MIMO antenna design and channel modeling," *International Journal of Antennas and Propagation*, vol. 2013, article ID 381081. doi:10.1155/2013/3810811.

[24] O. Edfors and A. J. Johansson, "Is orbital angular momentum (OAM) based radio communication an unexploited area?" *IEEE Transactions on Antennas and Propagation*, vol. 60, no.2, pp. 1126– 1131, Feb. 2012. doi: 10.1109/TAP.2011.2173142.

[25] Y. Yan, G. Xie, M. P. J. Lavery, *et al.*, "High-capacity millimeter wave communications with orbital angular momentum multiplexing," *Nature Communications*, vol. 5, article no. 4876, 2014, doi:10.1038/ncomms5876.

[26] S. M. Mohammadi, L. K. S. Daldorff, J. E. S. Bergman, *et al* "Orbital angular momentum in radio—a system study," *IEEE Transactions on Antennas and Propagations*, vol. 58, no.2, pp. 565– 572, Feb. 2010. doi: 10.1109/TAP.2009.2037701.

[27] T. S. Rappaport, S. Shu, R. Mayzus, and Z. Hang, "Millimeter wave mobile communications for 5G cellular: it will work," *IEEE Access*, vol.1, pp. 335-349, 2013. doi: 10.1109/ACCESS.2013.2260813.

[28] S. Hong, J. Brand, C. Jung, *et al.*, "Applications of self-interference cancellation in 5G and beyond," *IEEE Communications Magazine*, vol.52, no.2, pp. 114–121, Feb. 2014. doi: 10.1109/MCOM.2014.6736751.

[29] T. Irnich, J. Kronander, Y. Selen, and L. Gen, "Spectrum sharing scenarios and resulting technical requirements for 5G systems," in *Proc. IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications*, London, United Kingdom, Sept. 2013, pp. 127– 132. doi: 10.1109/PIMRCW.2013.6707850.

[30] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Fronthaul compression for cloud radio access networks: signal processing advances inspired by network information theory," *IEEE Signal Processing Magazine*, vol. 31, no.1, pp. 69–79, Nov. 2014. doi: 10.1109/MSP.2014.2330031.

[31] Y.-H. Nam, B. L. Ng, K. Sayana, *et al.*, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Communications Magazine*, vol.51, no.6, pp.172–179, Jun. 2013. doi: 10.1109/MCOM.2013.6525612.

[32] B. L. Ng, Y. Kim, J. Lee, *et al.*, "Fulfilling the promise of massive MIMO with 2D active antenna array," in *Proc. IEEE Globecom Workshops*, Anaheim, USA, Dec. 2012, pp. 691–696. doi: 10.1109/GLOCOMW.2012.6477658.

[33] B. Monday, T. Thomas, E. Visotsky, *et al.*, "3D Channel Model in 3GPP", *IEEE Communications Magazine*, 2015 (to appear).

[34] F. Khan, Z. Pi, and J. Zhang, "Techniques for millimeter wave mobile communication", US Patent App. 12/916,019, 2010.

[35] J. G. Andrews, S. Buzzi, W. Choi, *et al.*, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol.32, no.6, pp. 1065– 1082, Jun. 2014. doi: 10.1109/JSAC.2014.2328098.

Manuscript received: 2014-08-08

Biography

Fa-Long Luo (f.luo@ieee.org) is chief scientist at Element CXI, California. He was the founding editor-in-chief of *International Journal of Digital Multimedia Broadcasting*. He was also the chairman of the IEEE Industry DSP Standing Committee and technical board member of IEEE SPS from 2011 to 2012. He is an elected fellow of the IET. He has been an associate editor of a number of IEEE periodicals, including *IEEE SP Magazine* and *IEEE IoT Journal*. He has received many international recognitions in related fields and has published four books, more than 100 technical papers, and 18 patents.

DOI: 10.3969/j. issn. 1673-5188. 2015. 01. 004
http://www.cnki.net/kcms/detail/34.1294.TN.20150212.0858.001.html, published online February 12, 2015

Energy-Efficient Large-Scale Antenna Systems with Hybrid Digital-Analog Beamforming Structure

Shuangfeng Han[1], Chih-Lin I[1], Zhikun Xu[1], Qi Sun[1], and Haibin Li[2]
(1. Green Communication Research Center, China Mobile Research Institute, Beijing 100053, China;
2. Department of Planning and Construction, China Mobile, Beijing 100053, China)

Abstract

A large-scale antenna system (LSAS) with digital beamforming is expected to significantly increase energy efficiency (EE) and spectral efficiency (SE) in a wireless communication system. However, there are many challenging issues related to calibration, energy consumption, and cost in implementing a digital beamforming structure in an LSAS. In a practical LSAS deployment, hybrid digital-analog beamforming structures with active antennas can be used. In this paper, we investigate the optimal antenna configuration in an $N \times M$ beamforming structure, where N is the number of transceivers, M is the number of active antennas per transceiver, where analog beamforming is introduced for individual transceivers and digital beamforming is introduced across all N transceivers. We analyze the green point, which is the point of maximum EE on the EE-SE curve, and show that the log-scale EE scales linearly with SE along a slope of -lg2/N. We investigate the effect of M on EE for a given SE value in the case of fixed NM and independent N and M. In both cases, there is a unique optimal M that results in optimal EE. In the case of independent N and M, there is no optimal (N, M) combination for optimizing EE. The results of numerical simulations are provided, and these results support our analysis.

Keywords

digital beamforming; analog beamforming; hybrid beamforming; energy efficiency; spectral efficiency

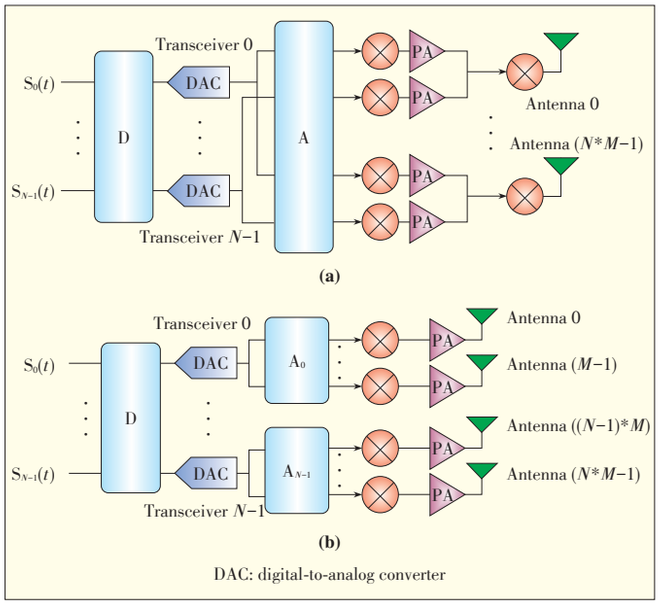
1 Introduction

Wireless communication systems have developed from first generation to fourth generation to accommodate ever-increasing and diversified mobile traffic. The anticipated thousand-fold increase in wireless traffic by 2020 and the push for green communication worldwide create some very tough challenges for 5G system design [1]. Massive MIMO, also known as large-scale antenna system (LSAS) [2], [3], is a promising green 5G communication scheme that improves both energy efficiency (EE) and spectral efficiency (SE). With full digital beamforming (BF) LSAS can, in theory, perform optimally. When many antennas are implemented to increase beamforming gain, it may not be feasible to implement the same number of transceivers because of excessive demand on real-time signal processing when there is large BF gain [4] and also because of cost and power consumption, especially for mixed-signal devices in a millimeter-wave system. A beamforming structure with a much smaller number of digital transceivers than antennas is therefore more practical and cost-effective.

To reduce complexity in a LSAS, analog beamforming with active antennas can be considered [5], [6]. With analog beamforming, the signal phase on each antenna is controlled by a network of analog phase shifters. In [7]–[9], hybrid analog-digital beamforming strategies were investigated for pre-coding multiple data streams and increasing beamforming gain. In [7], the transmitted signal on each of N digital transceivers travels along all NM RF paths (mixer, PA, and phase shifter), where M is the number of active antennas per transceiver. The signal is summed up before being connected with each antenna element (**Fig. 1a**). Analog beamforming is then introduced over NM RF paths per transceiver, and digital beamforming is introduced over N digital transceivers. The complexity of this structure is high.

In **Fig. 1b**, the $N \times M$ hybrid beamforming structure has N transceivers connected to M antennas. This structure is more practical for base station antenna deployment in 3G and 4G LTE systems, where each transceiver is connected to a column of antennas. With active antennas on each RF path, elevation beamforming can be introduced by applying different phases to each antenna in each column.

Recently, there has been growing interest in hybrid beamforming design. The structure in Fig. 1a features a precoding



▲**Figure 1. Hybrid beamforming structure**

solution where only some aspects of the channel (e.g., angle of arrival and departure) are known at the base station and mobile station [7]. The spatial structure of millimeter-wave channels has been further exploited to formulate the single user precoding/combining problem as a sparse reconstruction problem [8]. In [4], the authors propose an angle-of-arrival estimation algorithm and beamforming algorithm. In [9], the authors propose a beam-domain RS design that results in better performance than a design based on pure analog beamforming. An outdoor trial of the $N \times M$ beamforming structure has been carried out in South Korea [10], but the optimal configuration of M remains an open and very important issue. An improper M may reduce EE even if the SE is satisfactory.

In this paper, we focus on the $N \times M$ hybrid beamforming structure. In particular, we investigate the optimization of both EE and SE in the cases of fixed NM and independent N and M. In section 2, we discuss the relationship between EE and SE. Then, we investigate this relationship at the "green" points, i.e., the points of maximum EE on the EE-SE curve. We discuss the effect of M on EE for a given SE. We investigate the optimal (N,M) combination that results in the highest EE in the case of independent N and M. In particular, we discuss the optimal M when there is severe inter-user interference. In section 3, we present and discuss numerical simulation results. In section 4, we draw some conclusions from our analyses.

2 Energy Efficiency and Spectral Efficiency

2.1 Relationship

In the $N \times M$ structure in Fig. 1b, perfect analog beamforming is assumed within M antennas per transceiver, which points to one user (there are N users in total). Assuming there is no inter-user interference, i.e., there is proper user scheduling (the BS schedules users with orthogonal channels), then the sum capacity of this structure for N users is:

$$C = W \times N \times \log\left(1 + \frac{M\eta_{PA}P}{WN_0}\right) \tag{1}$$

where W is the bandwidth, P is transmit power of each transceiver (the total power of M antenna PAs), η_{PA} is the PA efficiency, and N_0 is the thermal noise density. Without loss of generality, the channel gain is assumed to be the unity. The SE of this structure is

$$\eta_{SE} = C/W = N \times \log\left(1 + \frac{M\eta_{PA}P}{WN_0}\right). \tag{2}$$

Because the accurate power model is non-trivial, the following simple power model is used:

$$P_{total} = NP + P_{static} = NP + NP_0 + P_{common} + NMP_{rf\_circuit} \tag{3}$$

where P_{total} is the total power; NP is the RF power of N transceivers; P_{static} is the static power of the BS, including NP_0, which scales with N; P_{common}, which is common for any number of transceivers; and $NMP_{\text{rf\_circuit}}$, which scales with NM . The relationship between EE and SE is

$$\eta_{EE} = C/P_{total} = \frac{\eta_{SE}}{\left(2^{\frac{\eta_{SE}}{N}} - 1\right)\frac{N_0}{\eta_{PA}}\frac{N}{M} + \frac{NP_0 + P_{common} + NMP_{rf\_circuit}}{W}} \tag{4}$$

Therefore, for a required SE, the hybrid LSAS beamforming should be designed to maximize EE through joint design of N, M, P_0, P_{common}, $P_{\text{rf\_circuit}}$ and η_{PA}. This paper focuses on the design for an optimal number of active antennas per transceiver M^* that ensures the best EE for a given SE.

2.2 Relationship at Green Points

When we take the circuit power into consideration, there is a "green" point on the EE-SE curve where EE is at its maximum and is denoted η^*_{EE} [11]. Here, we discuss two cases for the $N \times M$ hybrid beamforming structure: $NM = L$ (i.e., the total number of antennas is fixed as L, but N and M are variable), and 2) N and M are independent. In the former case, we allow the first-order derivative of EE over SE to be zero:

$$\eta'_{EE} = \frac{aN^2\left(2^{\frac{\eta_{SE}}{N}} - 1\right) + bN + c - \eta_{SE}aN2^{\frac{\eta_{SE}}{N}}\ln 2}{\left(aN^2\left(2^{\frac{\eta_{SE}}{N}} - 1\right) + bN + c\right)^2} = 0 \tag{5}$$

Combining (5) with (4), the relationship between the η^*_{EE} and corresponding SE η^*_{SE} is

$$\eta^*_{EE}=\left(\frac{n_0 N 2^{\frac{\eta^*_{se}}{N}}\ln 2}{L\eta_{PA}}\right)^{-1}. \quad (6)$$

The relationship between η^*_{EE} and η^*_{SE} is further given as

$$\lg\left(\eta^*_{EE}\right)=-\frac{\lg 2}{N}\eta^*_{SE}+\lg\left(\frac{L\eta_{PA}}{n_0 N\ln 2}\right), \quad (7)$$

which indicates that $lg(\eta^*_{EE})$ scales linearly with η^*_{SE} and has a slope of -lg2/N. Similar to the EE-SE relationship in classic Shannon theory, higher η^*_{SE} always leads to lower η^*_{EE}. The relationship between η^*_{EE} and η^*_{SE} does not depend on P_0, P_{common}, $P_{rf\_circuit}$ and W, although from (4), we see that η^*_{SE} and η^*_{EE} are based on all the other parameters.

It is expected, therefore, that the system operates at the green point. Also, it is important that η^*_{SE} satisfies the system SE requirement, and η^*_{EE} should be high enough. This requires careful design of P_0, P_{common}, $P_{rf\_circuit}$, W, η_{PA}, N and M. For example, when other parameters are given, M can be designed to maximize EE.

2.3 Optimal M for Maximizing EE for a Given SE

2.3.1 When N and M are Independent

It is of practical importance to know how M affects EE for a given SE. If there is one optimal M that results in the highest EE, it is not necessary to deploy too many antennas per transceiver. In the following, we derive the optimal M to maximize EE. We denote the denominator of (4) as $f(M)$:

$$f(M)=\left(2^{\frac{\eta_{se}}{N}}-1\right)\frac{N_0}{\eta_{PA}}\frac{N}{M}+\frac{NP_0+P_{common}+NMP_{rf\_circuit}}{W}. \quad (8)$$

The first- and second-order derivatives of $f(M)$ are

$$f'(M)=\frac{NP_{rf\_circuit}}{W}-\left(2^{\frac{\eta_{se}}{N}}-1\right)\frac{N_0}{\eta_{PA}}\frac{N}{M^2} \quad (9)$$

and

$$f''(M)=2\left(2^{\frac{\eta_{se}}{N}}-1\right)\frac{N_0}{\eta_{PA}}\frac{N}{M^3}\geq 0. \quad (10)$$

Then $f(M)$ is a quasi-convex function of M. The M^* that gives the minimum $f(M)$ is derived by making $f'(M)=0$:

$$M^*=\sqrt{\frac{WN_0}{\eta_{PA}P_{rf\_circuit}}\left(2^{\frac{\eta_{se}}{N}}-1\right)}. \quad (11)$$

Because of the definition of η_{EE} in (4), EE is a quasi-concave function of M, and the EE is at its maximum when $M=M^*$. When $M\leq M^*$, EE monotonically increases with M. When $M>M^*$, EE monotonically decreases with M. In practical system design, for a given SE there is one M* that results in the highest EE. As in (11), M* increases with SE and bandwidth, but decreases with PA power efficiency and $P_{rf\_circuit}$. For a given number of transceivers N, more antennas per transceiver are needed for higher SE. If W increases, the noise power increases correspondingly, and a larger M is needed to achieve the SE. A larger $P_{rf\_circuit}$, however, reduces M* because the increased circuit power may reduce EE.

2.3.2 When NM is Fixed

Assume $NM = L$, the denominator of (4) is written as

$$f(M)=\left(2^{\frac{M}{L}\eta_{se}}-1\right)\frac{N_0}{\eta_{PA}}\frac{L}{M^2}+\frac{LP_0/M+P_{common}+LP_{rf\_circuit}}{W}. \quad (12)$$

For simplicity of derivation, $f(M)$ is rewritten as

$$f(M)=a\left(2^{\frac{M}{L}\eta_{se}}-1\right)\frac{1}{M^2}+\frac{b}{M}+c \quad (13)$$

where $a=\frac{N_0 L}{\eta_{PA}}, b=\frac{LP_0}{W}$, $c=\frac{P_{common}+LP_{rf\_circuit}}{W}$.

The first-order derivative of $f(M)$ is

$$\begin{aligned}f'(M)&=M^{-3}\left(a2^{\frac{M}{L}\eta_{se}}\frac{\eta_{SE}}{L}M\ln 2-2a\left(2^{\frac{M}{L}\eta_{se}}-1\right)-bM\right)\\&=M^{-3}g(M)\end{aligned} \quad (14)$$

where $g(M)=\left(a2^{\frac{M}{L}\eta_{se}}\frac{\eta_{SE}}{L}M\ln 2-2a\left(2^{\frac{M}{L}\eta_{se}}-1\right)-bM\right)$.

Respectively, the first- and second-order derivatives of $g(M)$ are

$$g'(M)=a2^{\frac{M}{L}\eta_{se}}\left(\frac{\eta_{SE}}{L}\right)^2M(\ln 2)^2-a2^{\frac{M}{L}\eta_{se}}\frac{\eta_{SE}}{L}\ln 2-b \quad (15)$$

and

$$g''(M)=a2^{\frac{M}{L}\eta_{se}}\left(\frac{\eta_{SE}}{L}\right)^3M(\ln 2)^3\geq 0. \quad (16)$$

Therefore, $g'(M)$ monotonically increases with M. In addition,

$$g'(0)=-\frac{a\eta_{SE}}{L}\ln 2-b \quad (17)$$

$$g'(\infty)=\infty. \quad (18)$$

Thus, there is unique positive M_0 so that $g'(M_0) = 0$. When $M < M_0$, $g'(M) < 0$, $g(M)$ monotonically decreases with M. When $M > M_0$, $g'(M) > 0$, $g(M)$ monotonically increases with M. Because $g(0) = 0$ and $g(\infty)=\infty$, there is a unique positive M_1 that is larger than M_0 and satisfies $g(M_1) = 0$. From (14), g

(M) determines monotonicity of $f(M)$; therefore, when $M < M_1$, $g(M) < 0$, $f'(M) < 0$. When $M > M_1$, $g(M) > 0$, $f'(M) > 0$. Also,

$$f'(0) = \lim_{M\to 0} \frac{a2^{\frac{M}{L}\eta_{se}} \frac{\eta_{SE}}{L} M \ln 2 - 2a\left(2^{\frac{M}{L}\eta_{se}} - 1\right) - bM}{M^3} = \frac{a}{6}\left(\frac{\eta_{SE}\ln 2}{L}\right)^3 \geq 0 \quad (19)$$

Therefore, when $M \leq M_1$, EE monotonically increases with M. When $M > M_1$, EE monotonically decreases with M. In the case of fixed MN, EE is maximum at $M = M_1$, where M_1 can be obtained by solving $g(M_1) = 0$.

2.3.3 Optimal (N, M) Combination

An important issue is finding the N and M that results in the highest EE for a given SE when N and M are not fixed. Combining (11) and (4) the maximum EE for a given SE and N is

$$\eta_{EE}^*(N,\eta_{SE}) = \frac{W\eta_{SE}}{2N\sqrt{\frac{WN_0P_{rf\_circuit}}{\eta_{PA}}}\sqrt{\left(2^{\frac{\eta_{se}}{N}} - 1\right)} + NP_0 + P_{common}} \quad (20)$$

The optimal N can then be calculated:

$$N^* = \arg\max\left(\eta_{EE}^*(N,SE)\right), N \geq 1 \ . \quad (21)$$

We denote the denominator in (4) $f(N,M)$:

$$f(N,M) = \left(2^{\frac{\eta_{se}}{N}} - 1\right)\frac{N_0}{\eta_{PA}}\frac{N}{M} + \frac{NP_0 + P_{common} + NMP_{rf\_circuit}}{W} . \quad (22)$$

There is no extreme point for $f(N,M)$. The partial derivative of $f(N,M)$ over M is

$$\frac{\partial f}{\partial M} = \frac{NP_{rf\_circuit}}{W} - \left(2^{\frac{\eta_{se}}{N}} - 1\right)\frac{N_0}{\eta_{PA}}\frac{N}{M^2} = 0 \quad (23)$$

which leads to

$$2^{\frac{\eta_{se}}{N}} = \frac{M^2\eta_{PA}P_{rf\_circuit}}{WN_0} + 1 \quad (24)$$

The partial derivative of $f(N,M)$ over N is

$$\frac{\partial f}{\partial N} = \frac{MP_{rf\_circuit} + P_0}{W} + \frac{N_0}{\eta_{PA}}\frac{1}{M}\left(\left(2^{\frac{\eta_{se}}{N}} - 1\right) - 2^{\frac{\eta_{se}}{N}}\frac{\eta_{SE}}{N}\ln 2\right) = 0 \quad (25)$$

which leads to

$$\frac{MP_{rf\_circuit} + P_0}{W} + \frac{N_0}{\eta_{PA}}\frac{1}{M}\left(\frac{M^2\eta_{PA}P_{rf\_circuit}}{WN_0} - \left(\frac{M^2\eta_{PA}P_{rf\_circuit}}{WN_0} + 1\right)\frac{\eta_{SE}}{N}\ln 2\right) = 0. \quad (26)$$

The optimal M and N should satisfy (24) and (26). Combining (24) and (26), we get

$$M^2\eta_{PA}P_{rf\_circuit}(2N - \eta_{SE}\ln 2) + \eta_{PA}MNP_0 - WN_0\eta_{SE}\ln 2 = 0 \ . \quad (27)$$

This is equivalent to

$$M = \frac{-\eta_{PA}NP_0 + \sqrt{(\eta_{PA}NP_0)^2 - 4\eta_{PA}P_{rf\_circuit}(2N - \eta_{SE}\ln 2)WN_0\eta_{SE}\ln 2}}{2\eta_{PA}P_{rf\_circuit}(2N - \eta_{SE}\ln 2)} \quad (28)$$

However, M in (28) cannot not exist because when $2N - \eta_{SE}\ln 2 > 0$, $M < 0$, and when $2N - \eta_{SE}\ln 2 < 0$, $M < 0$. This is not feasible because M must be positive.

2.3.4 When Inter-User Interference is Taken into Account

In subsection 2.3.3, it is assumed there is no inter-user interference. However, in practical systems, inter-user interference may exist. For simplicity, we assume that interference from the kth beam to the nth beam is $M\eta_{PA}P\alpha_{k,n}$. Then, EE can be expressed as

$$\eta_{EE} = \frac{W\sum_{n=0}^{N-1}\log\left(1 + \frac{M\eta_{PA}P}{WN_0 + \sum_{k\in[0,N-1],k\neq n} M\eta_{PA}P\alpha_{k,n}}\right)}{NP + NP_0 + P_{common} + NMP_{rf\_circuit}} \quad (29)$$

Note that $\alpha_{k,n}$ can be a function of N and M. For example, consider a linear antenna array with NM elements, where the antenna spacing is half a wavelength. The main beam direction (azimuth) of the analog beamforming for the nth transceiver is $\phi_n = n\Delta/N$, $n=0,\cdots,N$-1, and N users are located on the N different main beam directions with same channel gain. We approximate $\alpha_{k,n}$:

$$\alpha_{k,n} = \frac{1}{M^2}\left|\frac{1 - \exp\left(jM\pi(\cos\phi_n - \cos\phi_k)\right)}{1 - \exp\left(j\pi(\cos\phi_n - \cos\phi_k)\right)}\right|^2 = \frac{1}{M^2}\left|\frac{1 - 2\cos\left(2\pi M\sin\frac{(n+k)\Delta}{2N}\sin\frac{(n-k)\Delta}{2N}\right)}{1 - 2\cos\left(2\pi\sin\frac{(n+k)\Delta}{2N}\sin\frac{(n-k)\Delta}{2N}\right)}\right| . \quad (30)$$

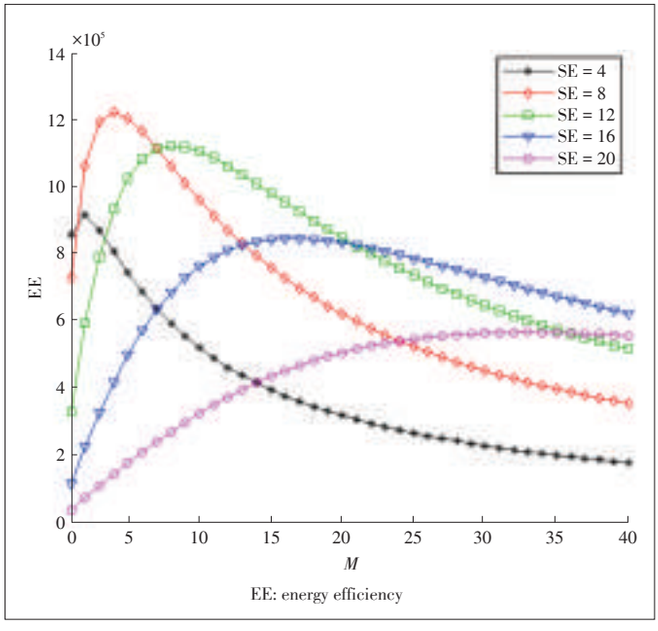
It seems difficult to determine how M affects EE in the cases that fixed NM and independent N and M are used. In some special cases, for example, in the interference-limited region, increasing the transmit power P does not improve spectral efficiency and actually reduces energy efficiency. When inter-us-

er interference is negligible, the analysis in previous subsections holds.

3 Simulation Results

3.1 *M* vs EE When There is No Inter-User Interference

Assume P_{common} = 50 W, $P_{rf\_circuit}$ = 1 W, P_0 = 1 W, $W = 2 \times 10^7$ Hz, $N_0 = 10^{-17}$ dBm/Hz, η_{PA} = 0.375 and the channel gain is -100 dB. **Fig. 2** shows the effect of M on EE for N = 2 and

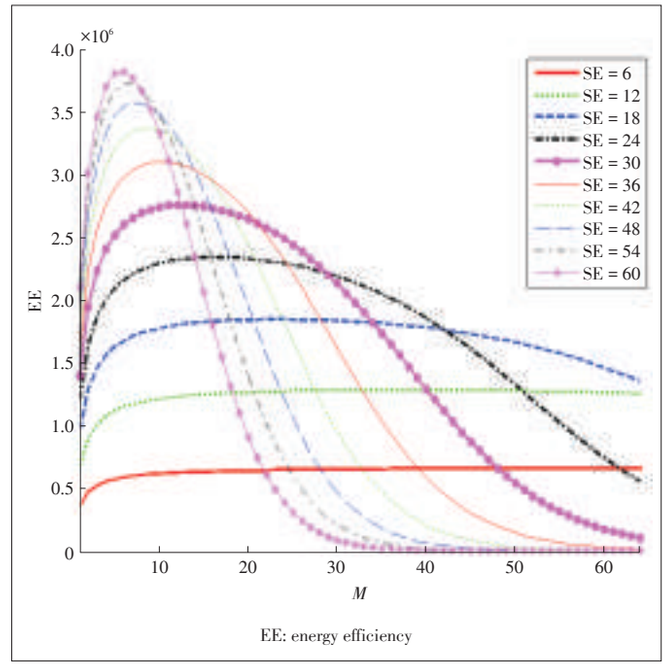


▲**Figure 2. *M* vs EE with different SE values (*N* = 2).**

where M is variable. Five spectral efficiencies between 4 bps/Hz to 20 bps/Hz are simulated. On each M versus EE curve, there is a unique M that results in the highest EE. For example, when SE is 20 bps/Hz, the M^* is 33. When SE is 12bps/Hz, M^* is 8. When M is smaller than optimal, more antennas per transceiver improve EE by providing beamforming gain. When M is greater than optimal, the extra power in the circuit needed by more antennas per transceiver negates any reduction in transmit power so that EE is reduced.

Fig. 3 shows the effect of M on EE when NM is fixed, e.g., NM = 128, and other parameters are the same as those in Fig. 2. Actually, M can only be 1, 2, 4, 8, 16, 32, 64 and 128 (not shown), because N and M are both integers. As in the case of independent N and M, there is a unique M on each curve that results in the highest EE. For example, when SE is 48 bps/Hz, the optimal M is 8. In Fig. 2, the optimal M increases as SE increases; however, in Fig. 3, the optimal M increases as SE decreases. The reason for this is: as SE increases, more transceivers are needed to make the system more energy efficient, and a smaller M (M=L/N) is required.

The above analysis can be referred to when designing an optimal LSAS. In a practical system, the required SE may vary according to the traffic load and service types. For example, in Fig. 2, the M^* for a maximum required SE of 20 bps/Hz is 33. However, when the required SE is reduced to 12 bps/Hz, M^* is 8. Therefore, it is important that, in the case of independent N and M, the system is designed with the largest M^* for the possible SE range, and the best M is chosen according to the SE requirement via antenna on/off. This can help increase EE according to system traffic load.

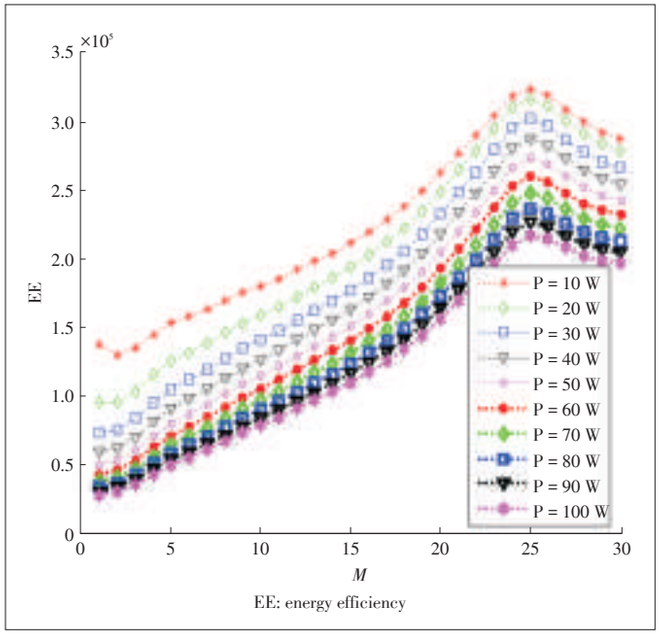


▲**Figure 3. *M* vs EE for different SE values (*NM* = 128).**

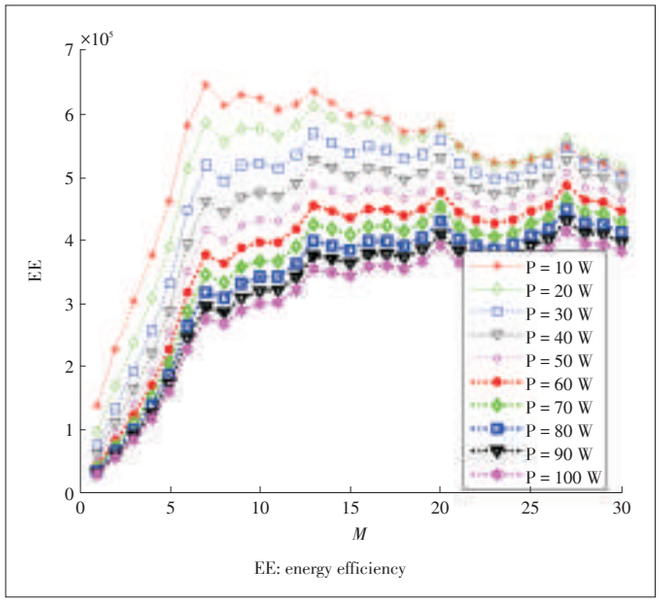
3.2 *M* vs EE When There is Inter-User Interference

We assume P_{common} = 50W, $P_{rf\_circuit}$ = 1 W, P_0 = 10 W, $W = 2 \times 10^7$ Hz, $N_0 = 10^{-17}$dBm/Hz, η_{PA} = 0.375, and channel gain is 10^{-10}. A linear antenna array with N = 10 and half-wavelength antenna spacing is considered. The effect of M on EE is shown in **Fig. 4**, and the inter-user interference is calculated according to (30) (where $\Delta = \pi/3$). Ten power levels between P = 10 W to P = 100 W are simulated. One power level corresponds to one SE value. At each power (and corresponding SE) level, increasing M from 1 to 25 increases EE, but if M goes beyond 25, EE decreases. This is quite different from when there is no inter-user interference and M^* is generally different for different SE values. The possible reason for this is that the coverage of N (N = 10) beams is only $\Delta(\pi/3)$, and there is too much inter-beam interference. Therefore, M has to be large enough to reduce this interference and increase EE.

When Δ increases to π, the beam spacing increases from $\pi/30$ to $\pi/10$, resulting in less inter-beam (inter-user) interference. **Fig. 5** shows the effect of M on EE. The trend is similar to that in the case of no inter-user interference: at each pow-



▲**Figure 4. *M* vs EE with different power levels.**



▲**Figure 5. *M* vs EE for different power levels.**

er level, there is one M^* that results in the highest EE. Inter-user interference can be mitigated via digital precoders whose design is based on certain channel assumptions and that result in increased EE. One straightforward method is to use beam domain downlink reference signals via analog beamforming to estimate 1) the angle of departure of each user, 2) the effective channel with analog beamforming, and 3) the inter-beam (inter-user) interference. Then, digital precoding can further increase the multiuser beamforming gain. EE-SE optimization depends on different multiuser beamforming algorithms and is more complicated, especially in the case of fixed NM.

4 Conclusions

In this paper, an $N \times M$ hybrid analog-digital LSAS beamforming structure is investigated. In this structure, the number of transceivers can be much smaller than the number of antennas. We analyzed the relationship between EE and SE to determine the optimal design of the $N \times M$ beamforming structure. In particular, we analyzed two cases: fixed NM and independent N and M. We analyzed the EE-SE relationship at the green point and showed that the log-scale EE scales linearly with SE along a slope $-\lg 2/N$. In both cases, a unique number of antennas M per transceiver results in the optimal EE for a given SE. In the case of independent N and M, there is no optimal (N, M) combination that results in optimal EE. When inter-user interference is negligible, the above results hold; when there is severe inter-user interference, the optimal M can be quite similar for each SE value. The findings in this paper can be used as guidelines for optimizing an LSAS design.

Acknowledgement

The authors would like to thank the editors and the reviewers for their very helpful comments and review. The authors are also grateful to the team members in the Green Communication Research Center of China Mobile Research Institute.

References

[1] C.-L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Towards green & soft: a 5G perspective," *IEEE Communication Magazine*, vol. 52, no. 2, pp. 66–73, Feb. 2014. doi: 10.1109/MCOM.2014.6736745.

[2] F. Rusek, D. Persson, B. K. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan. 2013. doi: 10.1109/MSP.2011.2178495.

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010. doi: 10.1109/TWC.2010.092810.091092.

[4] X. Huang, Y. J. Guo, and J. D. Bunton, "A hybrid adaptive antenna array," *IEEE Transactions on Wireless Communications*, vol. 9, no. 5, pp.1770–1779, May 2010. doi: 10.1109/TWC.2010.05.091020.

[5] J. Wang, Z. Lan, C. Pyo, T. Baykas, C. Sum, M. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada, and S. Kato, "Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 8, pp. 1390–1399, 2009. doi: 10.1109/JSAC.2009.091009.

[6] V. Venkateswaran and A.-J. van der Veen, "Analog beamforming in MIMO communications with phase shift networks and online channel estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 8, pp. 4131–4143, 2010. doi: 10.1109/TSP.2010.2048321.

[7] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Hybrid precoding for millimeter wave cellular systems with partial channel knowledge," in *Information Theory and Applications Workshop*, San Diego, USA, Feb. 2013, pp. 1–5. doi: 10.1109/ITA.2013.6522603.

[8] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014. doi: 10.1109/TWC.2014.011714.130846.

[9] S. Han, C.-L. I, Z. Xu, and S. Wang "Reference signals design for hybrid analog and digital beamforming," *IEEE Communications Letters*, vol. 18, no. 7, pp. 1191–1193, Jul. 2014. doi: 10.1109/LCOMM.2014.2317747.
[10] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results," *IEEE Communication Magazine*, vol. 52, no. 2, pp. 106–113, Feb. 2014. doi: 10.1109/MCOM.2014.6736750.
[11] G. Y. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient wireless communications: tutorial, survey, and open issues," *IEEE Wireless Communications*, vol.18, no.6, pp. 28–35, Dec. 2011. doi: 10.1109/MWC.2011.6108331.

Manuscript received: 2014-09-15

Biographies

Shuangfeng Han (hanshuangfeng@chinamobile.com) received his MS and PhD degrees in electrical engineering from Tsinghua University in 2002 and 2006. He joined Samsung Electronics as a senior engineer in 2006 and worked on MIMO and MultiBS MIMO. Since 2012, he has been a senior project manager in the Green Communication Research Center of China Mobile Research Institute. His research interests include green 5G, massive MIMO, full-duplex, non-orthogonal multiple access, energy efficiency, and spectral efficiency co-design.

Chih-Lin I (icl@chinamobile.com) received her PhD degree in electrical engineering from Stanford University. She has worked for numerous world-class companies and research institutes, including AT&T Bell Labs, AT&T HQ, ITRI Taiwan, and ASTRI Hong Kong. She was awarded the Stephen Rice Best Paper Award from *IEEE Tranactions on Communications* and is a winner of the CCCP National 1000 Talent program. Currently, she is China Mobile's chief scientist of wireless technologies and has established the Green Communications Research Center, spearheading major initiatives including key 5G technology R&D; high EE system architectures, technologies and devices; green energy; and C-RAN and soft base stations. She was an elected Board Member of IEEE ComSoc, Chair of the ComSoc Meetings and Conferences Board, and Founding Chair of the IEEE WCNC Steering Committee. She is currently an Executive Board Member of GreenTouch and a Network Operator Council Member of ETSI NFV. Her research interests are green communications, C-RAN, network convergence, bandwidth refarming, EE-SE co-design, massive MIMO, and active antenna arrays.

Zhikun Xu (xuzhikun@chinamobile.com) received his BSE and PhD degrees in signal and information processing from Beihang University (BUAA), China in 2007 and 2013. After graduation, he joined the Green Communication Research Center of China Mobile Research Institute as a project manager. His current interests include green technologies, cross-layer resource allocation, advanced signal processing, and transmission techniques

Qi Sun (sunqiyjy@chinamobile.com) received her BSE and PhD degrees in information and communication engineering from Beijing University of Posts and Telecommunications in 2009 and 2014. After graduation, she joined the Green Communication Research Center of China Mobile Research Institute. Her research interests include MIMO, cooperative communication, and green communications.

Haibin Li (lihaibin@chinamobile.com) received her MS degree in project management from Beijing University of Posts and Telecommunications, China. She is currently the director of Division of Energy Conservation and Emission Reduction, Department of Planning and Construction, China Mobile Communications Corporation. She has been in charge of energy saving and emission reduction from April 2011 and has 19 years experience in the field of communications planning and construction. She is also the director of Resource Sharing Plan for CMCC and deputy head of the CCSA ST2 and ST4 group.

Roundup

Introduction to *ZTE Communications*

ZTE Communications is a quarterly, peer-reviewed international technical journal (ISSN 1673–5188 and CODEN ZCTOAK) sponsored by ZTE Corporation. The journal publishes original academic papers and research findings on the whole range of communications topics, including communications and information system design, optical fiber and electro-optical engineering, microwave technology, radio wave propagation, antenna engineering, electromagnetics, signal and image processing, and power engineering. The journal is designed to be an integrated forum for university academics and industry researchers from around the world. *ZTE Communications* was founded in 2003 and has a readership of 5500. The English version is distributed to universities, colleges, and research institutes in more than 140 countries. It is listed in Inspec, Cambridge Scientific Abstracts (CSA), Index of Copernicus (IC), Ulrich's Periodicals Directory, Norwegian Social Science Data Services (NSD), Chinese Journal Fulltext Databases, Wanfang Data — Digital Periodicals, and China Science and Technology Journal Database. Each issue of *ZTE Communications* is based around a Special Topic, and past issues have attracted contributions from leading international experts in their fields.

DOI: 10.3969/j. issn. 1673-5188. 2015. 01. 005
http://www.cnki.net/kcms/detail/34.1294.TN.20150216.1537.002.html, published online February 16, 2015

An Optimal Lifetime Utility Routing for 5G and Energy-Harvesting Wireless Networks

Gina Martinez[1], Shufang Li[2], and Chi Zhou[1]
(1. Department of Electrical and Computer Engineering, Illinois Institute of Technology, IL 60616, USA;
2. Telecommunication Engineering Institute, Beijing University of Posts and Telecommunication, Beijing 100876, China)

Abstract

Harvesting energy from environmental sources such as solar and wind can mitigate or solve the limited-energy problem in wireless sensor networks. In this paper, we propose an energy-harvest-aware route-selection method that incorporates harvest availability properties and energy storage capacity limits into the routing decisions. The harvest-aware routing problem is formulated as a linear program with a utility-based objective function that balances the two conflicting routing objectives of maximum total and maximum minimum residual network energy. The simulation results show that doing so achieves a longer network lifetime, defined as the time-to-first-node-death in the network. Additionally, most existing energy-harvesting routing algorithms route each traffic flow independently from each other. The LP formulation allows for a joint optimization of multiple traffic flows. Better residual energy statistics are also achieved by such joint consideration compared to independent optimization of each commodity.

Keywords

routing; 5G; energy-harvesting; wireless sensor networks

1 Introduction

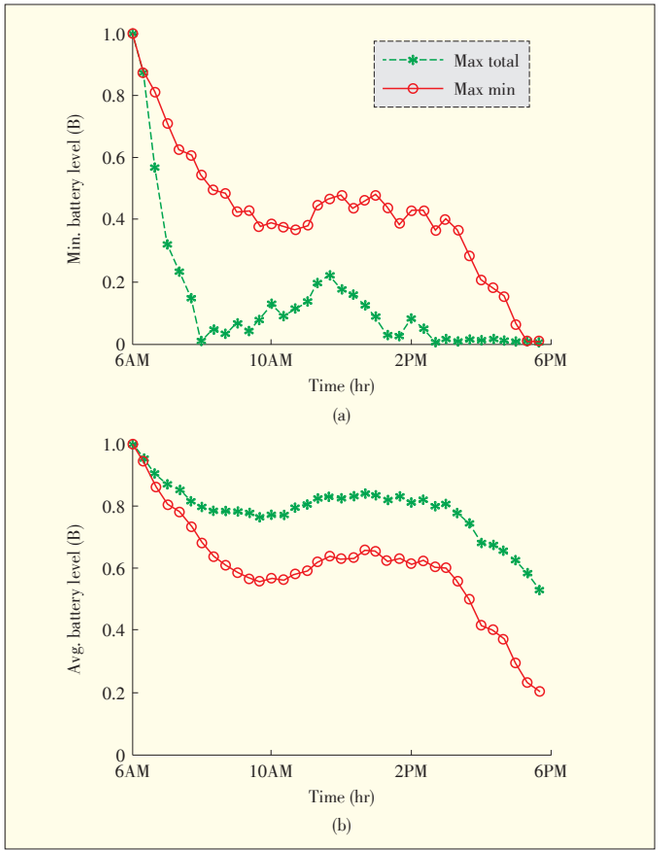
Wireless sensor networks (WSN) continue to be an affordable, convenient solution to a broad range of wireless communication challenges today. However, the limited energy capacity of wireless network components continues to be the main obstacle to WSN application. With research on 5G wireless networks already underway, there is a pressing need to ensure that components have sufficient energy capacity to support next-generation wireless technologies.

5G wireless networks will be characterized by high capacity and high data rate. The trend in wireless services is rapid increase in multimedia traffic and ever-increasing demand for bandwidth. Therefore, energy efficiency and power-saving system services are important in 5G research. Recently, energy harvesting has been more widely studied to satisfy the high energy demands of emerging wireless technologies. In an energy-harvesting wireless sensor network (EH-WSN), nodes are able to harvest energy, e.g., solar, thermal and mechanical, from the surrounding environment.

Two types of technologies make EH-WSN feasible: energy-harvesting and energy-management [1]. An energy-harvesting technology converts environmental energy into electrical energy; e.g., photovoltaic (PV) panels convert solar radiation; piezoelectric transducers convert mechanical stress; and wind turbines convert kinetic wind. Energy management involves intelligently constraining energy consumption while still meeting certain performance and energy objectives, such as maximizing network lifetime or packet delivery rate. Our work focuses on managing energy through harvest-aware routing.

Network lifetime can be defined in several ways. Here, we define it as time-to-first-node-death (TTFND). Maximizing the lifetime of a battery-powered network involves a tradeoff between two objectives: 1) minimizing the total amount of energy consumed in routing the data packets and 2) maximizing the minimum residual network battery level. In an EH-WSN, there is additional energy "consumption" in the form of energy wastage due to overcharge of finite-capacity energy buffers. Minimizing total energy consumption, including wastage, in routing is equivalent to maximizing the total residual energy of the network [2].

Our preliminary simulation results confirm the previously mentioned tradeoff. **Fig. 1a** shows the minimum battery level and **Fig. 1b** shows the average battery level in a network of 25 nodes. In the maximum total remaining (MAX-TOTAL) objective, nodes along shortest-path routes are depleted faster than in maximum minimum remaining (MAX-MIN) because these routes consume the least total energy. When these nodes are depleted, the network may become partitioned, which shortens



▲**Figure 1. (a) Minimum and (b) average battery levels for MAX-TOTAL and MAX-MIN routing.**

the network lifetime. On the other hand, the average battery level of the network for MAX-MIN is less than that for MAX-TOTAL because MAX-MIN may purposely select longer routes in order to avoid low-energy nodes, decreasing the overall available network energy.

In our previous work [2], we discussed the objective of maximizing the total remaining network energy. We proposed a simple yet effective route-selection method for maximizing a lifetime utility function, which seeks to balance the routing objectives of MAX-TOTAL and MAX-MIN energy level in the network. We show that this achieves higher lifetime than either objective alone. This method is applicable online because it takes into account the uncertainty of future traffic rather than requiring offline, deterministic traffic-arrival information.

We investigated the effects of network factors, such as topology, energy consumption, and prediction error, on the energy savings. In doing so, we gained insight into how some network design decisions and characteristics affect target performance and energy requirements.

2 Related Work

There have been several novel works published on routing for EH-WSNs. In [3], an energy-management and routing method that maximizes energy efficiency by modeling the energy buffer as a G/G/1(/*N*) queue is presented. Routes are chosen according to the probability of the buffer depletion of the nodes. The authors of [4] and [5] formulate link-cost metrics based on a node's energy availability, which generally encompasses initial battery and energy harvest. Routes are then found using Bellman-Ford, Directed Diffusion (DD), or other applicable shortest-path or least-cost algorithms with respect to these metrics. Most of the works in [6] discuss this method but not the exploitation of joint optimal routing of multiple commodities in the network.

The tradeoff between minimum energy route consumption and maximum residual energy routing mainly arises from unknown future traffic arrivals. That is, the online traffic arrival is unknown. For this reason, several heuristic online maximum lifetime algorithms for battery-powered WSNs have been investigated. In [7], conditional MAX-MIN battery capacity routing (CMMBCR) is introduced. With CMMBCR, routes that would result in a minimum battery below a threshold are discarded. Among the remaining routes, the minimum consumption route is chosen. Alternatively, in [8], the authors present a MAX-MIN zPmin algorithm. In this algorithm, of the routes with at most zPmin energy consumption, the route resulting in maximum minimum energy is chosen. As previously shown, this tradeoff also exists for EH-WSNs [2], [9]. Here, we introduce a simple maximum lifetime utility function that balances the two objectives.

We formulate multi-commodity routing as an LP problem that additionally takes into account finite battery capacities. Formulating the maximum lifetime routing problem as linear programming (LP) optimization has been well-explored for battery-based wireless sensor networks where the energy source is fixed and non-rechargeable. The authors of [10] provided a basis for flow conservation and energy constraints used in many LP-based routing formulations, including the one in this work. In [11] and [12], the LP formulation takes node operation modes, such as transmit, receive and idle, into account. In [13], routing is made more robust by taking into account uncertainties in the defined constants, such as the energy consumption costs and residual battery levels. In [14], the LP routing problem formulation specifically addresses multi-commodity traffic and analyzes the interaction of multiple routes. However, these works were not designed for energy harvesting sensor networks, where the harvest availability and finite capacity limits of the energy buffers need to be considered.

3 System Model

We consider a WSN network with a flat, multihop topology. The network can be described by a directed graph $G(\mathbf{V},\mathbf{E})$, where $\mathbf{V}$ is the set of vertices representing the nodes, and $\mathbf{E}$ is the set of edges representing the links. There is an edge

$e(i, j) \in \mathbf{E}$ between nodes v_i and v_j if v_j is within the transmission range of v_i. Each node v_i has an associated energy cost of e_{ij}^t for transmitting a packet to node v_j and an associated energy cost of e_{ki}^r for receiving a packet from v_k.

Our problem formulation is amenable to multiple sink nodes; however, in this case, we assume that the sink nodes freely and quickly exchange information, and one is designated to perform the routing optimization. Without loss of generality, we designate a single node v_d as a sink in the network. This node is assumed to be tethered with unlimited energy supply. Routing is performed as a centralized algorithm implemented at the sink. All other nodes $v_i \in \mathbf{V}$, $v_i \neq v_d$ can be either a source or router node.

Operation time is divided into time slots T_h of equal length. A commodity (or connection stream) c_m is associated with a source and sequence number (v_s, s_n), and $\mathbf{C} = \{c_1, \ldots, c_M\}$ is the set of M commodities active within a time slot. Additionally, energy harvest and packet origination rates of sources are assumed to be constant within a time slot.

4 Online Maximum Lifetime Utility Routing

Let e_i^* denote the resultant energy of v_i, which is the expected battery level after T_h has elapsed and is defined as

$$e_i^* = \min\left(B_i, e_i + e_i^h - \sum_{c_m \in C}\sum_{j:i \in S_j^{c_m}} e_{ji}^r q_{ji}^{c_m} - \sum_{c_m \in C}\sum_{j \in S_i^{c_m}} e_{ij}^t q_{ij}^{c_m}\right) \quad (1)$$

where B_i is the maximum battery capacity of v_i. The energy units e_i and e_i^h are, respectively, the current battery level and expected energy harvest of v_i for the future time horizon T_h.

This formulation, however, creates a "time coupling" [15] between time slots because e_i of the next time slot depends heavily on e_i^* in the current time slot. This results in a non-linear, sequential program that is very costly to solve. In this work, we instead formulate an online utility objective function that de-couples the dependency using the MAX-TOTAL and MAX-MIN tradeoff previously mentioned.

Each packet being routed in the network is identified to belong to a certain commodity $c_m \in \mathbf{C}$, where $\mathbf{C}$ is the set of commodities active in the current time slot. $S_i^{c_m}$ is the set of downstream neighbors of node v_i towards the destination of commodity c_m. $q_{ij}^{c_m}$ is the rate at which node v_i relays packets belonging to c_m to its downstream neighbor node v_j. e_{ij}^t is the energy cost of transmitting a packet from v_i to v_j, and e_{ji}^r is the corresponding cost at v_i for receiving a packet from v_j. For clarity, the notations used in this section's discussion are summarized in **Table 1**.

We formulate the online maximum lifetime utility (OMLU) routing objective function $\boldsymbol{U}$ as a weighted function between

Table 1. Notation

T_h	Time slot duration/prediction horizon length
B_i	Maximum/initial battery capacity of node v_i
e_i	Residual energy-current battery level
e_i*	Resultant energy-estimated battery level after T_h has elapsed
e_i^h	Predicted harvest-estimated energy harvest over T_h
e_{ji}^r	Energy cost in v_i to receive a packet from v_j
e_{ij}^t	Energy cost in v_i to transmit a packet to v_j
c_m	A traffic commodity associated with a source v_s and destination v_d
$\mathbf{C}$	Set of active traffic commodities in the time slot
$q_{ij}^{c_m}$	Designated packet relay rate from v_i to v_j for commodity c_m
$Q_i^{c_m}$	Packet generation rate of node v_i for commodity c_m
$S_i^{c_m}$	Set of node v_i's downstream neighbors for commodity c_m
$\mathbf{V}$	Set of nodes in the network

the average e_i^* and the minimum e_i^*:

$$\max_{\bar{q}}\left(U = \frac{\alpha}{N_V}\sum_{v_i \in V} e_i^* + \beta \min\left\{e_i^*\right\}\right) \quad \begin{matrix}\forall v_i \in V \\ v_i \neq v_d\end{matrix} \quad (2)$$

$$\text{s.t.} \quad q_{ij}^{c_m} \geqslant 0 \qquad \forall v_i \in V, \quad \forall c_m \in C, \quad \forall v_j \in S_i^{c_m} \quad (3)$$

$$\left(\sum_{v_j : v_i \in S_j^{c_m}} q_{ji}^{c_m}\right) + Q_i^{c_m} = \sum_{v_j \in S_i^{c_m}} q_{ij}^{c_m} \quad \begin{matrix}\forall v_i \in V \\ v_i \neq v_d \\ \forall c_m \in C\end{matrix} \quad (4)$$

$$\sum_{c_m \in C}\left(\sum_{v_j : v_i \in S_j^{c_m}} e_{ji}^r q_{ji}^{c_m} + \sum_{v_j \in S_i^{c_m}} e_{ij}^t q_{ij}^{c_m}\right) T_h \leqslant \min\left(B_i, e_i + e_i^h\right)$$
$$\forall v_i \in V, \; v_i \neq v_d \quad (5)$$

where α and β are weighting parameters, and N_V is the number of nodes in $\mathbf{V}$ excluding the sink. The vector $\bar{q}$ is the set of optimal packet rates of each link $e(i, j)$ with respect to the traffic of each commodity c_m.

The set of constraints in (4) guarantees the conservation of flows [10], which specifies that the total packet outflow of each node must equal the total inflow plus the self-generated traffic of the node $Q_i^{c_m}$ itself if it is the source of c_m. The constraint does not apply if the node is the destination.

The constraints in (5) satisfy the energy requirement that the total consumption of a node (due to reception and transmission) must not exceed the node's available energy, i.e., the minimum between the maximum battery capacity, and the effective energy, i.e., the sum of the residual battery level and expected harvest.

As previously mentioned, MAX-TOTAL is equivalent to the minimization of the sum of total route energy consumption and total wasted energy due to finite maximum battery capacities. Therefore, compared to simply minimizing the total route con-

sumption, MAX-TOTAL has an additional incentive to avoid low-energy nodes if there is energy wastage that can be minimized. However, especially in the case where there is little energy wastage to leverage, MAX-TOTAL still results in quick depletion of nodes along shortest-path routes [9] because these are the routes that result in minimum total energy consumption.

Energy-aware routing can also prioritize the avoidance of low-energy nodes, such as in MAX-MIN. However, in general, MAX-MIN consumes more energy and results in lower overall network residual energy, as will be seen in the simulation results. Therefore, the utility function in (2) attempts to balance these two conflicting objectives. Additionally, the LP optimization can be applied to both single commodity optimization and multi-commodity joint optimization. In section 5.5, we provide simulation results that show the advantage of the latter.

5 Simulation Results

5.1 Simulation Settings

In the following simulations, 25 nodes are placed within a fixed 500 × 500 m area (**Fig. 2**). The transmission range of each node is 150 m. Each node has an energy cost of e_{ij}^{t} = 10 μWh/bit and e_{ij}^{r} = 0 ∀ v_i and v_j. All the nodes have rechargeable batteries that are initially full a capacity $B_{\max}$ of 500 Wh.

In this work, we consider an EH-WSN that harvests solar energy to supplement the initial batteries of the nodes. We use the solar radiation data available from the National Solar Radiation Data Base (NSRDB) [16] for our closest geographical location, which is Chicago O'Hare International Airport. Fig. 2c shows the hourly solar radiation data for July 1, 2010. The PV panels are assumed to have a combined efficiency and sizing factor of 0.2, which means the nodes can only harvest 20% of the energy shown in Fig. 2c.

The simulation begins at 10 a.m. and has a duration T_s of 12 hours to coincide with the maximum availability of sunlight. This duration T_s is divided into timeslots with a duration T_h. The simulation is terminated when a battery is depleted, when the LP no longer has a feasible solution, or when the end of T_s is reached. Although a feasible solution is found for the timeslot, a battery may become depleted within the timeslot because of some variation in the residual battery level readings obtained by the RREQs or energy consumption of the control packets that are not taken into account in the LP formulation.

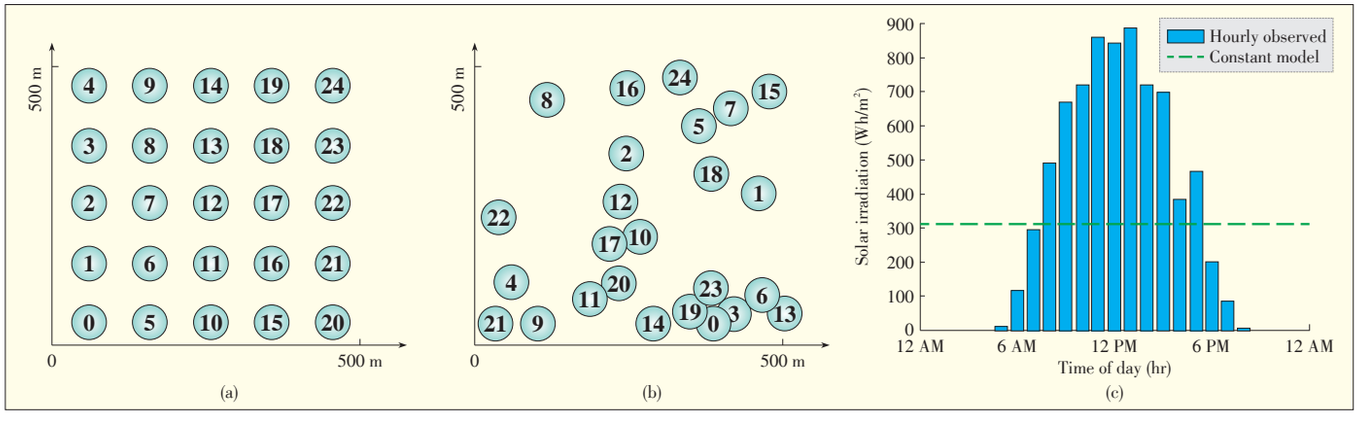
Traffic is characterized by a connection stream or commodity between a randomly chosen source and the fixed destination node. Within a commodity, v_i originates packets in a constant bit rate (CBR) manner at a specified packet rate Q_i. The source continues to produce packets within the timeslot, after which a new source is randomly chosen. All packets are 512 bytes. Last, the weights α and β are both set to 1 so that in the following simulations, OMLU equally balances both objectives.

5.2 On-Demand Routing

We run the simulations using NS2 by extending the on-demand routing protocol Dynamic Source Routing (DSR) in order to implement OMLU. DSR's source routing mechanism allows us to collect per-node energy availability information ($e_i + e_i^h$) without significantly modifying the Route Request (RREQ) header structure, which already collects per-hop node addresses. Moreover, the set of downstream neighbors $S_i^{c_m}$ of v_i for c_m, can be easily inferred from the source routes received at the sink. The route replies (RREP) include a packet rate assigned to that route. Unlike traditional DSR, the resulting route replies are derived from the optimal solution $\bar{q}$ and may not correspond to specific source routes received from an RREQ. Currently, it is assumed that all RREPs are received by the source.

5.3 Performance Metrics

To evaluate performance, we compare TTFND of MAX-TOTAL, MAX-MIN, and OMLU, which are denoted $ttfnd_{MT}$, $ttfnd_{MM}$, and $ttfnd_{OMLU}$, respectively. Additionally, the average and minimum of network battery levels at $ttfnd_0$ are compared, where $ttfnd_0$ = ***min***{$ttfnd_{MT}$, $ttfnd_{MM}$, $ttfnd_{OMLU}$}. Performance metrics may also be measured at $ttfnd_1$, which is the second-lowest



▲**Figure 2. Simulation settings: (a) grid topology, (b) random topology instance, and (c) solar irradiation for July 1, 2010.**

value in the set {$ttfnd_{MT}$, $ttfnd_{MM}$, $ttfnd_{OMLU}$}.

5.4 Simulation Results for Single-Commodity Traffic

In this section, we investigate the advantage of OMLU routing objective over MAX-TOTAL or MAX-MIN by considering single commodity traffic. That is, at each time slot, only a single source is chosen to originate packets for the destination. The timeslot duration T_h is 200 s, and the destination node v_{18}.

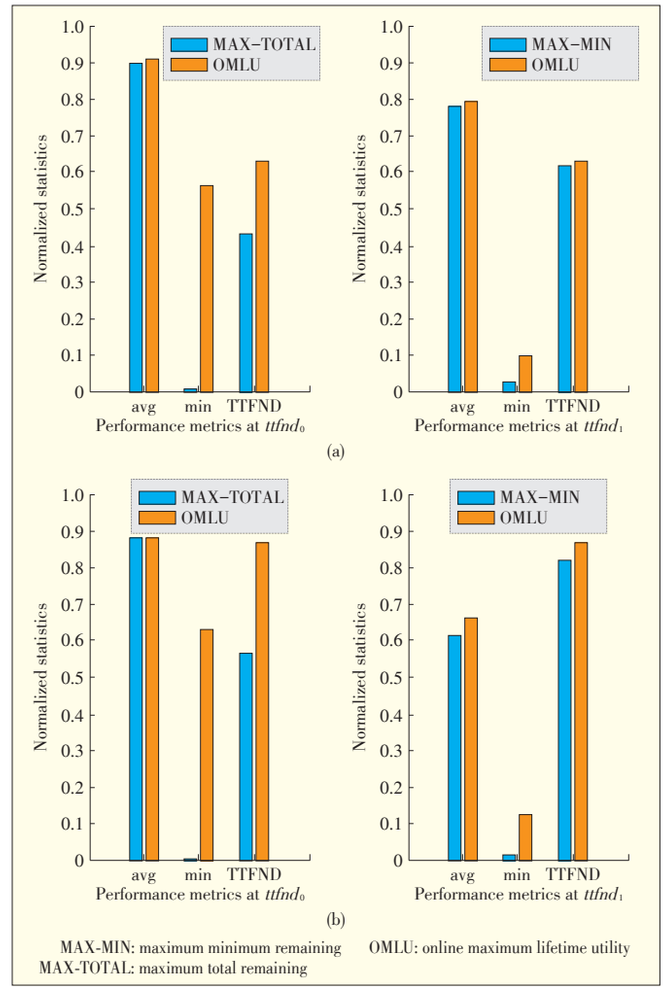
5.4.1 Topology and Traffic

Network topology and traffic variation can affect the performance of routing protocols. Truly optimal performance can be achieved through offline optimization only if future traffic is known; however, this is generally not realistic. Therefore, we run simulations on different topologies and with different traffic realizations to determine whether OMLU's performance improvement can be seen across traffic and topology variations. In the following single-commodity simulations, the source at each timeslot originates packets at a rate of 3 packets per second, that is, $Q_i = 3$ for source node v_i

We first run a simulation on the grid topology shown in Fig. 2a. In this network, only non-diagonal neighbors are able to communicate directly with each other. **Fig. 3a** shows the resulting performance metrics discussed in section 5.3. **Fig. 3b** compares the MAX-TOTAL and OMLU statistics measured at $ttfnd_0$. The right graph compares MAX-MIN and OMLU statistics measured at $ttfnd_1$. The average and minimum battery levels are normalized to B_{max}, and $ttfnd$ is normalized to T_s. The presented results are averaged over five simulations of random traffic on the grid network.

OMLU results in 46.48% longer TTFND than MAX-TOTAL. Moreover, at $ttfnd_0$, which is when the first node is depleted in MAX-TOTAL and occurs around 18,500 seconds into the simulation, the average battery level of the network with OMLU is 1.3% higher than that of MAX-TOTAL. In general, we can expect the average battery of OMLU to be comparable or perhaps slightly lower than that of MAX-TOTAL because maximizing the total (or average) battery is the objective of MAX-TOTAL whereas it is not the only objective of OMLU. Also, at $ttfnd_0$, the minimum battery of OMLU is still more than 50% of B_{max} whereas one or more batteries in MAX-TOTAL have been depleted.

By definition, TTFND itself is very close to the definition of MAX-MIN. Therefore, we expect MAX-MIN to perform well with respect to this metric. Even so, Fig. 3a shows that OMLU results in a longer lifetime than MAX-MIN, although the increase is smaller at 2.06%. Because MAX-MIN aggressively focuses on the minimum network battery without regard for the overall energy availability in the network, more total energy is consumed due to longer routes. However, in the long term, this leads to a shorter lifetime. At $ttfnd_1$, which is the time of first node death for MAX-MIN for these particular set of simulations, we also see higher average and minimum battery levels



▲**Figure 3. Averaged TTFND and battery level statistics on (a) grid and (b) random topologies.**

for OMLU compared to MAX-MIN.

These results highlight the effect of unknown future traffic on current decisions. Within a time slot, MAX-MIN may choose longer routes in order to avoid low-energy nodes. However, this reduces the energy of nodes that may potentially be critical routes of future traffic. Therefore, it is advantageous to be conservative in prioritizing either objective in anticipation of the randomness of future traffic.

Here, we determine performance across various network topologies by running simulations on networks with nodes randomly placed within the area according to a uniform distribution. Fig. 3b shows the same performance metrics averaged over five simulations on randomly generated topologies, such as shown in Fig. 2b, as well as randomly generated traffic. In this case, OMLU results in 63% and 6.24% longer lifetime than MAX-TOTAL or MAX-MIN routing, respectively. These results show a similar trend to that in the grid topology. OMLU results in comparable or higher average battery than MAX-TOTAL at $ttfnd_0$, and also achieves much higher minimum battery

levels at both $ttfnd_0$ and $ttfnd_1$ compared to the other two schemes.

5.4.2 Energy Consumption Rate

Here, Q_i is varied to determine the effect of network energy consumption rate on OMLU performance. **Table 2** shows the resulting TTFND values for the three routing schemes, averaged over five simulations of random traffic and normalized to T_s. Of all packet rates, MAX-TOTAL results in the shortest lifetime compared with MAX-MIN and OMLU.

▼**Table 2. Lifetime values normalized to T_s for various packet rates**

	$Q_i=2$	$Q_i=3$	$Q_i=4$	$Q_i=5$	$Q_i=6$
MAX-TOTAL	0.73	0.43	0.18	0.114	0.078
MAX-MIN	0.868	0.616	0.357	0.205	0.157
OMLU	0.864	0.627	0.417	0.245	0.161

MAX-MIN: maximum minimum remaining OMLU: online maximum lifetime utility
MAX-TOTAL: maximum total remaining

For very low packet rates, the lifetime of OMLU is comparable or slightly shorter than that of MAX-MIN because nodes are able to replenish energy very easily due to the low energy consumption rate. Moreover, the energy consumed by the longer routes in MAX-MIN may otherwise be wasted if not used due to battery storage limit. Therefore, the larger total consumption does not adversely affect the overall residual energy of the network.

However, as the packet rate increases, OMLU achieves a much longer lifetime than either MAX-TOTAL or MAX-MIN. For example, at 3 packets/s, OMLU's lifetime is 46.48% longer than that of MAX-TOTAL and 2.06% longer than MAX-MIN. At 4 packets/s, the performance gain increases to 140.8% and 20.3%, respectively. At 5 packets/s, the lifetime gain remains high at 128.8% and 26.6%, respectively. Finally, at 6 packets/s, OMLU still achieves longer lifetime than either scheme, but the respective gains are lower at 115.8% and 4.1%. The OMLU objective function, like that of MAX-TOTAL, includes the maximization of total or average residual energy, which is equivalent to the minimization of the sum of total energy consumption and network energy wastage. At high consumption rates, there is hardly any wastage to minimize; therefore, the advantage is reduced. Moreover, as consumption increases, the $\min\{e_i^*\}$ term of the objective function decreases faster than the $avg\{e_i^*\}$ term, making the OMLU results more comparable to that of MAX-MIN.

5.4.3 Harvest Prediction

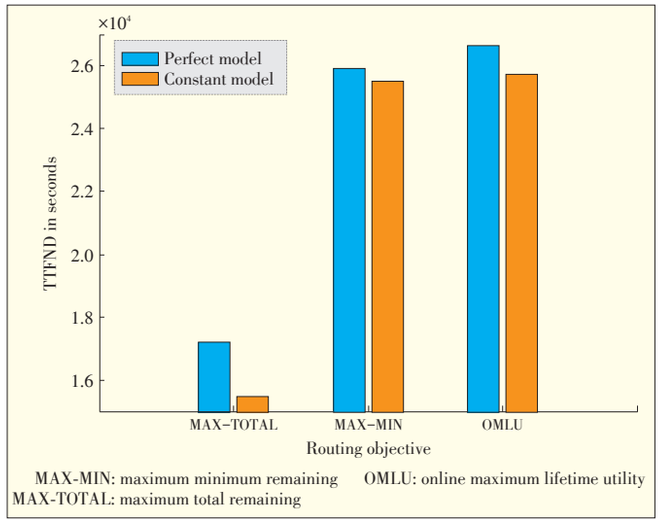
The formulations in (1)–(5) rely on the predicted energy harvest e_i^h. In this subsection, the effect of harvest prediction error is explored. The previous simulations used a perfect prediction model in which e_i^h, obtained at the start of the time slot, reflects the actual total energy harvested by the node within the time slot. To investigate prediction error, we compare OMLU performance with results derived from using a constant prediction model of 309.25 Wh/m^2 (Fig. 2c) that has an RMSE of 331 Wh/m^2. As in previous sections, these results are averaged over five simulations of random traffic.

Fig. 4 shows the resulting TTFND values achieved by the corresponding routing scheme using the perfect and constant prediction models. As can be expected, the error-free prediction model resulted in longer lifetimes than the error-prone constant model, indicating that the accuracy of harvest prediction has non-negligible effect on performance.

Additionally, a comparison of the resulting TTFND values and battery statistics among MAX-TOTAL, MAX-MIN and OMLU using a constant harvest prediction model show similar trends to that seen in the error-free case. That is, OMLU still achieves longer TTFND compared to MAX-TOTAL and MAX-MIN when there is prediction error. Moreover, its average and minimum battery levels are also higher compared to MAX-TOTAL at $ttfnd_0$ and MAX-MIN at $ttfnd_1$. Last, although not shown in these results, harvest prediction errors introduce another adverse effect in that the routing LP problem may yield an infeasible solution due to available energy being underestimated, leading to premature termination.

5.5 Simulation Results for Multi-Commodity Traffic

Another advantage of formulating routing as an LP problem, such as in OMLU, is that we can exploit the joint optimization of multi-commodity traffic. Optimal multi-commodity routing fundamentally differs from separately routing each traffic flow without consideration of the interactions of other traffic flows in the network. In order to jointly optimize multiple commodities, the respective routing requests must be available. While this cannot be assumed in many practical settings, there are al-



▲**Figure 4. TTFND values for the error-free and constant prediction models.**

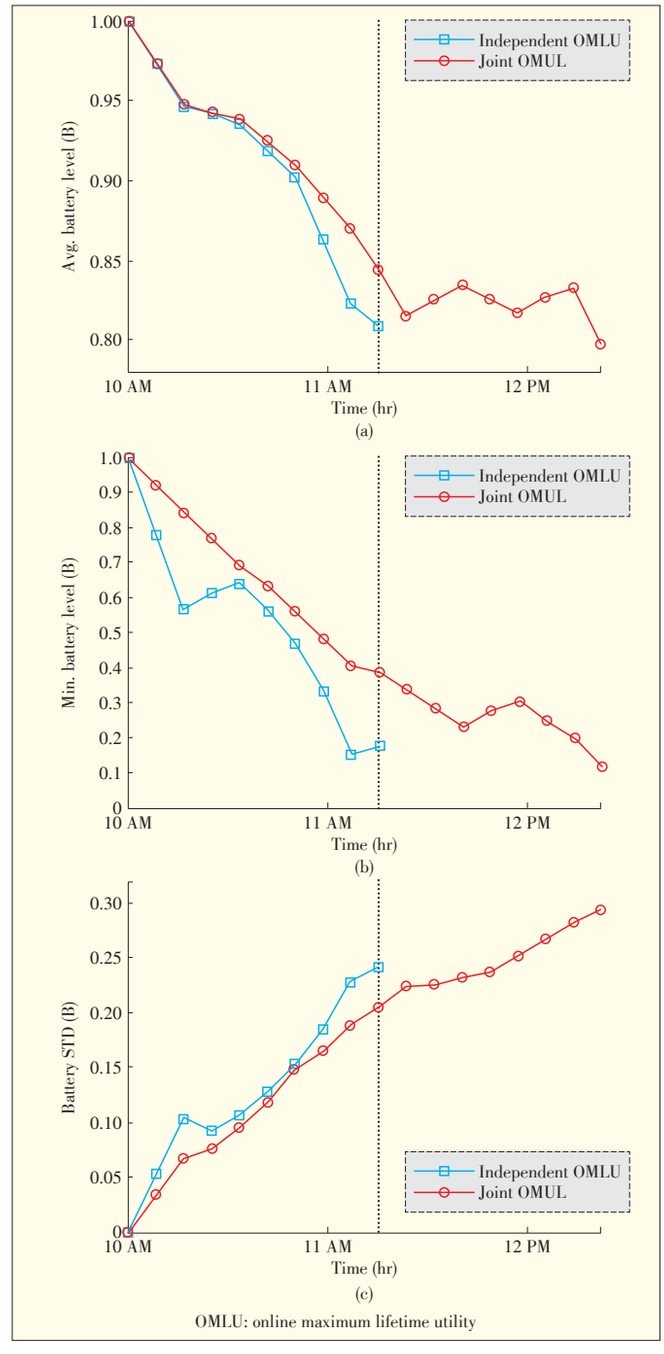
so many cases in which this scenario is applicable, especially in event monitoring WSNs in which nodes within a given area detect a certain event and simultaneously produce data to be sent to the server. In this section, we investigate the performance improvement of OMLU when multiple traffic commodities are jointly optimized as opposed to when commodities are optimized independently.

To simulate multi-commodity traffic, M nodes are randomly chosen at the start of each timeslot to be respective sources for commodities $\mathbf{C} = \{c_1, \ldots, c_M\}$, at a specified packet generation rate of $Q_i^{c_m}$, which is constant within the time slot. In this section, M is set to 3, each with uniform packet rate of 2 packets/s. Also, in order to emphasize the advantage of joint optimization, the timeslot duration is increased to T_h = 1000s.

Again, the results are averaged over five simulations of random traffic. Due to the heavy total packet rate in the network, the simulations terminate after about two or three hours. The battery level observed during the runtime of one such simulation is shown in **Fig. 5**. These results show longer lifetime and higher average and minimum battery levels for joint OMLU as well as lower standard deviation of network battery levels compared to independent OMLU.

Independent OMLU solves the LP routing problem for each commodity, and Joint OMLU waits for the RREQs of all three commodities for each timeslot before solving the LP problem. The independent case terminates after a little more than one hour whereas the joint case has about 2.5 hours of runtime. This indicates the advantage of joint optimization of commodities. In Fig. 5b, both cases terminate with the minimum battery still at about 10% of B_{max} because of the large T_h. Because the LP formulation seeks a solution that theoretically ensures operation throughout the entire timeslot, the simulations terminate due to the infeasibility of the solution.

On average, the joint OMLU yields 35% longer TTFND. At the first node death of the independent case, the average battery level of the joint case is about 0.38% higher. The minimal increase is also mainly due to the relatively large T_h compared to the short runtime of two to three hours. Because routes are "held" in use within the entire timeslot, only several nodes are ever used, leaving most other nodes to remain full. This results in a high average level for both cases. However, the performance improvement is much more evident in the minimum battery where it remains at about 17% of B_{max} for the joint case when the first node of independent OMLU is depleted.



▲**Figure 5. Average, minimum, and standard battery level for multi-commodity traffic for independent and joint OMLU optimization of routes.**

6 Conclusion

Energy-harvesting technology is a viable solution for aggressive energy-saving in 5G wireless networks. Energy-harvesting enables WSNs to support power-hungry applications that may otherwise be impractical because of current battery capacity limits. In this paper, we have considered a WSN in which initial battery is augmented by solar energy harvested using PV panels.

We proposed a linear programming-based solution to routing with a simple utility-based objective function that seeks a balance between the two objectives of maximizing the total and the minimum residual energy. These two objectives are generally conflicting because MAX-TOTAL can deplete nodes along shortest path routes relatively quickly. On the other hand, MAX

-MIN can consume more total energy because longer routes may be selected in order to avoid low-energy nodes, resulting in larger total consumed energy.

The simulation results showed that MAX-TOTAL yielded the shortest lifetime compared to MAX-MIN and our online maximum lifetime utility routing scheme, OMLU. Comparable lifetimes were obtained for MAX-MIN and OMLU in lower energy consumption rates; however, very large lifetime gains over MAX-MIN were achieved by OMLU in high packet rates. Moreover, in the presence of harvest-prediction error, OMLU still resulted in performance gain over the other two routing schemes. OMLU achieves the best of both worlds in that its average battery level is comparable to that of the MAX-TOTAL and the minimum battery level is comparable or higher that that of MAX-MIN.

In this paper, we also investigated the joint optimization of multi-commodity traffic in which multiple traffic flows between different source and destination nodes are jointly optimized. We showed that a joint solution can lead to better performance compared to an independent solution for each traffic flow.

In future work, we plan to formulate network lifetime formally across multiple time slots. However, because energy levels across time slots are not independent, the problem is no longer linear but probabilistic because future traffic is unknown. Moreover, we plan to more extensively investigate practical settings that were not considered in this work, such as mobility and stochastic solar harvest availability.

References

[1] Z. G. Wan, Y. K. Tan, and C. Yuen, "Review on energy harvesting and energy management for sustainable wireless sensor networks," in *IEEE 13th International Conference on Communication Technology*, Jinan, China, Sept. 2011, pp. 362–367. doi: 10.1109/ICCT.2011.6157897.

[2] G. Martinez, S. Li, and C. Zhou, "Maximum residual energy routing in wastage-aware energy harvesting wireless sensor networks", *IEEE 79th Vehicular Technology Conference*, Seoul, South Korea, May 2014, pp. 1–5. doi: 10.1109/VTCSpring.2014.7022982.

[3] L. X. Cai, Y. Liu, T. H. Luan, *et al.*, "Sustainability analysis and resource management for wireless mesh networks with renewable energy supplies," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 2, pp. 345–355, Feb. 2014. doi: 10.1109/JSAC.2014.141214.

[4] L. Lin, N. Shroff, and R. Srikant, "Asymptotically optimal energy-aware routing for multihop wireless networks with renewable energy sources," *IEEE/ACM Transactions on Networking*, vol. 15, no. 5, pp. 1021–1034, Oct. 2007. doi: 10.1109/TNET.2007.896173.

[5] M. K. Jakobsen, J. Madsen, and M. Hansen, "DEHAR: a distributed energy harvesting aware routing algorithm for ad-hoc multi-hop wireless sensor networks," in *IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks*, Montreal, Canada, June 2010, pp. 1–9. doi: 10.1109/WOWMOM.2010.5534899.

[6] D. Hasenfratz, A. Meier, C. Moser, *et al.*, "Analysis, comparison, and optimization of routing protocols for energy harvesting wireless sensor networks," in *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, Newport Beach, USA, June 2010, pp. 19–26. doi: 10.1109/SUTC.2010.35.

[7] C. K. Toh, "Maximum battery life routing to support ubiquitous mobile computing in wireless ad hoc networks," *IEEE Communications Magazine*, vol. 39, no. 6, pp. 138–147, Jun. 2001. doi: 10.1109/35.925682.

[8] Q. Li, J. Aslam, and D. Rus, "Online power-aware routing in wireless ad-hoc networks," in *7th Annual International Conference on Mobile Computing and Networking*, Rome, Italy, Jul. 2001, pp. 97–107. doi: 10.1145/381677.381687.

[9] G. Martinez, S. Li, and C. Zhou, "Wastage-aware routing in energy-harvesting wireless sensor networks," *IEEE Sensors Journal*, vol. 14, no. 9, pp. 2967–2974, Apr. 2014. doi: 10.1109/JSEN.2014.2319741.

[10] J. Chang and L. Tassiulas, "Maximum lifetime routing in wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 4, pp. 609–619, Aug. 2004. doi: 10.1109/TNET.2004.833122.

[11] K. R. Krishnan, D. Shallcross, and L. Kant, "Joint optimization of transmission schedule and routing for network capacity," in *IEEE Military Communications Conference*, San Diego, USA, Nov. 2008, pp. 1–7. doi: 10.1109/MILCOM.2008.4753383.

[12] B. K. Cetin, N. R. Prasad, and R. Prasad, "A novel linear programming formulation of maximum lifetime routing problem in wireless sensor networks," in *7th International Wireless Communications and Mobile Computing Conference*, Istanbul, Turkey, Jul. 2011, pp. 1865–1870. doi: 10.1109/IWCMC.2011.5982819.

[13] I. C. Paschalidis and R. Wu, "On robust maximum lifetime routing in wireless sensor networks," in *47th IEEE Conference on Decision and Control*, Cancun, Mexico, Dec. 2008, pp. 1684–1689. doi: 10.1109/CDC.2008.4738804.

[14] V. Kolar and N. B. Abu-Ghazaleh, "A multi-commodity flow approach for globally-aware routing multi-hop wireless networks," *Fourth Annual IEEE International Conference on Pervasive Computing and Communications*, Pisa, Italy, Mar. 2006, pp. 308–317. doi: 10.1109/PERCOM.2006.3.

[15] S. Chen, P. Sinha, N. Shroff, and C. Joo, "Finite-horizon energy allocation and routing scheme in rechargeable sensor networks," in *IEEE INFOCOM*, Shanghai, China, pp. Apr. 2011, 2273–2281. doi: 10.1109/INFCOM.2011.5935044.

[16] U.S. Dept. of Energy. *National solar radiation database* [Online]. Available: http://rredc.nrel.gov/solar/old_data/nsrdb/

Manuscript received: 2014-07-15

Biographies

Gina Martinez (gmartine@iit.edu) received her BS degree in Computer Engineering from the University of Illinois at Urbana-Champaign in 2005. She received her MS in Computer Engineering from Illinois Institute of Technology (IIT) in 2009. She is currently a PhD student in the Department of Electrical and Computer Engineering, IIT. Her research interests include wireless sensor networks (WSN), green communications, WSN routing, energy-harvesting and resource management in WSNs. She is an IEEE student member.

Shufang Li (lisf@bupt.edu.cn) received her PhD degree in electrical engineering from Tsinghua University, China, in 1997. She is professor and PhD adviser in School of Information and Communication Engineering, Beijing University of Posts and Communications (BUPT). She is also the director of the Joint Lab of BUPT and State Radio Monitoring Center (SRMC), China. Her primary research interests include EMI/EMC, simulation and optimization for radiation interference on high-speed digital circuit, simulation on electro-magnetic environment and countermeasure for interference control, simulation on SAR specific absorption rate) for electromagnetic radiation of cellular phones, and testing technology on EMC, and the theory and design of radio frequency circuits in wireless communication. Professor Li has been a senior member of the IEEE since 2009.

Chi Zhou (zhou@iit.edu) received double degrees in both automation and business administration from Tsinghua University, China, in 1997. She received her MS and PhD degrees in electrical and computer engineering from Northwestern University, USA, in 2000 and 2002. Since 2006, she has been working in the Department of Electrical and Computer Engineering, Illinois Institute of Technology, and is currently an associate professor there. Her primary research interests include wireless sensor networks, scheduling for OFMA/MIMO systems, network coding for wireless mesh networks, and integration of optical and wireless networks. She is a senior member of the IEEE.

DOI: 10.3969/j. issn. 1673-5188. 2015. 01. 006
http://www.cnki.net/kcms/detail/34.1294.TN.20150216.1101.001.html, published online February 16, 2015

Interference-Cancellation Scheme for Multilayer Cellular Systems

Wei Li[1], Yue Zhang[1], and Li-Ke Huang[2]
(1. University of Bedfordshire, Luton, LU1 3JU, UK;
2. Aeroflex UK, Stevenage, SG1 2AN, UK)

Abstract

A 5G network must be heterogeneous and support the co-existence of multilayer cells, multiple standards, and multiple application systems. This greatly improves link performance and increases link capacity. A network with co-existing macro and pico cells can alleviate traffic congestion caused by multicast or unicast subscribers, help satisfy huge traffic demands, and further extend converge. In order to practically implement advanced 5G technology, a number of technical problems have to be solved, one of which is inter-cell interference. A method called Almost Blank Subframe (ABS) has been proposed to mitigate interference; however, the reference signal in ABS still causes interference. This paper describes how interference can be cancelled by using the information in the ABS. First, the interference-signal model, which takes into account channel effect, time and frequency error, is presented. Then, an interference-cancellation scheme based on this model is studied. The timing and carrier frequency offset of the interference signal is compensated. Afterwards, the reference signal of the interfering cell is generated locally and the channel response is estimated using channel statistics. Then, the interference signal is reconstructed according to previous estimation of channel, timing, and carrier frequency offset. The interference is mitigated by subtracting the estimated interference signal. Computer simulation shows that this interference-cancellation algorithm significantly improves performance under different channel conditions.

Keywords

5G; cell edge interference; almost-blank subframe; eICIC

1 Introduction

With the rapid development of 5G wireless networks, heterogeneous links, which support the co-existence of multilayer cells, multiple standards, and multiple applications, are playing an important role in increasing capacity and coverage and satisfying huge traffic demand [1]. This paper discusses technical issues, in particular, interference cancellation, in a heterogeneous network with macro and pico cell. In the topology of a network with macro and pico cells, the high-power 1~40 W macro cell provides basic coverage and the low power 250 mW pico cell is the complementary cell. The pico cell extends network coverage and offloads data traffic of the macro-cell. This reduces cost and increases frequency efficiency. However, the user equipment (UE) served by the pico cell also receives RF signals from neighboring high-power macro cells. This interference is even more severe when users in the pico cell stay within the coverage area of macro cells with range-extension enabled [2].

Enhanced inter-cell interference coordination (eICIC) addresses this issue [2]. eICIC involves two techniques. First, the signal strength is biased to the pico cell, which reduces the interference power. Second, the macro cell remains silent for a certain period, called Almost-Blank Subframe (ABS) [2]. In the ABS, the physical downlink shared channel (PDSCH) is emptied. Therefore, UE does not receive PDSCH during the ABS, and interference can be alleviated. However, users may still receive the cell-specific reference signal (CRS), paging channel (PCH), physical broadcast channel (PBCH), and synchronization channels (PSS/SSS), all of which degrade performance. Further eICIC (FeICIC) has been proposed to eliminate CRS interference.

Some research has been done on CRS interference cancellation (IC). The authors of [3] and [4] investigate direct IC and log-likelihood ratio (LLR) puncturing methods. The simulation results show that direct IC results in better performance. The authors of [5] propose a receiver algorithm that combines IC with a direct-decision channel estimation (CE) algorithm for colliding CRS. The authors of [6] propose a space-alternating gener-

alized expectation-maximization (SAGE) with a maximum a-posteriori (MAP) method for estimating the interfering channel. This method involves reduced computation complexity compared with the linear minimum mean square error (LMMSE) method. However, timing error and frequency offsets can severely degrade performance.

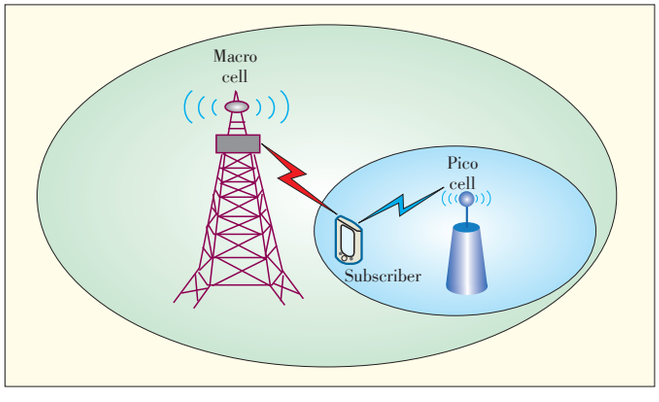
In this paper, we theoretically analyze and run simulations on the CRS interference-cancellation algorithm in a non-colliding scenario where channel statistics are taken into consideration. First, we analyze and model the interference signal and then discuss the interference-cancellation algorithm based on this model. The algorithm makes use of the primary synchronization signal (PSS) and secondary synchronization signal (SSS) to obtain the timing offset (TO) and carrier frequency offset (CFO). Then, the channel response is estimated using channel statistics. Then, the interference signal is reconstructed taking into account the channel effect, TO and CFO. Interference is alleviated by subtracting the interference signal from the received signal.

The rest of this paper is organized as follows. In section 2, the interference is analyzed and modeled. In section 3, we discuss IC algorithms. In section 4, results of the computer simulation are presented. In section 5, we sum up.

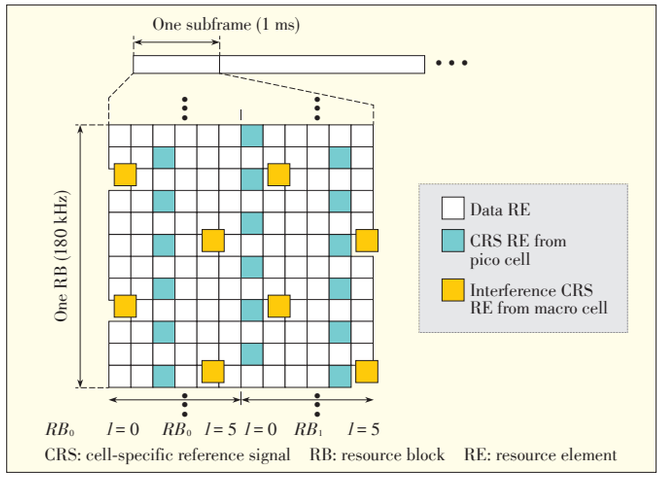
2 Interference Analysis and Model

Fig. 1 shows typical non-colliding inter-cell interference between macro and pico cells. The wireless data service is delivered to the subscriber via pico cell, and the downlink signal from the macro cell interferes with the subscriber at the edge of the pico cell. To alleviate the inter-cell interference, the ABS is transmitted by the macro cell. During the ABS, only certain control signals, such as CRS , are transmitted.

However, the CRS still causes interference for the subscriber. **Fig. 2** shows the received signal of one resource block (RB) with one interference cell. The CRS from a neighboring macro cell overlaps the resource elements (REs) from a serving cell (SC). The SC RE can be divided into data RE and CRS RE. Because of the TO and CFO between the interfering cell and subscriber, the received interfering signal suffers TO and CFO (Fig. 2).



▲Figure 1. Inter-cell interference between macro and pico cells.



▲Figure 2. Received signal in time and frequency grid.

Four modulation schemes are being considered for 5G: generalized frequency-division multiplexing (GFDM), filter bank multicarrier (FBMC), universal filtered multicarrier (UFMC), and biorthogonal frequency-division multiplexing (BFDM). However, these four schemes are all generalizations of OFDM, so we address the original OFDM modulation in the following way (for simplicity's sake).

In the downlink side of OFDM modulation, the frequency-domain signal of the i^{th} symbol is transferred to time domain signal $x_i(n)$ via N-point Inverse Fast Fourier Transform (IFFT):

$$x_i(n)=\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}X_{i,k}e^{j2\pi nk/N}=\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}\left(d_{i,k}+p_{i,k}\right)e^{j2\pi nk/N} \quad (1)$$

where $d_{i,k}$ and $p_{i,k}$ are the data and pilot, respectively.

Then the signal is transmitted over a multipath propagation channel that takes into account additive white Gaussian noise (AWGN). At the receiver side, the received signal is given by

$$y_i(n)=\sum_{l=0}^{L-1}h_l x_i\left(n-\tau_l\right)+\omega(n) \quad (2)$$

where h_l and τ_l are the gain and delay of the lth path, respectively; and $\omega(n)$ is the AWGN. Because of the TO and CFO, the corrupted receiver signal in the case of inter-cell interference is

$$y_i(n)=y_i^{(0)}(n)+y_i^{(1)}(n)+\omega(n)=\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}H_{i,k}^{(0)}X_{i,k}^{(0)}e^{j2\pi nk/N}+\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}H_{i,k}^{(1)}X_{i,k}^{(1)}e^{j2\pi(n-d)(k+f_\Delta)/N}+\omega(n) \quad (3)$$

where $y_i^{(0)}$ and $y_i^{(1)}$ are the desired signal and interference signal, respectively; $H_{i,k}^{(0)}$ and $H_{i,k}^{(1)}$ are the frequency response of the serving channel and interfering channel, respectively; and

d and f_Δ are the relative timing and frequency offset, respectively, between the macro and pico cell. After applying the N-point Fast Fourier Transform (FFT), the OFDM symbol is [7]:

$$Y_{i,k} = Y_{i,k}^{(0)} + Y_{i,k}^{(1)} = H_{i,k}^{(0)} X_{i,k}^{(0)} + \sum_{n=-N/1}^{N/2} e^{\frac{j2\pi nd}{N}} H_{i,n}^{(1)} X_{i,n}^{(1)} \Phi_n + W_{i,k} \quad (4)$$

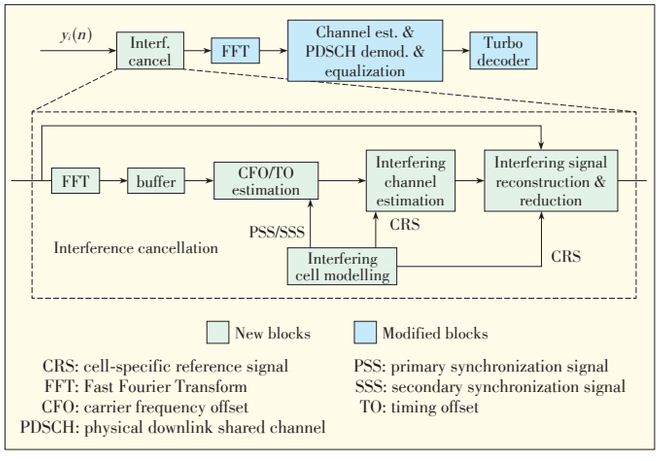
where $H_{i,k}^{(0)}$ and $H_{i,k}^{(1)}$ are the channel coefficients of the serving and interfering cell at kth subcarrier, respectively; and Φ_n is the inter-carrier interference (ICI). During the ABS, only certain control signals are transmitted, and the CRS from the macro cell overlaps the data REs (Fig. 2). At the data REs, the signal model of the serving cell with interference is

$$Y_{i,k} = H_{i,k}^{(0)} d_{i,k}^{(0)} + \sum_{n=-N/2}^{N/2} e^{\frac{j2\pi nd}{N}} H_{i,n}^{(1)} d_{i,n}^{(1)} \Phi_n + W_{i,k} \quad (5)$$

According to (5), the relative timing offset d between interfering and serving cells causes phase shift $e^{\frac{j2\pi nd}{N}}$ on the kth subcarrier. The terms Φ_n in (5) arises from the CFO term f_Δ, which results in intercarrier interference (ICI). Therefore, the CFO and TO need to be compensated. In addition, this model shows the case of single-input single-output (SISO) antenna only. The case of multiple-input multiple-output (MIMO) antenna can be easily derived from (5). However, the number of REs increases because the number of interference CRSs increases with number of antenna ports, which results in more severe interference [8]. These problems will be addressed in section 3.

3 IC Algorithm

The proposed inter-cell IC algorithm is shown in **Fig. 3**. This algorithm includes estimation of CFO and TO, estimation of the interfering channel, modeling of the interfering cell, and reconstruction and reduction of the interfering signal. With CFO and TO estimation, the relative frequency offset and timing offset between the interfering cell and serving cell is estimated using the PSS or SSS generated by modeling the interfering cell. Next, the interfering channel is estimated according to the compensated signal. The interfering signal is then reconstructed according to the previous estimation and subtracted from the received signal.



▲**Figure 3. IC receiver architecture.**

3.1 CFO and TO Estimation

The objective of this module is to retrieve OFDM symbol timing and estimate the CFO of the interfering cell. Many timing- and frequency-synchronization algorithms have been developed. Most of these exploit the periodic nature of the time-domain signal by using cyclic prefix (CP) [9]–[11] or pilot data [12]–[13]. However, there are no data REs in an ABS, which severely reduces the power of the CP. The low SNR of the CP makes both timing and frequency synchronization difficult. Apart from the CP and pilot, there are still the PSS and SSS, which are dedicated to timing and frequency synchronization in the downlink. The PSS and SSS are located at the last and second-last symbol in the time slot 0 and 10. The PSS $pss(n)$ and SSS $sss(n)$ are given by

$$pss(n) = \begin{cases} e^{\frac{j\pi\mu n(n+1)}{63}} & n = 0,1,...,30 \\ e^{\frac{j\pi\mu(n+1)(n+2)}{63}} & n = 31,32,...,61 \end{cases}$$

and

$$sss(2n) = \begin{cases} s_0^{(m_0)}(n)c_0(n) & time\ slot\ 0 \\ s_1^{(m_1)}(n)c_0(n) & time\ slot\ 10 \end{cases}$$

$$sss(2n+1) = \begin{cases} s_1^{(m_1)}(n)c_1(n)z_1^{(m_0)}(n) & time\ slot\ 0 \\ s_0^{(m_0)}(n)c_1(n)z_1^{(m_1)}(n) & time\ slot\ 10 \end{cases}$$

where μ is 25, 29 or 34 and corresponds to the physical layer identity $N_{ID}^{(2)}$; and m_0 and m_1 are derived from the physical layer cell identity group $N_{ID}^{(1)}$; $s_0^{(m_1)}$, $s_1^{(m_0)}$, $c_0(n)$, $c_1(n)$, $z_1(m_0)$ and $z_1(m_1)$ are defined in [8]. The timing and frequency offset can be estimated using the cross-correlation of PSS/SSS [14]:

$$\left\{\hat{d},\hat{f}_\Delta\right\} = \arg\max_{d,f_\Delta}\left(\left|C(d,f_\Delta)\right|\right) \quad (6)$$

where

$$C\left(d,f_\Delta\right) = \sum_{m=1}^{N/2} s_i^*(n) r(n+m) e^{\frac{-2\pi f\Delta n}{N}} \quad (7)$$

The generation of PSS/SSS is based on the assumption of an ideal cell search. The cell-search algorithm in the case of inter-cell interference is beyond the scope of this paper. After the timing and frequency synchronization of the interfering signal, the interfering-channel response can be estimated.

3.2 Interfering-Channel Estimation

Before interference cancellation, it is essential to estimate the interfering-channel response. The channel estimation can be based on least squares (LS) or minimum mean-square error (MMSE) [15], [16]. The MMSE algorithm gives 10–15 dB gain

in signal-to-noise ratio (SNR) for the same mean-square error of CE over LS estimation [15]. However, the MMSE is more complex than the LS algorithm. After timing and frequency offset compensation, (5) can be rewritten as

$$Y_{i,k} = H_{i,k}^{(0)} d_{i,k}^{(0)} + H_{i,k}^{(1)} d_{i,k}^{(1)} + W_{i,k} \tag{9}$$

From (10), the interfering CRS sequence $p^{(1)}$ can be expressed as

$$p^{(1)} = \frac{1}{\sqrt{2}}(1 - 2c(2n)) + j\frac{1}{\sqrt{2}}(1 - 2c(2n+1)) \tag{10}$$

where $c(n)$ is generated by Gold Sequence with a length of 31, the state of which is initialized according to the cell ID, slot number, and antenna port. Assuming that the user conducts ideal cell research, the interfering CRS $p_{i,k}^{(1)}$ can be generated locally. Applying LS CE, the interfering channel can be estimated with

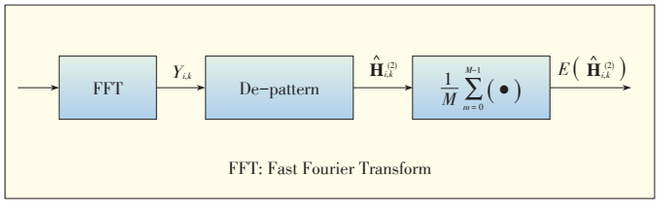
$$\hat{H}_{i,k}^{(1)} = Y_{i,k}/p_{i,k}^{(1)} = H_{i,k}^{(1)} + H_{i,k}^{(0)} d_{i,k}^{(0)}/p_{i,k}^{(1)} + W_{i,k}/p_{i,k}^{(1)} \tag{11}$$

According to (11), the data RE of serving cell $H_{i,k}^{(0)} d_{i,k}^{(0)}/p_{i,k}^{(1)}$ becomes interference with relatively high power. Thus, the estimation in (11) is inaccurate. Numerical studies in [17] show that the distribution of the interference signal is close to Gaussian for a larger RB and non-Gaussian for a smaller. However, the mean of the distribution converges to 0. Therefore, the expectation of (11) can be derived:

$$\begin{aligned} E\{\hat{H}_{i,k}^{(1)}\} &= E\left\{Y_{i,k}/p_{i,k}^{(1)}\right\} = E\left\{H_{i,k}^{(1)} + H_{i,k}^{(0)} d_{i,k}^{(0)}/p_{i,k}^{(1)} + W_{i,k}/p_{i,k}^{(1)}\right\} \\ &= E\left\{H_{i,k}^{(1)}\right\} + E\left\{H_{i,k}^{(0)} d_{i,k}^{(0)}/p_{i,k}^{(1)}\right\} + E\left\{W_{i,k}/p_{i,k}^{(1)}\right\} \\ &\approx E\left\{H_{i,k}^{(1)}\right\} \end{aligned} \tag{12}$$

Equation (12) provides a good estimation of mean value of the interfering channel. This value can be estimated by using a moving-average window in the time dimension (**Fig. 4**). If the moving-average window of length M is within the coherence time of the channel, $\hat{H}_{i,k}^{(1)}$ could be approximated by $E\{\hat{H}_{i,k}^{(1)}\}$. The procedure of the interfering-channel estimation is show in Fig. 4.

The IC algorithm should set the correct antenna number and bandwidth of the interfering cell for interfering-cell CE and interference modelling block. Usually this information is not available at the UE unless the UE decides to hand over to that cell. Therefore, the antenna number and bandwidth of the interfering cell need to be estimated at the UE.



▲**Figure 4. Interfering-channel estimation.**

A straightforward method for interfering-cell CE is to enable the IC control block to always set the maximum possible bandwidth and number of antennas, i.e., 20 MHz and 4 antennas, so that the interfering-cell CE and interference modelling block estimates the channel accordingly. If the actual bandwidth is less than 20 MHz, the power of the pilots outside the signal band will be zero. In the mathematical form, the estimation of the channel that is out of the signal bandwidth is

$$\begin{aligned} &E\left\{\hat{H}_{i,k,out}^{(1)}\right\} = E\left\{H_{i,k}^{(1)}\right\}_{out} + E\left\{H_{i,k}^{(0)} d_{i,k}^{(0)}/p_{i,k}^{(1)}\right\}_{out} + \\ &E\left\{W_{i,k}/p_{i,k}^{(1)}\right\}_{out} \approx 0 + 0 + 0 \end{aligned} \tag{13}$$

Equation (13) indicates that the estimation of the neighbouring cell channel could filter out the interference and noise by moving average. Therefore, the power derived from the channel estimation is reliable way of detecting the signal bandwidth. A similar approach could be taken for detecting the number of antennas as well.

3.3 Interfering-Signal Reconstruction and Reduction

After estimating CFO, TO, and the channel response, the estimated interference signal can be reconstructed on the basis of the local time-domain CRS. The relative timing offset d is potentially larger than the duration of CP, which causes ISI within the OFDM window of a desired signal. Thus, reconstructing a frequency-domain interference signal symbol by symbol could result in inaccurate IC. This algorithm reconstructs the interference signal in the time domain and subtracts it from the received signal in time domain:

$$\hat{y}^{(0)}(n) = y(n) - \hat{y}^{(1)}(n + \hat{d}) e^{\frac{-2\pi j \hat{f}_\Delta n}{N}} * \hat{h}_l \tag{14}$$

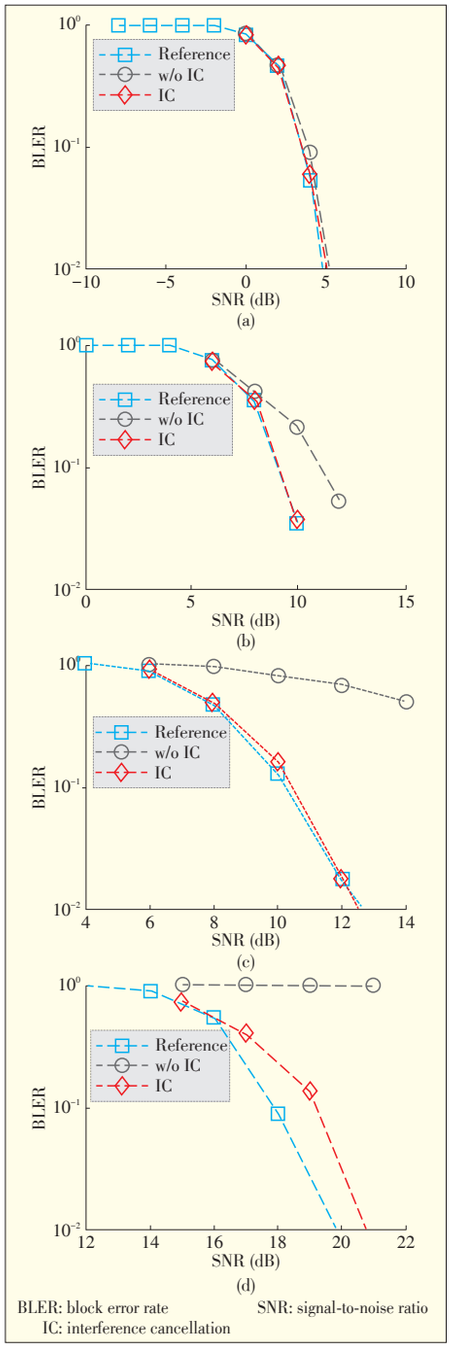
where $\hat{h}_l = FFT\left\{\hat{H}_i^{(1)}\right\}$.

4 Simulation Results

In this section, we evaluate the performance of the IC algorithm using Monte Carlo simulation. We simulate a typical two-cell interference scenario (Fig. 1). The serving cell is set to work in MBSFN mode with 10 MHz bandwidth and different modulation and coding schemes to deliver the service. The interfering cell transmits a normal ABS with a bandwidth of 5 MHz. During the ABS, the CRS overlaps the data RE of desired signal, which causes inter-cell interference. The desired and interfering signal both pass through the time-varying channel with a delay spread smaller than the duration of CP. In the simulation, the WINNER II C2 (EVA) [18] channel model is used with different Doppler frequency determine the effective-

ness of IC under different channel conditions. The arrival time of desired and interfering signal is adjusted to determine the effect of relative timing offset. In addition, different CFOs are applied to the interfering signal to evaluate the effect of CFO. To generate the correct PSS, SSS, and CRS for IC, the user is assumed to conduct an ideal cell search.

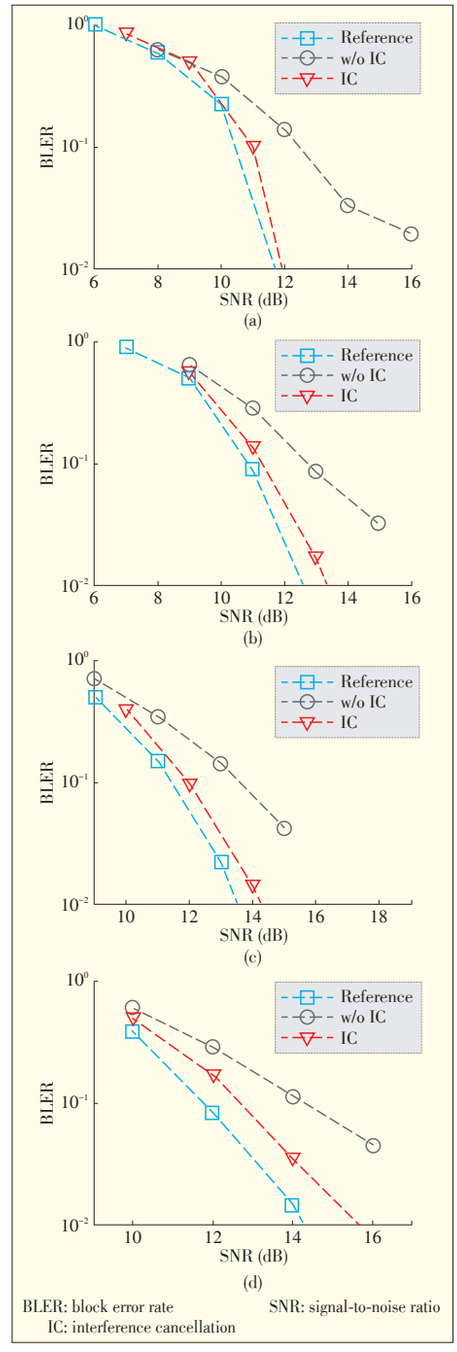
Figs. 5a to **d** show BLER versus SNR for different IC scenarios. MCS 8 and MCS 16 are used. The signal is transmitted via EVA channel with 5 Hz Doppler frequency and with different SNRs. The antenna multiplex mode is set to SIMO and MIMO.



◀**Figure 5. BLER performance vs. SNR in different IC scenarios: a) SIMO, MCS 8, b) SIMO, MCS 16, c) MIMO, MCS 8, d) MIMO, MCS 16.**

The block error rate (BLER) is a performance criteria and is calculated on the basis of 10,000 block transmissions. The BLER of transmission without interference is used as the reference. Fig. 5 also shows the performance with and without IC (red and grey curves, respectively). The inter-cell interference degrades performance during the SNR range of interest. When the IC algorithm is used, BLER approaches that of transmission without interference.

Fig. 6 shows the BLER in different Doppler frequency scenarios. SIMO MCS 18 modulation is used in this simulation,

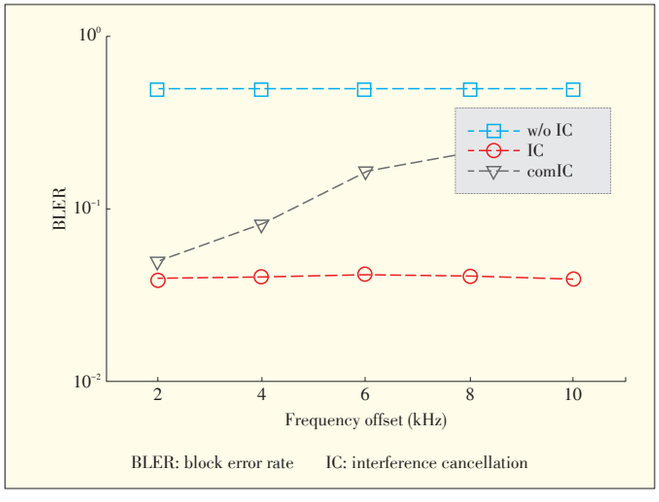


◀**Figure 6. BLER of different Doppler frequency scenarios: a) 5 Hz, b) 70 Hz, c) 150 Hz, d) 200 Hz.**

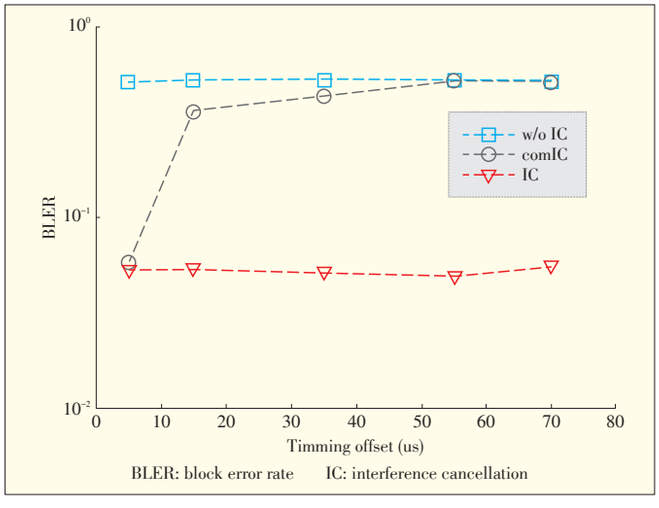
and the Doppler frequency varies from 5 Hz to 200 Hz. The BLER in the case of no interference is the reference (blue curve). The BLER in the case of interference and IC are shown by the grey and red curves, respectively. In Figs. 6a-d, IC significantly improves the BLER for different SNR and Doppler frequencies. This proves the robustness of the IC algorithm.

Fig. 7 shows the effect of CFO on BLER when the proposed IC algorithm and combined IC (comIC) algorithm in [5] are used. The performance of the algorithm in [5] gradually degrades as CFO increases. On the contrary, there is no significant degradation in performance using the proposed algorithm. This proves the effectiveness of frequency synchronization when the CFO is large.

Fig. 8 shows the effect of TO on BLER, when MCS 22 modulation is used. The channel is set at EVA 5Hz, and SNR is set at 16 dB. The performance of proposed IC algorithm is shown by the red curve, and the performance of the comIC algorithm in [5] is shown by the grey curve. The red curve shows that proposed IC algorithm greatly improves BLER when there is a short delay or a very long delay (the inference pilot almost overlaps the following symbol). When the delay is larger than half an OFDM symbol, the BLER for comIC increases, which means that timing synchronization is required. The red curve shows that IC with timing synchronization achieves results in robust performance within the TO range of interest.



▲**Figure 7. BLER performance versus frequency offset.**



▲**Figure 8. BLER performance versus timing offset.**

5 Conclusions

This paper discusses cancellation of inter-cell interference caused by the CRS at the edge of a cell in a multilayer cellular network. This paper describes a signal model that takes into account the interfering signal from a neighboring cell, channel effect, and timing and frequency offset. Using this model, we estimate the TO, CFO, and interfering channel. The interfering signal is then reconstructed locally. Finally, the interference is alleviated by subtracting the reconstructed interference signal. The computer simulation shows this IC algorithm significantly improves performance in different channel conditions. In future work, we will generalize the proposed scheme to non-OFDM cells, such as sparse codebook multiple-access (SCMA) cells and non-orthogonal multiple-access (NOMA) cells, which will also be used in 5G networks.

References

[1] H. Baligh, M. Hong, W.-C. Liao, *et al.*, "Cross layer provision of future cellular networks," *IEEE Signal Processing Magazine*, vol.31, no.6, pp. 56–68, Nov. 2014.

[2] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets," *IEEE/ACM Transaction on Networking*, vol. 22, no. 1, pp. 137–150, Feb. 2014. doi: 10.1109/TNET.2013.2246820.

[3] Qualcomm Inc, "Enabling communication in harsh interference scenarios," R4-102350, 3GPP-RAN WG4 AH#10-03, Bratislava, Jul. 2010.

[4] Qualcomm Inc, "Link level simulations for FeICIC with 9dB cell range expansion," R4-123313, 3GPP-RAN WG4 #63, Prague, May 2012.

[5] M. Huang and W. Xu, "Macro-femto inter-cell interference mitigation for 3GPP LTE-A downlink," in *IEEE Wireless Communications and Networking Conference Workshops*, Paris, France, Apr. 2012, pp. 75–80. doi: 10.1109/WCNCW.2012.6215544.

[6] B. E. Priyanto, S. Kant, F. Rusek, *et al.*, "Robust UE receiver with interference cancellation in LTE advanced heterogeneous network," in *IEEE 78th Vehicular Technology Conference*, Las Vegas, USA, Sept. 2013, pp. 1–7. doi: 10.1109/VTC-Fall.2013.6692396.

[7] H. Nguyen-Le, T. Le-Ngoc, and C. C. Ko, "RLS-basedjoint estimation and tracking of channel response, sampling, and carrier frequency offsets for OFDM," *IEEE Transaction on Broadcasting*, vol. 55, no. 1, pp.84–94, Mar. 2009. doi: 10.1109/TBC.2008.2012361.

[8] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation*, 3GPP TS 36.211, Feb. 2013.

[9] J.-J. van de Beek, M. Sandell, and P. O. Börjesson, "ML estimation of time and frequency offset in OFDM systems," *IEEE Transactions on Signal Processing*, vol. 45, no. 7, pp. 1800–1805, Jul. 1997. doi: 10.1109/78.599949.

[10] M. Speth, D. Daecke, and H. Meyr, "Minimum overhead burst synchronization for OFDM based broadband transmission," in *IEEE Global Telecommunica-*

tions Conference, Sydney, Australia, Nov. 1998, pp. 2777–2782. doi: 10.1109/GLOCOM.1998.776494.

[11] C. C. Ko, R. Mo, and M. Shi, "A new data rotation based CP synchronization scheme for OFDM systems," *IEEE Transaction on Broadcasting*, vol. 51, no. 3, pp. 315–321, Sept. 2005. doi: 10.1109/TBC.2005.851135.

[12] J.-S. Baek and J.-S. Seo, "Effective symbol timing recovery based on pilot-aided channel estimation for MISO transmission mode of DVB-T2 system," *IEEE Transaction on Broadcasting*, vol. 56, no. 2, pp. 193–200, Jun. 2010. doi: 10.1109/TBC.2010.2049054.

[13] X. Wang, T. T. Tjhung, Y. Wu, and B. Caron, "SER performance evaluation and optimization of OFDM system with residual frequency and timing offsets from imperfect synchronization," *IEEE Transaction on Broadcasting*, vol. 49, no. 2, pp. 170–177, Jun. 2003. doi: 10.1109/TGRS.2003.810271.

[14] Y.-H. Tsai and T.-H. Sang, "A new timing synchronization and cell search procedure resistant to carrier frequency offsets for 3GPP-LTE downlink," in *First IEEE International Conference on Communications in China*, Beijing, China, Aug. 2012, pp. 334–338. doi: 10.1109/ICCChina.2012.6356903.

[15] J.-J. van de Beek, O. Edfors, M. Sandell, *et al.*, "On channel estimation in OFDM systems," in *IEEE 45th Vehicular Technology Conference*, Chicago, USA, Jul. 1995, pp. 815–819. doi: 10.1109/VETEC.1995.504981.

[16] V. Srivastava, C. Ho, P. Fung, and Sumei Sun, "Robust MMSE channel estimation in OFDM systems with practical timing synchronization," in *IEEE* Wireless Communications and Networking Conference, Atlanta, USA, Mar. 2004, pp. 711–716. doi: 10.1109/WCNC.2004.1311273.

[17] C. Feng, H. Cui, M. Ma, and B. Jiao, "On statistical properties of co-channel interference in OFDM systems," *IEEE Communication Letters*, vol. 17, no. 12, pp. 2328–2331, Oct. 2013. doi: 10.1109/LCOMM.2013.101813.131297.

[18] *Evolved Universal Terrestrial Radio Access (E-UTRA), User Equipment (UE) Radio Transmission and Reception*, 3GPP TS 36.101, Jul. 2013.

Manuscript received: 2014-09-18

Biographies

Wei Li (wei.li@beds.ac.uk) received his BEng degree from the University of Electronic Science and Technology of China in 2010. He is currently working towards his PhD degree at the University of Bedfordshire, UK, and working with Aeroflex UK on a project looking at the baseband signal process problem in LTE networks. His research interests include signal processing for mobile communications, cognitive radio, OFDM channel estimation, and cooperative communications via relays.

Yue Zhang (yue.zhang@beds.ac.uk) is currently senior lecturer in the Department of Computer Science and Technology, University of Bedfordshire. He is also on industry secondment from the Royal Academy of Engineering working with Aeroflex UK on a high-throughput wireless measurement platform project. He obtained his BEng and MEng degrees from Beijing University of Post and Telecommunications in 2001 and 2004. He received his PhD degree from Brunel University, UK, in 2008. He has worked as a research engineer for the EU IST FP6 project called PLUTO. He then worked as a signal processing design engineer at Anritsu. He was responsible for RF/IF, digital, and DSP design for various wireless communication systems. His research interests include signal processing, wireless communications systems, MIMO-OFDM systems, radio propagation model, and multimedia and wireless networks. He is a member of IEEE and IET.

Li-Ke Huang (li-ke.huang@aeroflex.com) is a technical and research manager at Aeroflex UK. He develops testing and measurement technologies for wireless systems. He specializes in transceiver algorithms and architecture designs for all major wireless communication standards. He is responsible for products and technologies R&D. His research interests include communication system designs and signal processing algorithms and architectures. He received his BSc degree in electronic engineering at Shenzhen University, China, in 1998. He received his PhD degree in communication and signal processing from Imperial College London in 2003.

DOI: 10.3969/j. issn. 1673-5188. 2015. 01. 007
http://www.cnki.net/kcms/detail/34.1294.TN.20150228.1540.001.html, published online February 28, 2015

Big-Data Processing Techniques and Their Challenges in Transport Domain

Aftab Ahmed Chandio[1,3], Nikos Tziritas[1], and Cheng-Zhong Xu[1,2]

(1. Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China;
2. Department of Electrical and Computer Engineering, Wayne State University, MI 48202, USA;
3. Institute of Mathematics and Computer Science, University of Sindh, Jamshoro 70680, Pakistan)

Abstract

This paper describes the fundamentals of cloud computing and current big-data key technologies. We categorize big-data processing as batch-based, stream-based, graph-based, DAG-based, interactive-based, or visual-based according to the processing technique. We highlight the strengths and weaknesses of various big-data cloud processing techniques in order to help the big-data community select the appropriate processing technique. We also provide big data research challenges and future directions in aspect to transportation management systems.

Keywords

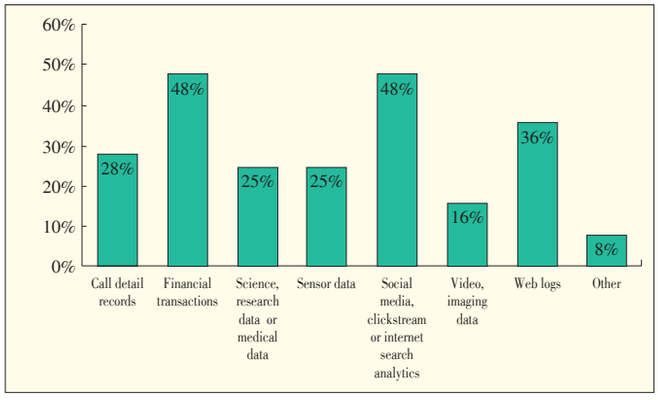
big-data; cloud computing; transportation management systems; MapReduce; bulk synchronous parallel

1 Introduction

In today's ICT era, data is more voluminous and multifarious and is being transferred with increasing speed. Some reasons for these trends are: scientific organizations are solving big problems related to high-performance computing workloads; various types of public services are emerging and being digitized; and new types of resources are being used. Mobile devices, global positioning system, computer logs, social media, sensors, and monitoring systems are all generating big data. Managing and mining such data to unlock useful information is a significant challenge [1]. Big data is huge and complex structured or unstructured data that is difficult to manage using traditional technologies such as database management system (DBMS). Call logs, financial transaction logs, social media analytics, intelligent transport services, location-based services, earth observation, medical imaging, and high-energy physics are all sources of big data. **Fig. 1** shows the results of a big-data survey conducted by Talend [2]. The survey revealed that many common real-world applications deal with big data.

Real-time monitoring traffic system (RMTS) is one of the most interesting examples of a transportation monitoring system (TMS) [3]. In such a system, information about vehicles, buildings, people, and roads are accessed to probe city dynamics. The data is often in the form of GPS location. Because of the real-time nature of data collected in a TMS, the amount of data can grow exponentially and exceed several dozen terabytes [4]. For example, there are 14,000 taxis in Shenzhen. With a 30 s sampling rate, these taxis generate a 40 million GPS records in a day. This GPS data is often used by numerous transportation services for traffic flow analysis, route planning and hot route finder, geographical social networking, smart driving, and map matching [3]–[5]. However, to extract and mine massive transportation data from a database comprising millions of GPS location records, TMS needs an effective, optimized, intelligent ICT infrastructure.

Cloud computing is one of the best potential solutions to dealing with big-data. Many big-data generators have been adapted to cloud computing. According to a survey by GigaSpaces [6], only 20% of IT professionals said their company had no plans to move their big data to the cloud, which indicates



▲Figure 1. Which applications are driving big-data needs at your organization? (multiple responses, n = 95) [2].

This work was supported in part by the National Basic Research Program (973 Program, No.2015CB352400) , NSFC under grant U1401258 and U.S NSF under grant CCF-1016966.

that most companies dealing with big data have turned to the cloud [4]. Several TMS applications, such as cloud-agent-based urban transportation systems, MapReduce for traffic flow forecasting, and cloud-enabled intensive FCD computation framework [7], [8], have been significant in bringing forward the cloud computing paradigm.

Cloud computing integrates with computing infrastructures, e.g., data centers and computing farms, and software frameworks, e.g., Hadoop, MapReduce, HDFS, and storage systems to optimize and manage big data [1]. Because of the importance and usability of cloud computing in daily life, the number of cloud resource providers has increased. Cloud resource providers offer a variety of services, including computation and storage, to customers at low cost and on a pay-per-use basis.

Currently, the cloud computing paradigm is still in its infancy and has to address several issues, such as energy efficiency and efficient resource use [9]–[11]. Unfortunately, as big-data applications are driven into the cloud, the research issues for the cloud paradigm become more complicated. Hosting big-data applications in the cloud is still an open area of research.

In this paper, we describe the fundamentals of cloud computing, and we discuss current big-data technologies. We categorize key big-data technologies as batch-based, stream-based, graph-based, DAG-based, interactive-based, or visual-based. To the best of our knowledge, the Hadoop big-data techniques that fall into these categories have not been covered in the literature to date. In this survey, we highlight the strengths and weaknesses of various Hadoop-based big-data processing techniques in the cloud and in doing so, intend to help people within the big-data community select an appropriate processing technique. We discuss challenges in big-data research as well as future directions in big data related to transportation.

In section 2, we give an overview of cloud computing. In section 3, we give an overview of big data. In section 4, we introduce state-of-the-art big-data processing technologies. In section 5, we discuss big-data research directions and challenges. Section 6 concludes the paper.

2 Cloud Computing and Data-Processing Platforms

Cloud computing is being adapted to every kind of real-world application. Over the next two decades, cloud computing technologies will be crucial to innovation in education, government, healthcare, transportation, traffic control, media, Internet-based business, manufacturing, and media. Cloud computing is the collection of computing resources that can be accessed via a digital network such as wide-area network (WAN) or the Internet. These resources can be accessed using a computer, tablet, notebook, smart phone, GPS device, or some other device. Cloud servers provide and manage the applications and also store data remotely [12].

Cloud computing has been defined in numerous ways by different researchers and experts. The authors of [12]–[14] all have their own opinions of what constitutes cloud computing. NIST [13] defines cloud computing as "a model for enabling ubiquitous, convenient, on-demand network access to shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." Cloud computing is not a new concept; it has been derived from several emerging trends and key technologies, including distributed, utility, and parallel computing [14]. In the following sections, we describe the architecture and key technologies of cloud computing, which is classified in **Fig. 2**.

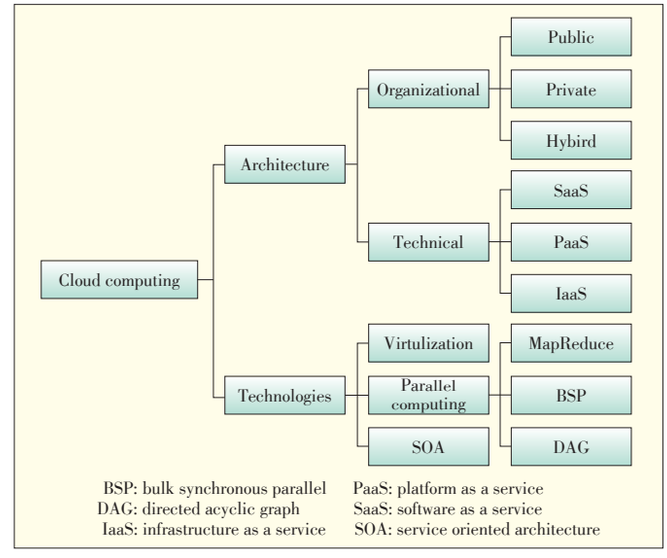
2.1 Cloud Deployment Models

Cloud architecture can be explained from organizational and technical perspectives. From an organizational perspective, cloud architecture can be categorized as public, private or hybrid [15] according to deployment model.

A public cloud deployment model is used for the general public or a large group of industries. Examples are Google App Engine, Microsoft Windows Azure, IBM Smart Cloud, and Amazon EC2. A private cloud deployment model is used for an organization. Examples are Eucalyptus, Amazon VPC, VMware, and Microsoft ECI data center. A hybrid cloud deployment model is a mixture of two or more clouds (i.e., public and private) for a unique domain. Examples are Windows Azure and VMware vCloud.

From a technical perspective, cloud architecture has three main service models: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) [16].

The first-layer IaaS provides access with an abstracted view of centrally located hardware, computers, mass storage, and networks. The most popular examples of IaaS are IBM IaaS,



▲**Figure 2. Classification of cloud computing.**

Amazon EC2, Eucalyptus, Nimbus, and Open Nebula.

The PaaS layer is designed for developers rather than end-users. It is an environment in which the programmer can execute services. Examples of PaaS are Google AppEngine and Microsoft Azure.

The SaaS layer is designed for the end-user, who does not need to install any application software on their local computer. In short, SaaS provides software for rent and is sometimes called on-demand applications over the Internet. The most popular examples of SaaS are Google Maps, Google Docs, Microsoft Windows Live, and Salesforce.com.

2.2 Key Aspects of Cloud Architecture

2.2.1 Service Orientation

In cloud computing, service-oriented architecture (SOA) is a software architecture that defines how services are offered and used. Functions and messages in the SOA model are used by end-users, applications, and other services in cloud computing. In other words, the SOA determines the way services are designed, deployed, and managed. SOA services are flexible, scalable, and loosely coupled [17]. In an SOA, services are interoperable, which means that distributed systems can communicate and exchange data with each another [17].

2.2.2 Virtualization

Virtualization involves creating an abstract, logical view of the physical resources, e.g., servers, data storage disks, networks and software, in the cloud. These physical resources are pooled, managed, and utilized. Virtualization has many advantages in terms of resource usage, management, consolidation, energy consumption, space-saving, emergency planning, dynamic behavior, availability, and accessibility [18]. Operating systems, platforms, storage devices, network devices, and software applications can all be virtualized.

2.2.3 Parallel Computing

The parallel computing paradigm in cloud computing is pivotal for solving large, complex computing problems. The current parallel-computing paradigms in cloud environments include MapReduce, bulk synchronous parallel (BSP), and directed acyclic graph (DAG). The jobs handled within these paradigms are computation requests from the end-user and may be split into several tasks.

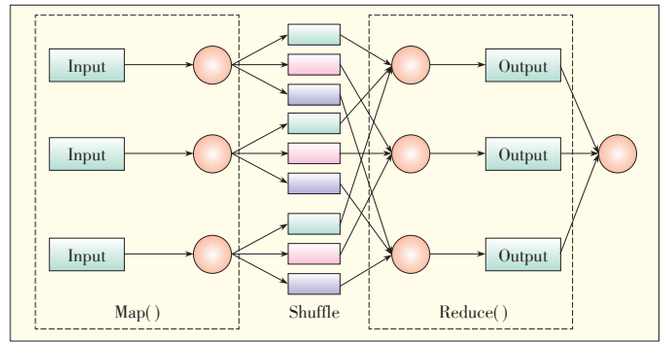
MapReduce was introduced by Google to process mass data on a large cluster of low-end commodity machines [19]. MapReduce is an emerging technique based on Hadoop. It is used to analyze big data and perform high-throughput computing. Hadoop [20] is an Apache project that provides a library for distributed and parallel processing. Each job is divided into several map and reduce tasks (**Fig. 3**). MapReduce takes input data in the form of $<key;value>$ pairs, which are distributed on computation nodes and then map-task produces intermediate $<key;value>$ pairs to distribute them on computation nodes. Finally, the intermediated data is processed by reduce-task to generate the final output $<key;value>$ pairs. During this process, input and output data are stored in Hadoop distributed file system (HDFS), which creates multiple copies of the data as blocks for distribution on nodes.

Bulk Synchronous Parallel (BSP) computing paradigm was introduced by Valiant and Leslie in [21]. A BSP algorithm [21], [22] generates a series of super-steps, each of which executes a user-defined function in parallel that performs computations asynchronously. At the end of every super-step, the BSP algorithm uses a synchronization barrier to synchronize computations within the system. **Fig. 4** shows a BSP program. The synchronization barrier is the state on which every super-step waits for other super-steps running in parallel. The BSP parallel paradigm is well suited to graph computation problems. In [22], BSP performs better than MapReduce for graph processing problems. Hama [23] and Pregel [24] are common technologies based on BSP graph-based processing for big-data analytics.

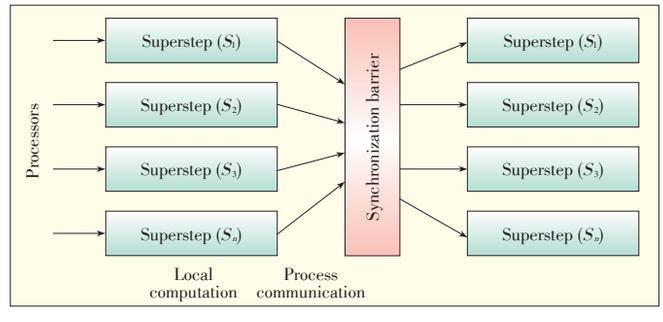
Directed acyclic graph (DAG) computing model describes complicated computing jobs according to dataflow graph processing. DAG is widely used in Dryad, which is a scalable parallel and distributed project of Microsoft [25]. In Dryad, a job is processed in a directed graph manner.

3 Big-Data

Big-data is a huge structured or unstructured data set that is



▲Figure 3. MapReduce framework.

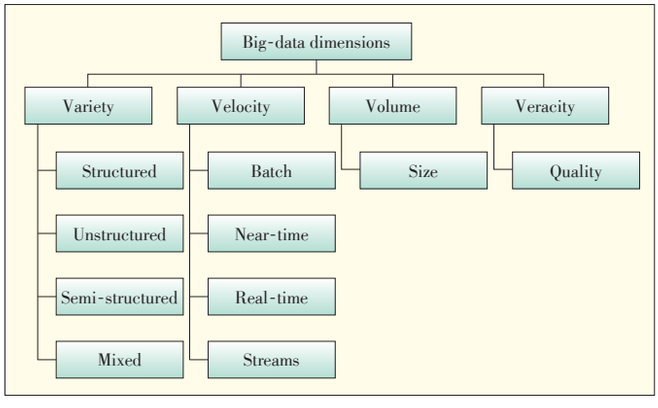


▲Figure 4. Bulk synchronous parallel computing paradigm.

difficult to compute using a traditional DBMS. An increasing number of organizations are producing huge data sets, the size of which start at a few terabytes. For example, in the U.S., Wal-Mart processes one million transactions an hour, which creates more than 2.5 PB of data [2]. In the following sections, we discuss the characteristics and lifecycle of big data.

3.1 Characteristics

Big-data characteristics are often described using a multi-V model (**Fig. 5**). Gartner proposed a 3V model of big data, but an additional dimension, veracity, is also important for data reliability and accuracy [15].



▲**Figure 5. Classification of big-data.**

3.1.1 Volume

Volume is a major dimension of big-data. Currently, the volume of data is increasing exponentially, from terabytes to petabytes and beyond.

3.1.2 Velocity

Velocity includes the speed of data creation, capturing, aggregation, processing, and streaming. Different types of big-data may need to be processed at different speeds [15]. Velocity can be categorized as

- Batch. Data arrives and is processed at certain intervals. Many big-data applications process data in batches and have batch velocity.
- Near-time. The time between when data arrives and is processed is very small, close to real time.
- Real time. Data arrives and is processed in a continuous manner, which enables real-time analysis.
- Streaming. Similar to real-time, data arrives and is processed upon incoming data flows.

3.1.3 Variety

Variety is one of the most important characteristics of big-data. Sources of big-data generate different forms of data. As new applications are developed, a new type of data format may be introduced. As the number of big-data forms grows, designing algorithms or logic for big-data mining and analysis becomes more challenging. Big data can be categorized as

- structured. Big-data in this form is very easy to input and analyze because there are several relational database management (RDBMS) tools that can store, query, and manage the data effectively. Structured big data comprises characters, numbers, floating points, and dates commonly used in customer relationship management systems.
- unstructured. Big-data in this form cannot be stored and managed using traditional RDBMS tools because it is not in a table (i.e., according to a relational model). Unstructured big-data includes location information, photos, videos, audio, emails, sensors data, social media data, biological data, and PDFs that are totally amorphous and very difficult to store, analyze and mine. Social media websites and sensors are major sources of unstructured big data. Eighty to ninety percent of today's data in the world is unstructured social media data [26]. HP Labs has estimated that by 2030 approximately 1 trillion sensors will be in use, monitoring phenomena such as energy consumption, cyberspace, and weather [26].
- semi-structured. Big-data in this form cannot be processed using traditional RDBMS tools. Semi-structured data is a type of structured data that is not organized in a table (i.e., according to a relational model).
- mixed. Big-data may also be a mixture of the above forms of data. Mixed big-data requires complex data capture and processing.

3.1.4 Veracity

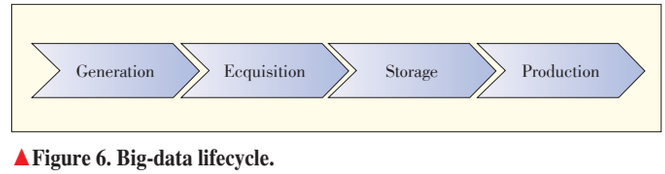
The veracity of big-data is the reliability, accuracy, understandability, and trustworthiness of data. In a recent report [27], it was found 40–60% of the time needed for big-data analysis was spent preparing the data so that it was as reliable and accurate as possible. In several big-data applications, controlling data quality and accuracy has proven to be a big challenge.

3.2 Big-Data Lifecycle

In this section, we describe the big data lifecycle and divide it into four major phases (**Fig. 6**).

3.2.1 Big-Data Generation

The first phase of the big-data lifecycle involves generation of big-data. Specific sources generate a huge amount of multifarious data. that can be categorized as enterprise data, related to online trading, operation, and analysis data managed by RDBMS; Internet of Things (IoT), related to transport, agriculture,



▲**Figure 6. Big-data lifecycle.**

government, healthcare, and urbanization; and scientific, related to bio-medical, computational biology, astronomy, and telescope data [28].

3.2.2 Big-Data Acquisition

Big-data acquisition is the second phase of the lifecycle and involves collection, pre-processing, and transmission of big-data. In this phase, raw data generated by different sources is collected and transmitted to the next stage of the big-data lifecycle. Log files, sensing, and packet capture library (i.e., Libpcap) are common techniques for acquiring big-data. Because big-data has many forms, an efficient pre-processing and transmission mechanism is required to ensure the data's veracity. In particular, before data is sent to the next phase, it is filtered during the acquisition phase to remove redundant and useless data. Data integration, cleaning, and redundant elimination are major methods for big data pre-processing. In that way, new data layout with a meaningful data can save storage space and improve overall computing efficiency for big data processing.

3.2.3 Big-Data Storage

As big data has grown rapidly, the requirements on storage and management has also increased. Specifically, this phase is a responsible of data availability and reliability for big data analytics. Distributed file system (DFS) is commonly used to store big-data originating from large-scale, distributed, data-intensive applications. A variety of distributed file systems have been introduced recently. These include GFS, HDFS, TFS and FastTFS by Taobao, Microsoft Cosmos, and Facebook Haystack. NoSQL database is also commonly used for big data storage and management. NoSQL databases have three different storage models: key-value model, i.e., Dynamo and Voldemort; document-oriented, i.e., MongoDB, SimpleDB, and CouchDB; and column-oriented, i.e., BigTable.

3.2.4 Big-Data Production

Big-data production is the last stage of the big-data lifecycle and includes big-data analysis approaches and techniques. Big-data analysis is similar to traditional data analysis in that potentially useful data is extracted and analyzed to maximize the value of the data. Approaches to big-data analysis include cluster analysis, factor analysis, correlation analysis, regression analysis, and data mining algorithms such as k-mean, Naïve Bayes, a priori, and SVM. However, these methods cannot be used with big-data because of the massive size of data. If any of these methods are to be leveraged by big data analysis, they must be re-designed to make use of parallel computing techniques, which may be batch-based (i.e., MapReduce-based), BSP-based, or stream-based.

4 Big-Data Processing

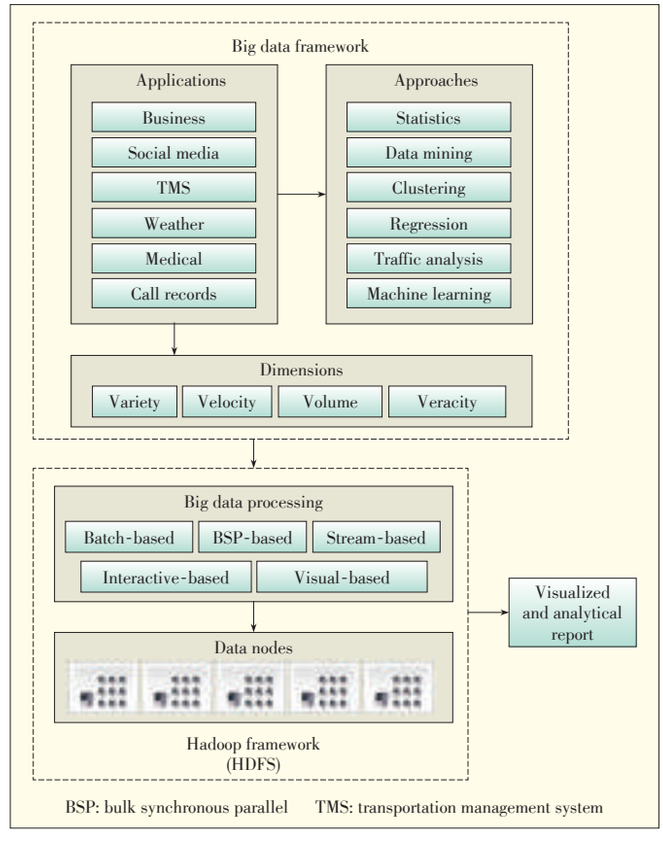
In this section, we explain big data processing approaches and techniques which are based on cloud environments. Firstly, we discuss about major analysis approaches used to analyze big data. Next, we explain five different categories of big data processing techniques in the next subsection. In **Fig. 7**, it is depicted a complete big data framework.

4.1 Analyitic Approaches

Big-data analysis approaches are used to retrieve hidden information from big data. Currently, many big data analysis approaches follow basic analysis approaches. Big data analysis approaches include mathematical approaches and data mining approaches. Basically, an analysis approach chosen by a big data application is totally dependent on the nature of the application problem and its requirements. Particularly, different big data analysis approaches provide different outcomes. We categorize big data analysis approaches into two broad categories: mathematical approaches and data mining approaches [29].

4.1.1 Mathematical Approaches

Mathematical analysis approaches for big data involve very basic mathematical functions including statistical analysis, factor analysis, and correlation analysis used in many fields (i.e., engineering, physics, economics, healthcare, and biology). In statistical analysis, big data can be completely described, summarized, and concluded for its further analysis. Applications



▲**Figure 7. Big-data framework.**

for economic and healthcare widely use statistical analysis approach for big data analysis. In factor analysis, a relationship among many elements presented in big data is analyzed with only a few major factors. In such analysis, most important information can be revealed. Correlation analysis is a common mathematical approach used in several big data applications. Basically, with the help of correlation analysis, we can extract information about a strong and weak dependence relationship among many elements contained in big data.

4.1.2 Data-Mining Approaches

Data mining involves finding useful information from big-data sets and presenting it in a way that is clear and can aid decision-making. Some approaches to data mining in big-data applications include regression analysis, clustering analysis, association rule learning, classification, anomaly or outlier detection, and machine learning.

Regression analysis is used to find and analyze tendency and dependency between variables. For example, in CRM big-data applications, different levels of customer satisfaction that affect customer loyalty can be determined through regression analysis. A prediction model can then be created to help make decisions on how to increase value for an organization.

Clustering analysis is used to identify different pieces of big data that have similar characteristics and understand differences and similarities between these pieces. In CRM, cluster analysis is used to identify groups of customers who have similar purchasing habits and predict similar products.

Association rule learning is used to discover interesting relationships between different variables and uncover hidden patterns in big-data. A business can use patterns and interdependencies between different variables to recommend new products based on products that were retrieved together. This helps a business increase its conversion rate.

Classification analysis is used to identify a set of clusters in data comprising different types of data. It is similar to clustering analysis. Anomaly (outlier) detection is a data-mining technique for detecting data with unmatched patterns and unexpected behavior. Detected data has to be analyzed because it may indicate fraud or risk within an organization.

4.2 Cloud-Based Big-Data Processing Techniques

Cloud-based Hadoop is used for processing in a growing number of big-data applications [20], each of which has a different platform and focus. For example, some big-data applications require batch processing and others require real-time processing. Here, we give a taxonomy of cloud-based big-data processing techniques (**Fig. 8**).

4.2.1 Batch

Big-data batch processing is a MapReduce-based parallel computing paradigm of cloud computing (section 0). There are several tools and techniques are based on batch processing and run on top of Hadoop. These include Mahout [30], Pentaho [31], Skytree [32], Karmasphere [33], Datameer [34], Cloudera [35], Apache Hive, and Google Tenzing.

Mahout [30] was introduced by Apache and takes a scalable, parallel approach to mining big-data. It is used in large-scale data-analysis applications. Google, IBM, Amazon, Facebook, and Yahoo have all used Mahout in their projects. Mahout uses clustering analysis, pattern analysis, dimension reduction, classification, and regression.

Skytree [32] is a general-purpose server with machine learning and advanced analytics for processing huge datasets at high speed. It has easy commands for users. Machine learning tasks in Skytree server include anomaly or outlier detections, clustering analysis, regression, classification, dimensions reductions, density estimation, and similarity search. Because its main focus is real-time analytics, it enables optimized implementation of machine-learning tasks on both structured and unstructured big data.

Pentaho [31] is a big-data software platform for generating business reports. It is enables data capturing, integration, exploration, and visualization for business users. With business analytics, the user can make data-based decisions and increase profitability. Pentaho uses Hadoop for data storage and management and provides a set of plugins to communicate with a document-oriented model of NoSQL databases (i.e., MongoDB) and Cassandra database.

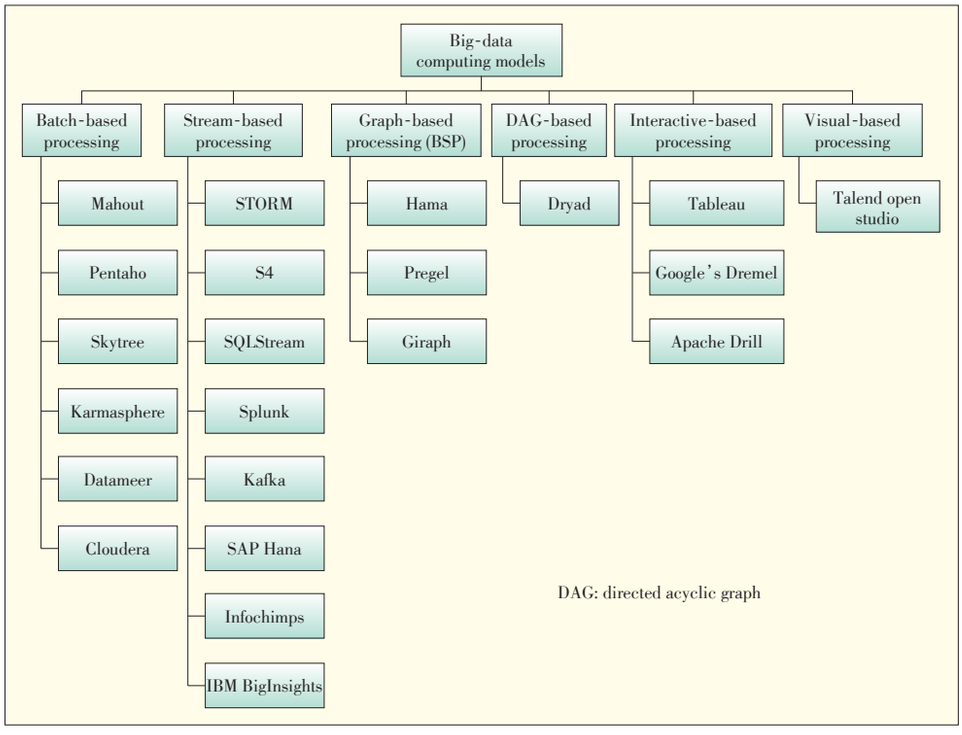
Karmasphere [33] is a platform for business big-data analysis. It is based on Hadoop. With Karmasphere, a program can be efficiently designed for big-data analytics and self-service access. Karmasphere is capable of big-data ingestion, reporting, visualization, and iterative analysis in order to gain business insight. It can process structured and unstructured big data on Hadoop embedded with Hive.

Datameer [34] provides a business integration PaaS, called Datameer Analytic Solution (DAS), which is based on Hadoop and is used to analyze a large volume of business data. DAS includes an analytics engine, data source integration, and data visualization. DAS services are deployed in other Hadoop distributions, such as Cloudera, Yahoo!, Amazon, IBM BigInsights, MR, and GreenplumHD. Because the main objective of Datameer is data integration, data can be imported from structured data sources, such as Oracle, MySQL, IBM, HBase, and Cassandra, as well as from unstructured sources, such as log files, LinkedIn, Twitter, and Facebook.

Cloudera [35] provides Hadoop solutions such as batch processing, interactive search, and interactive SQL. Cloudera is an Apache Hadoop distribution system called CDH that supports MR, Pig, Flume, and Hive. Cloudera also supports embedded plugins with Teradata, Oracle, and Neteza.

4.2.2 Stream

Stream-based processing techniques are used to compute continuous flows of data (data streams). Real-time processing



▲Figure 8. Big-data computing models.

overcomes the limitations of batch-based processing. Projects that use stream processing include Storm [36], S4 [37], SQLStream [38], Splunk, Kafka, SAP Hana, Infochimps, and BigInsights.

Storm [36] is a fault-tolerant, scalable, distributed system that provides an open-source and real-time computation environment. In contrast to batch processing, Storm reliably processes unbounded and limitless streaming data in real-time. Real-time analytics, online machine learning, interactive operating system, and distributed remote procedure call (RPC) are all implemented in Storm project. This project allows the programmer to create and operate an easy setup and process more than a million of tuples per second. Storm comprises different topologies for different Storm tasks created and submitted by a programmer in any programming language. Because Storm works through graph-based computation, it has nodes, i.e., spouts and bolts, in the topology. Each of these nodes contains a processing logic and processes in parallel. A source of streams is called a spout, and a bolt computes input and output streams. A Storm cluster system is managed by Apache ZooKeeper.

In 2010, Yahoo! introduced S4 [37], and Apache included it as an Incubator project in 2011. S4 is a platform that facilitates fault-tolerant, distributed, pluggable, scalable computing. It is designed to process large-scale continuous streams of data. Because its core library is written in Java, a programmer can easily develop applications in S4, which supports cluster management and is robust, scalable, and decentralized. It is used to process large-scale data streams. Analogous to Storm, S4 can also manage the cluster by using Apache ZooKeeper. Yahoo! has deployed S4 for computing thousands of search queries.

SQLStream [38] is a platform for processing large-scale unbound streaming data in real-time with the support of automatic, intelligent operations. Specifically, SQLStream is used to discover interesting patterns in unstructured data. The platform responds quite rapidly because the streaming data is processed in memory. Server 3.0 is a recently released version of SQLStream and is used for real-time big-data analytics and management.

Splunk [39] is a platform for analyzing real-time streams of machine-generated big data. Senthub, Amazon, and Heroku have all used a Splunk big-data intelligent platform to monitor and analyze their data via a web interface. Splunk can be used with structured or unstructured machine-generated log files.

Kafka [40] has been developed for LinkedIn. Kafka is a stream processing tool for managing large-scale streaming and messaging data and processing it using in-memory techniques. Kafka generates an ad hoc solution to the problems created by two different types of data, i.e., operational and activity, belonging to a website. Service logs, CPU/IO usage, and request times are examples of operational data that describes the performance of servers. Activity data, on the other hand, describes the actions of different online users' actions. These actions include clicking a list, scrolling through webpage content, searching keywords, or copying content. Kafka is used in several organizations.

SAP Hana [41] is a stream processing tool that also processes streaming data in-memory. SAP Hana is used for real-time business processes, sentiment data processing, and predictive analysis. It provides three real-time analytics: operational reporting, predictive and text analysis, and data warehousing. SAP Hana can also work with interactive demographic applications and social media.

Infochimps [42] cloud suite covers several cloud IaaS services, categorized as:

- cloud streams: real time analytics for multiple data sources,
- cloud queries: query capability for NewSQL and NoSQL (i.e., Apache Cassandra, HBase, MySQL, and MongoDB)
- cloud Hadoop: analysis of massive amount of data in HDFS.

Infochimps platform is suitable for both private and public

clouds. It can also control STORM, Kafka, Pig, and Hive.

BigInsights [43] is used in the Infosphere platform introduced by IBM. BigInsights manages and integrates information within Hadoop environment for big-data analytics. BigInsights leverages InfosphereStreams, a stream-based tool of the IBM Infosphere. BigInsights is used for real-time analytics on large-scale data streams. JAQL, Pig, Hive (for querying), Apache Lucene (for text mining), and Apache Oozie (job orchestration) are supported by BigInsights.

4.2.3 Graph

Graph-based big-data processing techniques work according to the BSP parallel computing paradigm of cloud computing (section 0). Several big-data applications are better suited to graph-based processing over batch processing [22]. Hama [23], Pregel [24], and Giraph [44] are common useful graph processing techniques for big-data analytics.

Hama [23] is a complete programming model introduced by Apache. It was inspired by BSP parallel computing paradigm running on the top of Hadoop. Hama is written in Java. Massive scientific computations, including matrix algorithms, graph functions, and network computation algorithms, can be easily implemented through Hama [23]. In Hama architecture, a graph is distributed over all the computational nodes, and the vertices are assumed to reside in the main memory during computation. The Hama architecture three main components: BSPMaster, groom servers, and ZooKeeper. BSPMaster maintains the status of groom servers, super-steps, and job progress. A groom server performs BSP tasks assigned by the BSPMaster, and the synchronization barrier is managed efficiently by the Zookeeper component.

Pregel [24] is a graph computational model for efficiently processing billions of vertices connected through trillions of edges. A Pregel program comprises sequences of iterations. In each of these iterations, a vertex may receipt messages, update state or dispatch messages. In this model, a problem is approached through the BSP processing model.

Apache Giraph [44] is an iterative graph-processing system built for high scalability. It is widely used within Facebook to analyze and process the social graph generated by users and their connections. Giraph originates from Pregel and is inspired by the BSP distributed computation model. Features of Giraph include master computation, out-of-core computation, and edge-oriented input.

4.2.4 DAG

Dryad [25] is a scalable parallel and distributed programming model based on dataflow graph processing. Similar to the MR programming model, a Dryad program can be executed in a distributed way on a cluster of multiprocessor or multicore computing nodes. Dryad computes a job in a directed-graph computation manner, wherein each vertex denotes a computational vertex, and an edge denotes a communication channel. This model can generate and dynamically update the job graph and schedule the processes on the resources. Microsoft Server 2005 Integration Services (SSIS) and Dryad-LINQ are built on Dryad.

4.2.5 Interactive

Tableau [45] sits between users and big-data applications by using an interactive mechanism for large-scale data processing. Tableau comprises three different tools: Tableau Desktop, Tableau Server, and Tableau Public. Tableau Desktop visualizes and normalizes data in different ways. Tableau Server offers browser-based analytics called a business intelligence system. Tableau Public is used for interactive visuals. Tableau uses the Hadoop environment and Hive to process queries.

Google Dremel [46] is an interactive analysis system proposed by Google and used for processing nested data. Dremel is a scalable system that complements batch processing tools such as MapReduce. This system is capable of scaling to thousands of processing units. It can process petabytes of data and respond to thousands of users. Dremel can also query very large tables.

Apache Drill [47] is also an interactive analysis system designed for processing nested data similar Google Dremel. It has the flexibility to support different queries and different data sources and formats. A Drill system can scale up to more than ten thousand servers that process petabytes of data (i.e., trillions of records) in seconds. Likely Dremel, Drill stores data in HDFS and performs batch analysis using a MapReduce tool.

4.2.6 Visual

Talend Open Studio [48] is specially designed for visual big-data analysis. Talend Open Studio has user's graphical platform that is completely open source software developed in Apache Hadoop. In this platform, programmer can easily build a program for Big Data problem without writing its Java code. Specifically, Talend Open Studio provides facilities of dragging and dropping icons for building up user's task in Big Data problem. It offers Really Simple Syndication (RSS) feed that may be collected by its components.

5 Big-Data Research Directions and Challenges

In this section, we highlight research directions and challenges in relation to big-data in transportation management systems, which is one of the emerging generators of big-data. In TMS, moving objects such as GPS-embedded taxis and buses generate GPS location data that exponentially increases the volume of big-data. Location data is required in numerous transportation services, such as map matching, to deal with the uncertainty of trajectories, visualize transport data, analyze traffic flow, mine driving patterns, and give smart driving directions. It is also used for crowd sourcing and geographical social net-

working. However, to handle and manage the big-data associated with these transportation services, which produce a massive number of GPS records, TMS needs an optimized, intelligent ICT infrastructure. Here, we describe major transportation services that require further research in terms of big-data management.

5.1 Map Matching

GPS location data are sometimes affected by two typical problems

1) Due to the limitations of positioning devices, moving objects mostly generate noisy and imprecise GPS location data that is called the measurement error. This leads uncertainty in acquiring original locations of the object.
2) Moving objects continuously update their location at discrete time intervals, which may lead to sampling error. The low sampling rate and long intervals between updates may reduce energy consumption and communication bandwidth at the expense of increasing the uncertainty of the actual location. On the contrary, the high-sampling-rate greatly increases the amount of extraneous data.

Therefore, map matching in TMS is used to accurately align the observed GPS locations on a road network in a form of a digital map [5]. Map matching from massive historical GPS location records is performed to predict a driver's destination, suggest the shortest route, and mine certain traffic patterns. However, [5] suggests that map matching is most accurate because of transition probability, which incorporates the shortest path between two consecutive observed GPS location points. On the other hand, the execution of the shortest path queries (SPQs) in the map-matching service involves high computational cost, which makes map-matching unaffordable for real-time processing [5]. Moreover, extraneous data (i.e., in case of a vehicle that stops many times, moves slowly, is trapped in a traffic jam, waits for traffic lights, and moves on a highway link) incurs an extra number of SPQs. The approaches in [49] and [50] are introduced to execute the SPQs by pre-computing the shortest path distances and splitting a road network into small portions so that the required portion can be loaded in the memory [49]. Due to the sequential execution of the SPQs, these approaches incur high pre-computation and storage costs [50].

To map match the huge number of moving objects with tremendous GPS location records (i.e., big data) there is a dire need to execute the SPQs in a computationally efficient environment. The SPQs can be implemented in graph-based big data processing paradigms (i.e., see Section 0) on a large cluster of low-end commodity machines. Consequently, pre-computations of the SPQs on a large cluster of low-end commodity machines benefits low wall-clock-time and storage cost.

5.2 Visualizing Transportation Data

Visualizing transportation data is crucial in TMS to present raw data and compute results generated by data-mining [3]. Such presentation of data reveals hidden knowledge which helps in decision making to solve a problem in the system. In this service, transportation data can be viewed from different perspectives to detect and describe patterns, trends, and relationships in data. Moreover, it provides an interactive way to present the multiple types of data in TMS called exploratory visualization for purpose of investigation. Exploratory visualization can help to detect the relevant patterns, trends, and relations, which can grow new questions that can cause to view the visualized data in more details [3].

Visualizing the massive amount of transportation data i.e., big data conveys a huge amount of information cannot be better visualized and presented in simple and traditional visualization tools. This service can be more challengeable when it visualizes multimodal data that leads to high dimensions of views such as social, temporal, and spatial [3]. In big data research, visualizing the tremendous transportation data is an open issue and needing a large concern on new techniques of big data management.

6 Conclusions

In this paper, we have described cloud computing and key big-data technologies. We categorized big-data key technologies as batch-based, stream-based, graph-based, DAG-based, interactive-based, or visual-based. In this survey, we have discussed the strengths of various Hadoop-based big-data cloud processing techniques that help the big-data community select an appropriate processing technique. Moreover, we have highlighted research directions and challenges in big data in the transportation domain.

References

[1] B. D. Martino, R. Aversa, G. Cretella, et al., "Big data (lost) in the cloud," *International Journal of Big Data Intelligence*, vol. 1, no. 1, pp. 3–17, 2014. doi: 10.1504/IJBDI.2014.063840.
[2] Talend. (2013). *How Big Is Big Data Adoption?* [Online]. Available: http://www.talend.com
[3] Y. Zheng, L. Capra, O. Wolfson, *et al.*, "Urban computing: concepts, methodologies, and applications," *ACM Transaction on Intelligent Systems and Technology*, vol. 5, no. 3, article no. 38, Sept. 2014. doi: 10.1145/2629592.
[4] A. A. Chandio, F. Zhang, and T.D. Memon, "Study on LBS for characterization and analysis of big data benchmarks," *Mehran University Research Journal of Engineering and Technology*, vol. 33, no. 4, pp. 432–440, Oct. 2014.
[5] Y. Lou, C. Zhang, Y. Zheng, *et al.*, "Map-matching for low-sampling-rate GPS trajectories," in *17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Seattle, USA, 2009, pp. 352–361. doi: 10.1145/1653771.1653820.
[6] GigaSpaces. (2013). *Big Data Survey* [Online]. Available: http://www.gigaspaces.com
[7] Q. Li, T. Zhang, and Y. Yu, "Using cloud computing to process intensive floating car data for urban traffic surveillance," *International Journal of Geographical Information Science*, vol. 25, no. 8, pp. 1303–1322, Aug. 2011. doi: 10.1080/13658816.2011.577746.
[8] Z. Li, C. Chen, and K. Wang, "Cloud computing for agent-based urban transportation systems," *IEEE Intelligent Systems*, vol. 26, no. 1, pp. 73–79, 2011. doi: 10.1109/MIS.2011.10.

[9] A. A. Chandio, K. Bilal, N. Tziritas, *et al.*, "A comparative study on resource allocation and energy efficient job scheduling strategies in large-scale parallel computing systems," *Cluster Computing*, vol. 17, no. 4, pp. 1349–1367, Dec. 2014. doi: 10.1007/s10586-014-0384-x.
[10] C.-Z. Xu, J. Rao, and X. Bu, "URL: a unified reinforcement learning approach for autonomic cloud management," *Journal of Parallel and Distributed Computing*, vol. 72, no. 2, pp. 95–105, Feb. 2012. doi:10.1016/j.jpdc.2011.10.003.
[11] A. A. Chandio, C.-Z. Xu, N. Tziritas, *et al.*, "A comparative study of job scheduling strategies in large-scale parallel computational systems," in *12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, Melbourne, Australia, 2013, pp. 949–957. doi: 10.1109/TrustCom.2013.116.
[12] A. A. Chandio, I. A. Korejo, Z. U. A. Khuhro, *et al.*, "Clouds based smart video transcoding system," *Sindh University Research Journal (Science Series)*, vol. 45, no. 1, pp. 123–130, 2013.
[13] P. M. Mell and T. Grance, "The NIST definition of cloud computing," National Institute of Standards and Technology, Gaithersburg, USA, Tech. Rep. SP 800-145, Sept. 2011.
[14] D. Hilley, "Cloud computing: A taxonomy of platform and infrastructure-level offerings," Georgia Institute of Technology, Tech. Rep. GIT-CERCS-09-13, 2009.
[15] M. D. Assunção, R. N. Calheiros, S. Bianchi, *et al.*, "Big data computing and clouds: trends and future directions," *Journal of Parallel and Distributed Computing*, Aug. 2014. doi: 10.1016/j.jpdc.2014.08.003.
[16] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, *et al.*, "The rise of "big data" on cloud computing: review and open research issues," *Information Systems*, vol. 47, pp. 98–115, Jan. 2015. doi: 10.1016/j.is.2014.07.006.
[17] R. Buyya, C. S. Yeo, S. Venugopal, *et al.*, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, Jun. 2009. doi: 10.1016/j.future.2008.12.001.
[18] C. Baun, M. Kunze, J. Nimis, *et al.*, *Cloud Computing: Web-Based Dynamic IT Services*, New York City, USA: Springer, 2011.
[19] J. Dean, and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008. doi: 10.1145/1327452.1327492.
[20] Apache. (2012). *Apache Hadoop Project* [Online] Available: http://www.hadoop.apache.org
[21] L. G. Valiant, "A bridging model for parallel computation," *Communications of the ACM*, vol. 33, no. 8, pp. 103–111, 1990. doi: 10.1145/79173.79181.
[22] T. Kajdanowicz, P. Kazienko, and W. Indyk, "Parallel processing of large graphs," *Future Generation Computer Systems*, vol. 32, pp. 324–337, Mar. 2014. doi:10.1016/j.future.2013.08.007
[23] S. Seo, E. J. Yoon, J. Kim, *et al.*, "Hama: an efficient matrix computation with the mapreduce framework," in *IEEE Second International Conference on Cloud Computing Technology and Science*, Indianapolis, USA, 2010, pp. 721–726. doi: 10.1109/CloudCom.2010.17.
[24] G. Malewicz, M. H. Austern, A. J. C Bik, et al., "Pregel: a system for large-scale graph processing," in *ACM SIGMOD International Conference on Management of Data*, Indianapolis, USA, 2010, pp. 135–146. doi: 10.1145/1807167.1807184.
[25] M. Isard, M. Budiu, Y. Yu, *et al.*, "Dryad: distributed data-parallel programs from sequential building blocks," in *EuroSys'07*, Lisboa, Portugal, 2007.
[26] StateTech. (2013). Breaking down big data by volume, velocity and variety: a new perspective on big data for state and local governments. *Business Intelligence* [Online]. Available: http://www.statetechmagazine.com/article/2013/06/breaking-down-big-data-volume-velocity-and-variety
[27] Paxata, "Ventana research: easing the pain of data preparation," Ventana Research, Feb. 2014.
[28] M. Chen, S. Mao, and Y. Liu, "Big data: a survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014. doi: 10.1007/s11036-013-0489-0.
[29] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: a survey on big data," *Information Sciences*, vol. 275, pp. 314–347, Aug. 2014. doi: 10.1016/j.ins.2014.01.015.
[30] Apache. (2013). *Apache Mahout* [Online]. Available: http://mahout.apache.org/
[31] Pentaho. (2013). *Pentaho Big Data Analytics* [Online]. Available: http://www.pentaho.com/product/big-data-analytics
[32] Skytree. (2013). *Skytree The Machine Learning Company* [Online]. Available: http://www.skytree.net/
[33] Karmasphere. (2012). *FICO Big Data Analyzer* [Online] Available: http://www.karmasphere.com/
[34] Datameer. (2013). *Datameer* [Online]. Available: http://www.datameer.com/
[35] Cloudera. (2013). *Cloudera* [Online]. Available: http://www.cloudera.com/
[36] Apache. (2012). *Apache Storm Project* [Online]. Available: http://www.storm-project.net
[37] L. Neumeyer, B. Robbins, A. Nair, *et al.*, "S4: distributed stream computing platform," in *IEEE International Conference on Data Mining Workshops*, Sydney, Australia, 2010, pp. 170–177. doi: 10.1109/ICDMW.2010.172.
[38] SQLstrean. (2012). *SQLstream s-Server* [Online]. Available: http://www.sqlstream.com/blaze/s-server/
[39] Splunk. (2012). *Storm Splunk* [Online] Available: https://www.splunkstorm.com/
[40] A. Auradkar, C. Botev, S. Das, *et al.*, "Data infrastructure at Linkedin," in *28th International Conference on Data Engineering*, Washington, USA, 2012, pp. 1370–1381. doi: 10.1109/ICDE.2012.147.
[41] S. Kraft, G. Casale, A. Jula, *et al.*, "WIQ: work-intensive query scheduling for in-memory database systems," in *IEEE 5th International Conference on Cloud Computing*, Honolulu, USA, 2012, pp. 33–40. doi: 10.1109/CLOUD.2012.120.
[42] Infochimps. (2013). *Infochimps* [Online]. Available: http://www.infochimps.com
[43] IBM. (2013). *IBM Infosphere BigInsights* [Online]. Available: http://www-01.ibm.com/software/data/infosphere/biginsights/
[44] Apache. (2011). *Apache Giraph* [Online]. Available: http://giraph.apache.org/
[45] Tableau. (2013). *Tableau* [Online]. Available: http://www.tableausoftware.com/
[46] S. Melnik, A. Gubarev, J. J. Long, *et al.*, "Dremel: interactive analysis of web-scale datasets," *Proceedings of the VLDB Endowment*, vol. 3, no. 1, pp. 330–339, 2010.
[47] Apache. (2013). *Apache drill* [Online]. Available: https://www.mapr.com/products/apache-drill
[48] Talend. (2009). *Talend Open Studio* [Online]. Available: https://www.talend.com/
[49] S. Tiwari and S. Kaushik, *Databases in Networked Information Systems: Scalable Method for k Optimal Meeting Points (k-OMP) Computation in the Road Network Databases*, New York City, USA: Springer, 2013, pp. 277–292.
[50] J. R. Thomsen, M. L. Yiu, and C. S. Jensen, "Effective caching of shortest paths for location-based services," in *Proc. 2012 ACM SIGMOD International Conference on Management of Data*, Scottsdale, USA, pp. 313–324. doi: 10.1145/2213836.2213872.

Manuscript received: 2015-01-26

Biographies

Aftab Ahmed Chandio (aftabac@siat.ac.cn) is a doctoral student at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He is also a lecturer at the Institute of Mathematics and Computer Science, University of Sindh, Pakistan. His research interests include cloud computing, big data, parallel and distributed systems, scheduling, energy optimization, workload characterization, and map-matching strategies for GPS trajectories.

Nikos Tziritas (nikolaos@siat.ac.cn) received his PhD degree from the University of Thessaly, Greece, in 2011. He is currently a postdoctoral researcher at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He researches scheduling, load-balancing and replication in CDNs as well as energy optimization and resource management in WSNs and cloud computing systems

Cheng-Zhong Xu (cz.xu@siat.ac.cn) received his PhD degree from the University of Hong Kong in 1993. He is currently a tenured professor at Wayne State University and director of the Institute of Advanced Computing and Data Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include parallel and distributed systems and cloud computing. He has published more than 200 papers in journals and conference proceedings. He was nominated for Best Paper at 2013 IEEE High Performance Computer Architecture (HPCA) and 2013 ACM High Performance Distributed Computing (HPDC). He serves on a number of journal editorial boards, including *IEEE Transactions on Computers*, *IEEE Transactions on Parallel and Distributed Systems*, *IEEE Transactions on Cloud Computing*, *Journal of Parallel and Distributed Computing* and *China Science Information Sciences*. He was a recipient of the Faculty Research Award, Career Development Chair Award, and the President's Award for Excellence in Teaching of WSU. He was also a recipient of the "Outstanding Overseas Scholar" award of NSFC.

DOI: 10.3969/j. issn. 1673-5188. 2015. 01. 008
http://www.cnki.net/kcms/detail/34.1294.TN.20141210.1619.001.html, published online December 10, 2014

Utility-Based Joint Scheduling Approach Supporting Multiple Services for CoMP-SU-MIMO in LTE-A System

Borui Ren, Gang Liu, and Bin Hou
(Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract

In this paper, we study utility-based resource allocation for users supporting multiple services in a LTE-A system with coordinated multi-point transmission for single-user multi-input multi-output (CoMP-SU-MIMO). We designed Joint Transmission Power Control (JTPC) for the selected clusters for minimizing power consumption in LTE-A systems. The objective of JTPC is to calculate the optimal transmission power for each scheduled user and subcarrier. Moreover, based on the convex optimization theory, we propose the dynamic sector selection method in which the average sector throughput and cell-edge users (UEs) rates are performed to achieve the optimal solution. Simulation results show that the system performance achieved by using the proposed suboptimal algorithm is close to that achieved by the dual decomposition method.

Keywords

CoMP-SU-MIMO; multiple services; scheduling algorithm

1 Introduction

Multiple-input multiple-output (MIMO) spatial multiplexing has the potential to dramatically improve spectral efficiency for future communications and networking [1]. However, it may be ineffective when interference levels are high. Suitable processing methods at the transmitter side are thus expected to suppress interference to successfully deploy MIMO spatial multiplexing in cellular environment [2].

Coordinated multi-point (CoMP) transmission/reception was proposed to mitigate inter-cell interference (ICI) by applying the signals transmitted from other cells to assist the transmission instead of acting as interference [3], [4]. According to information shared between coordinated base transceiver station (BTSs), CoMP is mainly characterized as coordinated scheduling (CS)/beamforming or joint processing/transmission (JP) [3]. The class of JP is employed in our work. In this class, data intended for a particular user (UE) is shared among different BTS and is jointly preprocessed at these BTSs. On the other hand, these BTSs can serve a single UE as in the CoMP-single-user (SU)-MIMO mode, or serve multiple UEs simultaneously in the CoMP-multi-user (MU)-MIMO [4]. We focus on the downlink CoMP-SU-MIMO transmission scheme.

However, beyond efficiency and robustness to ICI, next-generation wireless networks are challenged by the demand for multiple services [5]. Resource-scheduling algorithms have been created to solve subcarrier assignment problems for heterogeneous services in MIMO-Orthogonal frequency-division multiplexing (OFDM) systems [6], OFDM-based distributed antenna systems (DAS) [7], and OFDM-based cognitive radio multicast networks [8]. In a CoMP-SU-MIMO system, a series of efficient resource allocation algorithms have been studied. A dynamic clustering approach in wireless networks with multi-cell cooperative processing is proposed in [9]. In [10], a flexible frequency allocation plan (FFAP) has been proposed. A novel transmission scheme with joint Proportional Fairness (PF) scheduling algorithm for CoMP-SU-MIMO has been proposed in [9].

These schemes improve cell-edge efficiency and average sector throughput. However, the main limitation of the algorithms in [9]–[12] is that resource scheduling is designed according to full-buffer model and no multi-service scenario is undertaken.

In this paper, a utility-based scheduling algorithm supporting heterogeneous services for CoMP-SU-MIMO is proposed. The scheme considers not only guaranteeing the system spectral efficiency but also the greater gain improvement with the technology of CoMP. It does not allocate dedicated frequency to cell-edge UEs but treats cell-center UEs and cell-edge UEs equally in every time-frequency resource block (RB) and dynamic CoMP cluster selection method is employed. Two kinds of users are considered: delay-tolerant users (DT-UEs), who require delay-tolerant services such as FTP and email services; and delay-sensitive users (DS-UEs), who require delay-sensitive services such as video streaming. By using Queue Theory to transform delay constraints of DS-UEs into rate constraints, the utility-based scheduling problem is formulated into a mixed-integer nonlinear optimization problem. Using convex optimization theory and dynamic CoMP mechanism, the dual decomposition method combining with CoMP-SU-MIMO is proposed

This work was partially supported by the Fundamental Research Funds for the Central Universities under Grant No. 2012RC0401.

as a solution.

2 System Model

We focus on the downlink of a cellular network with M hexagonal cells, and each cell is partitioned into three base station sectors with k active users served within the coverage of each base station sector (BSS). The base station sectors are assumed to share the same bandwidth B, and the corresponding maximum transmission power P is regarded as uniformly distributed. According to service requirement, users are either homogeneous service users with the File Transfer Protocol (FTP) or multiple services user with video traffic and FTP (**Fig. 1**).

According to [13], when continuous rate adaptation is adopted, the achievable throughput of user k on the n th subcarrier is

$$R_{n,k} = \log_2(1+\beta\gamma_{n,k}) \tag{1}$$

where $\gamma_{n,k}$ is the current SNR for user k on the n th subcarrier and β is a constant related to the target BER by

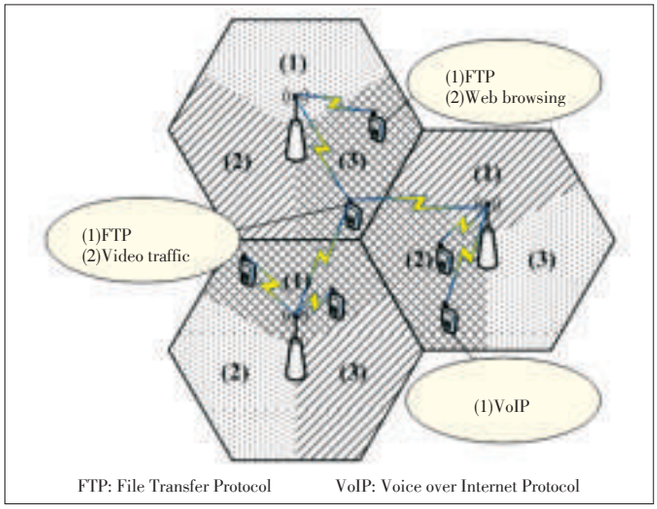
$$\beta = \frac{-1.5}{\ln(5\times \mathrm{BER})} \tag{2}$$

2.1 Non-CoMP SU-MIMO System

Considering of channel gain, the received signal of user k in BSS m based on non-CoMP SU-MIMO mode is:

$$y_m^{(k)} = H_m^{(k)} W_m^{(k)} s_m^{(k)} + \sum_{n\neq m}^{3M}\sum_{w=1}^{K} H_n^{(k)} W_n^{(w)} s_n^{(w)} + n_m^{(k)} \tag{3}$$

where $H_m^{(k)}$ is channel matrix from BSS m to user k and $W_m^{(k)}$ is the precoding matrix. Denote $s_m^{(k)}$ as the transmitted vector for user k served by BSS m, and l_k is the number of data streams. Define $s_m^{(k)} = \left[\sqrt{p_k}s_{m,1}^{(k)}, \sqrt{p_k}s_{m,2}^{(k)}, \cdots, \sqrt{p_k}s_{m,l_k}^{(k)}\right]^T$ and p_k is the transmit power of each data stream at user k. $n_m^{(k)}$ is the addictive white Gaussian noise with zero mean and variance $E(n_m^{(k)} n_m^{(k)*}) = \sigma^2 I_{N_r}$.



FTP: File Transfer Protocol VoIP: Voice over Internet Protocol

▲**Figure 1. Multiple services for the users in CoMP-SU-MIMO system.**

2.2 CoMP SU-MIMO System

According to the interference power queen, a dynamic cluster consisting of two or three base station sectors is considered in the CoMP SU-MIMO mode. The central unit (CU) determines user schedule and power control for all base station sectors in each cluster.

The received signal of user k in base station sector c based on CoMP SU-MIMO mode can be expressed as:

1) Two-BSS cluster

$$y_c^{(k)} = H_c^{(k)} W_c^{(k)} s_c^{(k)} + \sum_{r\neq c}^{2K}\sum_{w=1} H_r^{(k)} W_r^{(w)} s_r^{(w)} + \sum_{r\neq c}^{2}\sum_{n=1}^{K}\sum_{w=1} H_{r,n}^{(k)} W_{r,n}^{(w)} s_{r,n}^{(w)} + n_c^{(k)} \tag{4}$$

2) Three-BSS cluster

$$y_c^{(k)} = H_c^{(k)} W_c^{(k)} s_c^{(k)} + \sum_{r\neq c}^{3K}\sum_{w=1} H_r^{(k)} W_r^{(w)} s_r^{(w)} + \sum_{r\neq c}^{3}\sum_{n=1}^{K}\sum_{w=1} H_{r,n}^{(k)} W_{r,n}^{(w)} s_{r,n}^{(w)} + n_c^{(k)} \tag{5}$$

$H_c^{(k)}$ is the channel matrix from cluster c to user k, and $H_{r,n}^{(k)}$ is the channel matrix from base station sector n in cluster r to user k.

We focus on the downlink of a dynamic CoMP cluster consisting of two or three neighboring base station sectors. According to the long-term channel gain, users are cell-center users (CCUs) or cell-edge users (CEUs). Joint transmission can only be applied to CEUs. In this paper, we focus only on CEUs. The base station sectors are assumed to have one directional transmit antenna each with the same fixed maximum transmission power P and share the same cell-edge bandwidth B. The CEUs are further divided into BE users and VoIP users based on the services they require. Each CEU has one receive antenna and can receive signals from a subset of the base station sectors of the CoMP cluster.

3 Problem Formulation for Multiple Services

3.1 Utility-Based Subcarrier Allocation and Sharing for Homogeneous Services

There are k users in each BSS, and the frequency band consists of N subcarriers. We consider the set of users as $K=\{1,2,\cdots,k\}$ and the set of subcarriers as $N=\{1,2,\cdots,n\}$. Let A_k represent the best-effort services set of user k and s_k be the cardinality of the set. According to Utility Theory, the utility function for best-effort service is given as algorithm form with respect to average data rate. Define U_k to be the total utility function of user k, which can be expressed as:

$$U_k = \sum_{j=1}^{s_k} \alpha_{k,j}\left[a + b*\ln\left(r_{k,j}^{(t)} - c\right)\right]\left(c < r_{k,j}^{(t)}\right) \tag{6}$$

where $\alpha_{k,j}$ is the utility weight of service j, which depends on the priority level of service, and $\sum_{j\in A_k}\alpha_{k,j}=1$. In case of best-effort services, $\alpha_{k,j}=\frac{1}{s_k}$. $r_{k,j}^{(t)}$ denotes average data rate of service j of user k at the t th time slot and a, b, c are constants.

We define $g_{n,k}$ as a subcarrier allocation indicator variable. $g_{n,k}=1$ means that subcarrier n is allocated to user k for packet transmission, or else $g_{n,k}=0$. For any subcarrier n, $\sum_{k=1}^{K}g_{n,k}=1$. To avoid co-channel interference, each subcarrier is allocated one user at most. $b_{k,j}$ is assumed to be the assigned transmission bits for service j of user k and $b_{k,j}\geqslant 0$. Considering that the available throughput of user k on all allocated subcarriers is generally equal to the assigned transmission bits for all services of user k, we have $\sum_{n=1}^{N}g_{n,k}R_{n,k}=\sum_{j=1}^{s_k}b_{k,j}$. Then, $r_{k,j}^{(t)}$ can be updated:

$$r_{k,j}^{(t)}=\frac{t-1}{t}r_{k,j}^{(t-1)}+\frac{b_{k,j}}{t\times T} \tag{7}$$

where T is the time slot duration .

Substituting (7) into (6), we get:

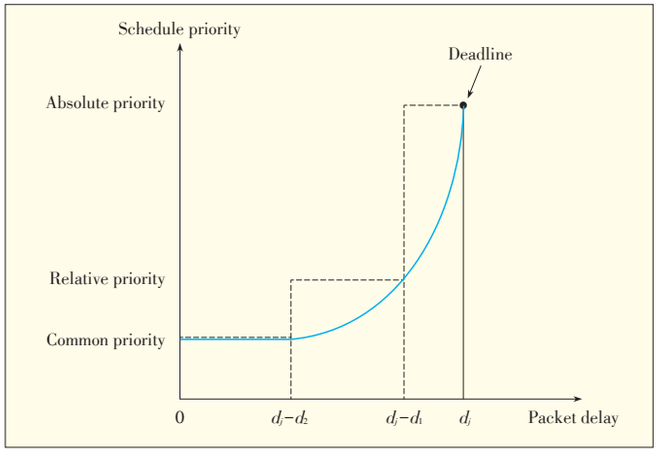
$$U_k=\sum_{j=1}^{s_k}\alpha_{k,j}\ln\left[1+\frac{b_{k,j}}{P_{k,j}^{(t-1)}}\right] \tag{8}$$

where

$$P_{k,j}^{(t-1)}=\left[(t-1)r_{k,j}^{(t-1)}-tc\right]\times T \tag{9}$$

3.2 Scheduling Algorithm for Multiple Services

Compared with best-effort service, user experience mainly depends on delay restriction, which is related to deadline. Owing to the provision of buffer for each service, the scheduling scheme for delay-sensitive service should transmit packets before deadline rather than as soon as possible. In terms of head-of-line (HOL) packet delay (**Fig. 2**), the delay-sensitive service is divided into three parts: absolute priority services (APS), relative priority services (RPS), and common priority services (CPS). d_j is the transmit deadline of the HOL packet of delay sensitive service j, and J_1 and J_2 $(0\leqslant J_1\leqslant J_2)$ are two parameters.



▲**Figure 2. Priority of delay sensitive service.**

The available subcarriers for the SU are firstly assigned to APS until the HOL packet delay is out of $[d_j-d_1,d_j]$. The remaining subcarriers are allocated among best-effort services, RPS, and CPS, which are all reckoned as best-effort services. The optimal resource allocation problem is given by:

$$\begin{aligned}
&P1:\max_{g_{n,k},b_{k,j}}\sum_{k=1}^{K}\sum_{j\in A_k'}U_k\\
&s.t.\sum_{n=1}^{N}g_{n,k}R_{n,k}=\sum_{j=1}^{s_k}b_{k,j},\forall k\in K\\
&\sum_{k=1}^{K}g_{n,k}=1,g_{n,k}\in\{0,1\},\forall n\in N\\
&\qquad\qquad P_i\geqslant 0\\
&\sum_{i=1}^{N}P_i\leqslant P_t
\end{aligned} \tag{10}$$

where A_k' denotes the service set of user k less APS sensitive services, and P_t is the maximum transmit power of the SU transmitter. The optimization problem is a non-linear integer solution with $N\times K$ optimal variables.

Theorem 1: Define k^* as the user that subcarrier n is allocated to, the optimal SU with best effort services to maximize the utility function is:

$$k^*=\arg\max_{k\in K}u_k\left(\sum_{n=1}^{N}\rho^*R_{n,k}\right)R_{n,k} \tag{11}$$

where $u_k\left(\sum_{n=1}^{N}\rho^*R_{n,k}\right)=dU_k\left(\sum_{n=1}^{N}\rho^*R_{n,k}\right)/d\left(\rho^*R_{n,k}\right)$ and $\rho_{n,k^*}=1$, $\rho_{n,k}=0$, while $k\in K$, $k\notin k^*$. ρ^* is the optimal solution of **Algorithm 1**.

Algorithm 1 the MUMS algorithm

1: subcarrier set: $N'=\{remainingavailablesubcarriers\}$,user set: $K''=K-K'$
2: **while** $(N'\neq\varnothing and K'\neq\varnothing)$ **do**
3: 1) randomly select a subcarrier $n^*\in N'$, and identify an optimal user $k^*\in K''$ such that

$$k^*=\arg\max_{k\in K}u_k\left(\sum_{n=1}^{N}\rho^*R_{n,k}\right)R_{n,k}$$

2) assign subcarrier n^* to user k^*
3) $N'=N'-n^*$
4: **end while**

Theorem2: In a heterogeneous service cognitive radio (CR) system, set the subcarrier allocation $\{\rho_{1,k},\rho_{2,k},\cdots,\rho_{N,k}\}$ for user k. In order to maximize the utility of user k, the optimal subcarrier sharing B_k is

$$b_{k,j}=\left[\alpha_{k,j}\frac{b}{\mu}-\left((t-1)r_{k,j}^{(t-1)}-tc\right)\times T\right]^{+} \tag{12}$$

where μ satisfies the maximum available throughput constraint $\sum_{n=1}^{N}\rho_{n,k}R_{n,k}=\sum_{j=1}^{s_k}b_{k,j}$. The Maximum Utility for Multiple Services (MUMS) resource allocation scheme is presented in Algorithm 1.

4 CoMP-SU-MIMO Transmission Scheme

4.1 Dynamic Selection of Coordinated BSS Set

1) Interference Power

If user C in one base station sector S_0 has a service requirement, it may receive interference power from neighboring base station sectors. List the interference power from the maximum to minimum. Base station sectors with the top two interference power S_1 and S_2 are selected.

2) Rate Comparison

Three modes are considered for user C: non-CoMP, BSS S_0 coordinated with BSS S_1, and BSS S_0 coordinated with BSS S_1, S_2 mode. We calculate transmission rate R_1, R_2 and R_3, respectively, for three modes.

3) Mode Selection

We Compare R_1 with $R_2/2$ and $R_3/3$ and chose the most appropriate model with biggest corresponding value out of R_1, $R_2/2$, $R_3/3$.

4.2 Power Control Analysis

In this paper, we designed Joint Transmission Power Control (JTPC) for the selected clusters for minimizing power consumption in LTE-A systems. Unlike other scheduling papers based on uniform full power allocation for each subcarrier, the objective of JTPC is to calculate the optimal transmission power for each scheduled user and subcarrier.

Let $K=\{1, 2,\cdots, k\}$ and $M=\{1, 2,\cdots, m\}$ denote the set of scheduled users and base station sectors, respectively. N_0 denotes the power of the additive white Gaussian noise (AWGN). Let G_{km} denote the channel gain between base station sector m and scheduled user k, consisting of path loss, large-scale fading, and small-scale fading. For a selected coordinated cluster, we ensure joint power control on the basis of signal-to-interference-and-noise ratio (SINR). We update the SINR of the nth scheduled user in the following way:

1) Two-BSS cluster or three-BSS cluster with non-JTPC

$$SINR_k=\frac{\bar{P}G_{km}}{\sum_{i\neq k}\bar{P}G_{im}+N_0} \tag{13}$$

2) Two-BSS cluster with JTPC

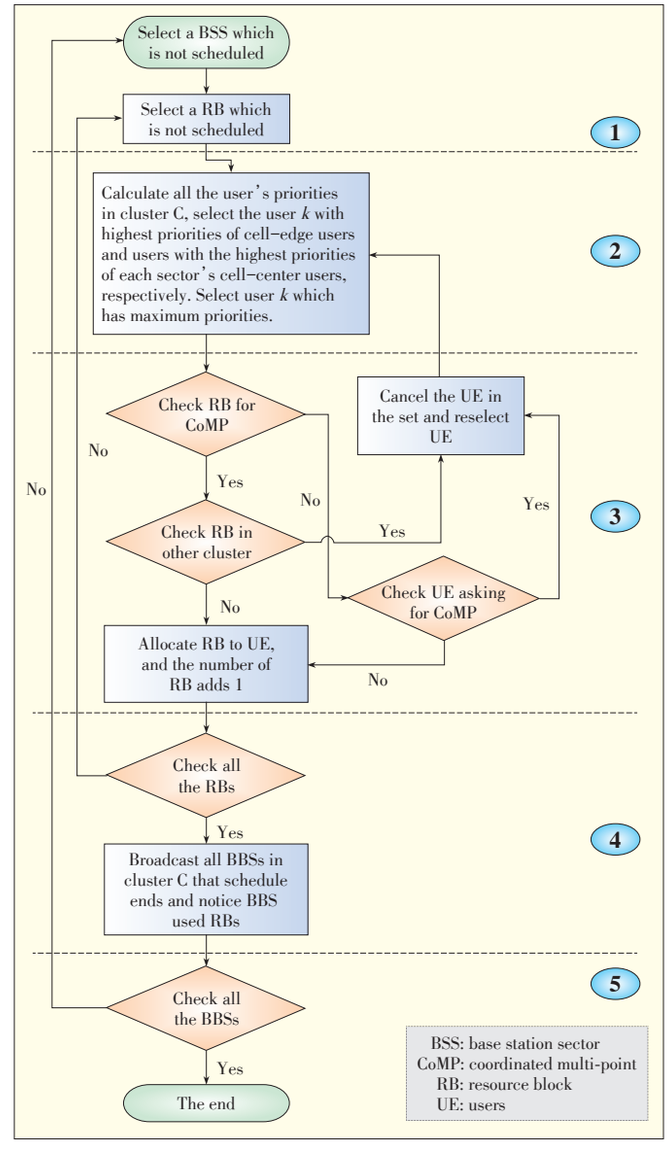
$$SINR_k=\frac{J_{k_1}*\bar{P}G_{km}}{J_{k_2}*\bar{P}G_{km}+\sum_{i\neq k_1,k_2}\bar{P}G_{im}+N_0} \tag{14}$$

3) Three-BSS cluster with JTPC

$$SINR_k=\frac{J_{k_1}*\bar{P}G_{km}}{J_{k_2}*\bar{P}G_{km}+J_{k_3}*\bar{P}G_{km}+\sum_{i\neq k_1,k_2,k_3}\bar{P}G_{im}+N_0} \tag{15}$$

4.3 Proposed CoMP-SU-MIMO Transmission Scheme with Utility-Based Scheduling Algorithm

Based on dynamic selection of transmission mode (Non-CoMP SU-MIMO or CoMP SU-MIMO), a novel CoMP-SU-MIMO transmission scheme (**Fig. 3**) is proposed. Each user in a base station sector has access to every time-frequency resource block. Then there is no need to partition the frequency band dedicated to the cell-edge UEs, and the frequency diversity gain can be fully utilized.



▲Figure 3. Flow chart of the proposed scheme for CoMP SU-MIMO.

With the utility-based scheduling algorithm, the proposed scheme does not need to operate the CoMP SU-MIMO transmission especially for cell-edge UEs within the CoMP frequency zone arranged at the beginning of the frequency band. Instead, it selects the better transmission mode flexibly in each resource block to maximize the frequency diversity gain of the system. Compared to a traditional scheme for CoMP, this scheme, which treats the cell-center UEs and cell-edge UEs equally in every RB, ensures fairness within the system and increases average sector throughput and cell-edge UE rates.

4.4 Scheduling Algorithm for Multiple Services

In this subsection, MUMS, a dynamic subcarrier assignment algorithm for multiple services, is formulated. According to the solution for homogeneous services in section 3.1, the scheduling algorithm for multiple services consists of three parts:

Part 1: Because the APS is prior to other services, the available subcarriers are allocated to the service at first until the HOL packet is out of $[d_j - d_1, d_j]$, adopting the greedy subcarrier assignment described as **Algorithm 2**.

Algorithm 2 greedy subcarrier allocation

1: subcarrier set: $N = \{1,2,\cdots,n\}$, user set: $K' = \{userswhohaveAPS\}$
2: **while** $(N \neq \varnothing$ and $K' \neq \varnothing)$ **do**
3: 1) find $(n^*, k^*) = \arg\max R_{n,k}$
2) assign subcarrier n^* to user k^*
3) **if** the HOL packet delay is out of $[d_j - J_1, d_j]$
then $K = K' - \{k^*\}$
end if
4) $N = N - n^*$
4: **end while**

Part 2: The remaining subcarriers are allocated among best-effort services, RPS, and CPS. Because of the total power's uniform distribution to all subcarriers, the sum power constraint of each cell can always be satisfied. Then, the optimal subcarrier allocation problem can be solved by P1.

Part 3: If user k has RPS sensitive service, the available throughput is assigned to RPS at first until the HOL packet delay is out of $[d_j - d_2, d_j - d_1]$. The remaining transmission bits are shared as (13).

5 Simulations Results and Discussion

5.1 Simulation Design

By conducting system-level simulations and the corresponding parameters are listed in **Table 1**, the performance of the cell average throughput, the cell-edge average throughput, and the utility fairness are evaluated for the proposed MUMS

Table 1. Simulation parameters

Parameters	Values
Cell layout	19 cells / 57 sectors
Radius of cell	500 m
Carrier frequency	2 GHz
Channel bandwidth	10 MHz
eNB transmitting power	31–49 dBm
Noise power	-174 dBm/Hz
Traffic model	FTP/video
User distribution	Uniform
User moving speed	3 km/h
Simulation TTIs	100
CoMP cluster	2/3 sectors
FTP utility function	$U(\bar{r}) = 0.16 + 0.18\ln(\bar{r} - 0.3)$
Path loss model	$128.1 + 37.6\log_{10}(D)$
Antenna pattern	$A(\theta) = -\min\left[12(\frac{\theta}{65})^2, 20\,dB\right]$

eNB: enhanced Node B CoMP: coordinated multi-point
TTI: transmission time intervals FTP: File Transfer Protocol

scheduling scheme in the CoMP-SU-MIMO system with joint transmission power control (MUMS-CJ) and in the CoMP-SU-MIMO system without joint transmission power control (MUMS-CNJ). For comparison, PF scheduling scheme in the non-CoMP system and in the CoMP-SU-MIMO system without joint transmission power control are considered and are called PF-NCNJ and PF-CNJ, respectively. A CoMP-SU-MIMO system with 10MHz bandwidth and 19-cell hexagonal grid layout is considered. A cluster of two or three BSSs is taken into account. The path-loss model is based 3GPP TR 36.942[14]:

$$L = 40\cdot(1 - 4\cdot 10^{-3}\cdot Dhb)\cdot \log_{10}(R) - 18\cdot\log_{10}(Dhb) + 21\cdot\log_{10}(f) + 80dB \quad (16)$$

where L is the pass-loss value, f is the carrier frequency in MHz, R is the distance between the BS and the user in kilometers [12], and Dhb is the height of the transmit antenna relative to the average height of the roof in meters. All base stations and users have two omni-directional antennas. In the simulations, two types of representative services are considered: video streaming (on behalf of delay-sensitive real-time service) and FTP (on behalf of best-effort service). These two kinds of service are modeled according to 3GPP recommendation. For video streaming service, the average data rata and maximum packet delay are assumed to be 128 kbps and 100 ms, respectively. Two kinds of users are assumed in the system. Kind A is a BE service user requiring only a FTP service, and kind B

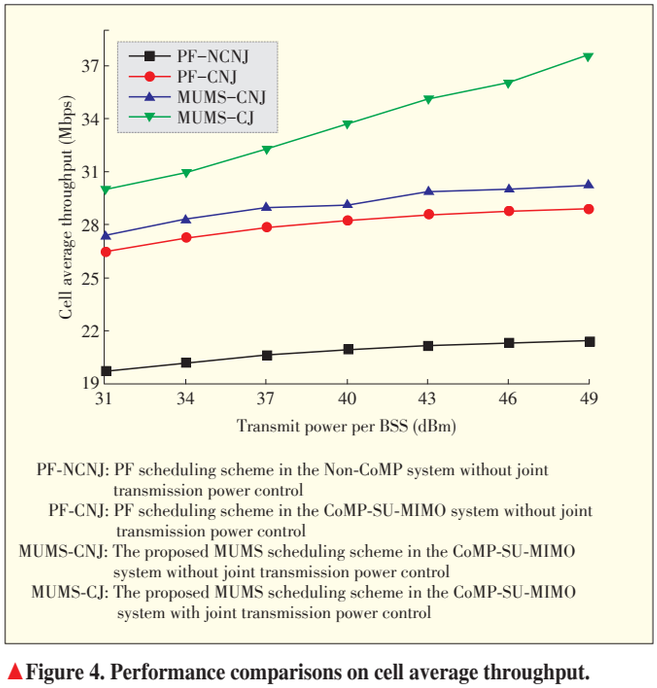
is heterogeneous service user in need of a video streaming service and FTP service. We set the utility weight αj for video streaming and FTP service to 0.7 and 0.3, respectively.

5.2 Simulation Results

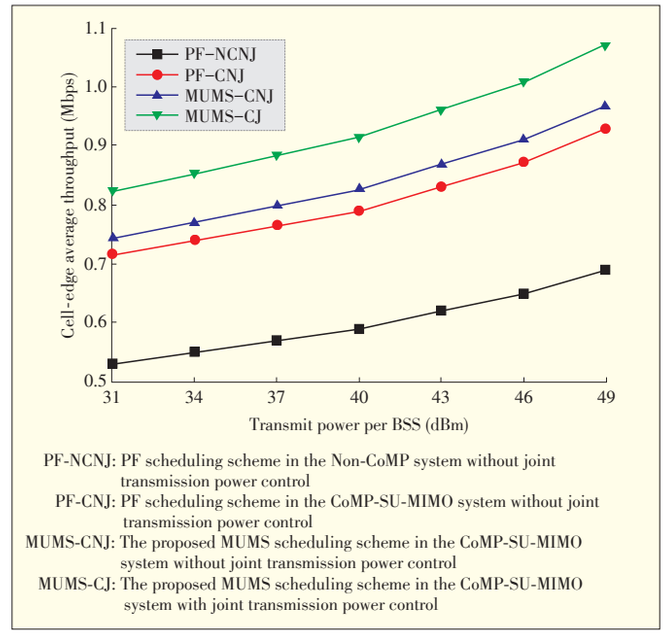
Fig. 4 shows the cell average throughput for the four different schemes considered in this paper. The cell average throughput increases as the transmit power per base station sector increases in all four schemes. MUMS-CJ yields 17% higher cell average throughput than MUMS-CNJ, which is consistent with the joint transmission power control analysis in section 4. After adopting the joint transmission power control, the scheduled users get the biggest SINR.

Fig. 5 shows the cell-edge average throughput for the four different schemes with respect to the transmit power per base station sector. First, the three schemes MUMS-CJ, MUMS-CNJ and PF-CNJ achieve higher cell-edge average throughput, which is the result of supporting CoMP joint transmission, and it can mitigate inter-cell interference and improve the cell-edge user throughput. The improved cell-edge performance is yielded by a better trade-off between joint transmission and interference coordination. The joint transmission power control is much more important to the better throughput performance by observing the 15.75% gain between MUMS-CNJ and MUMS-CJ. By contrasting the PF-CNJ and MUMS-CNJ we find that the proposed MUMS scheduling scheme outperforms the proportional fairness algorithm in terms of throughput.

To increase our understanding of the proposed MUMS scheduling algorithm designed for the CoMP-SU-MIMO system (**Fig. 6**), we plot the video traffic throughput of cell-edge users with respect the heterogeneous service user number in the system.



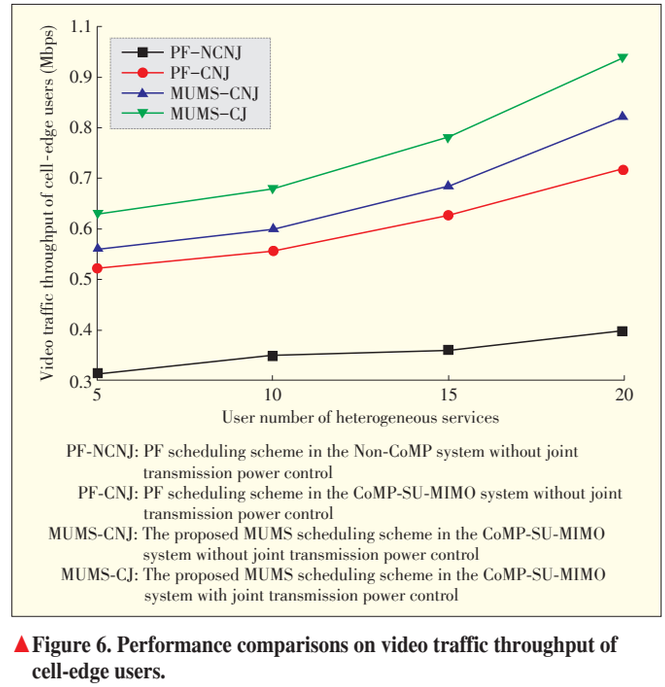
▲Figure 4. Performance comparisons on cell average throughput.



▲Figure 5. Performance comparisons on cell-edge average throughput.

The video traffic throughput of all these scheduling schemes increases as the number of heterogeneous service users increases. That of the proposed MUMS-CJ increases fastest and is obviously higher than that of the others, which indicates the MUMS can satisfy the QoS requirements of RTs well while improving the performance of the system.

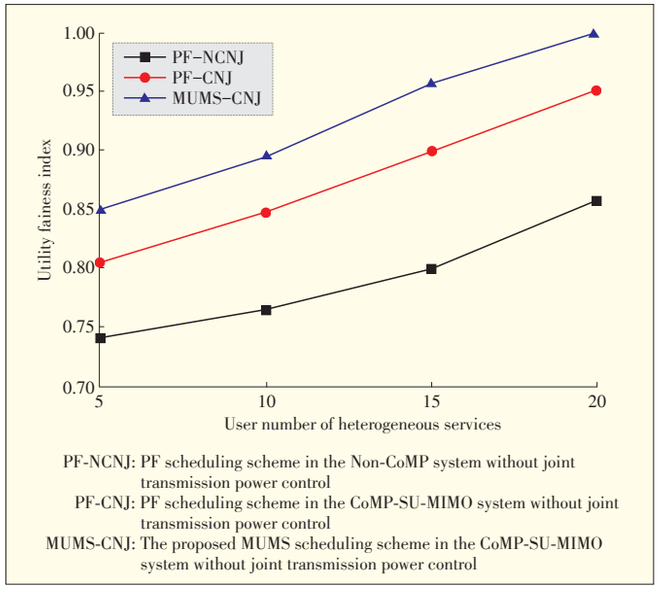
We use Jain's Fairness Index (JFI) to measure the fairness of the resource scheduling scheme. The JFI is based on user utility:



▲Figure 6. Performance comparisons on video traffic throughput of cell-edge users.

$$F=\frac{\left(\sum_{k=1}^{K}U_k\right)^2}{K\sum_{k=1}^{K}U_k^2} \quad (17)$$

The utility JFIs of the four schemes is shown in **Fig. 7**. By exploiting different utility functions, the proposed MUMS scheduling algorithm achieves better user utility fairness than



▲**Figure 7. Performance comparisons on utility fairness index.**

the PF algorithm. Additionally, the joint transmission power control can also help to achieve higher utility fairness by contrast to the MUMS-CNJ and MUMS-CJ algorithms.

6 Conclusion

In this paper, we focus on the downlink of multi-cluster CoMP-SU-MIMO networks with multiple traffic patterns and provisioning QoS requirements. A utility-based joint scheduling algorithm is proposed to address the integrated problem of subcarrier allocation and dynamic subcarrier sharing between multiple services for one user. The goals of the MUMS scheme are to maximize system throughput, fulfill QoS requirements, and reduce computational complexity while considering multiple service classes. We apply JTPC to find the optimal power allocation for any selected user, which could provide most of the achievable throughput gain. The simulation results indicate a significant improvement in cell average throughput, cell-edge average throughput, video traffic throughput of cell-edge users and fairness criterion of the proposed MUMS-CJ.

References

[1] G. J. Foschini and M. J. Gans, "On the limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, no.3, pp. 311–335, 1998. doi: 10.1023/A:1008889222784.
[2] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, 1999. doi: 10.1002/ett.4460100604.
[3] W.-H. Park, S. Cho, and S. Bahk, "Scheduler design for multiple traffic classes in OFDMA networks," *Comput. Commun.*, vol. 31, no. 1, pp. 174–184, Jul. 2007. doi: 10.1016/j.comcom.2007.10.041.
[4] M. Tao, Y. C. Liang, and F. Zhang, "Resource allocation for delay differentiated traffic in multiuser OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2190–2201, Jun. 2008. doi: 10.1109/TWC.2008.060882.
[5] Z. Chen, K. Xu, F. Jiang, Y. Wang, and P. Zhang, "Utility based scheduling algorithm for multiple services per user in mimo ofdm system," in *IEEE Int. Conf. Commun.*, Beijing, China, May 2008, pp. 4734– 4738. doi: 10.1109/ICC.2008.887.
[6] B. Song, R. L. Cruz, and B. D. Rao, "Network duality for multiuser MIMO beamforming networks and applications," *IEEE Trans. Commun.*, vol. 55, no 3, pp. 618–630, 2007. doi: 10.1109/TCOMM.2006.888889.
[7] B. Mielczarek and W. Krzymien, "Throughput of realistic multiuser MIMO-OFDM systems," in *ISSSTA*, Sydney, Australia, Sep. 2004, pp. 434–438. doi: 10.1109/ISSSTA.2004.1371737.
[8] B. Huang, J. Li, and T. Svensson, "A utility-based scheduling approach for multiple services in coordinated multi-point networks," in *14th Int. Symp. Wireless Personal Multimedia Commun.*, Brest, France, Oct. 2011, pp. 1–5.
[9] N. Reider, A. Rácz, and G. Fodor, "On scheduling and power control in multi-cell coordinated clusters," in *IEEE Global Telecommun. Conf.*, Honolulu, USA, 2009, pp. 1–7. doi: 10.1109/GLOCOM.2009.5425622.
[10] L. Zhang and P. Lu, "A utility-based adaptive resource scheduling scheme for multiple services in downlink multiuser MIMO-OFDMA systems," in *IEEE 77th Veh. Technol. Conf.*, Dresden, Germany, Jun. 2013, pp. 1–5. doi: 10.1109/VTCSpring.2013.6691852.
[11] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, pp. 171–178, Feb. 2003. doi: 10.1109/JSAC.2002.807348.
[12] 3GPP. (2004). *3GPP TR 25.8923 V2.0.0*: *Feasibility Study for OFDM for UTRAN Enhancement*. [Online]. Available: http://www.docin.com/p-93113950.html
[13] C. X. Shi, Y. Wang, T. Wang, and L. S. Ling, "Resource allocation for heterogeneous services per user in OFDM distributed antenna systems," in *IEEE 71th Veh. Technol. Conf.*, Taipei, China, May 2010, pp. 1–5. doi: 10.1109/VETECS.2010.5493992.
[14] 3GPP. (2009). *3GPP TR 36.942 v8.2.0*: *Radio Frequency (RF) system scenarios*. [Online]. Available: http://www.docin.com/p-376005379.html

Manuscript received: 2014-10-08

Biographies

Borui Ren (zxzrbr@163.com) received his BS degree from Beijing University of Posts and Telecommunications (BUPT) in 2013.He is now a postgraduate student in Beijing University of Posts and Telecommunications. His research interests include mobile communication theory, mobile internet, and other technologies.

Gang Liu (liugang@bupt.edu.cn) received his PhD degree from BUPT in 2003. Over the past few years, he has participated as principal investigator in two National Natural Science Foundation programs. His research interests include cloud computing, speech recognition, and other technologies. He has published more than 50 papers and holds five nation invention patents.

Bin Hou (robinhou@163.com) received his PhD degree from BUPT in 2007. He currently works as a lecturer at Beijing University of Posts and Telecommunications. He has co-authored *Hadoop Open Source Cloud Computing Platform*. His research interests include information and network security, intelligent information processing, and data mining.

ZTE Communications Guidelines for Authors

• Remit of Journal

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

• Manuscript Preparation

Manuscripts must be typed in English and submitted electronically in MS Word (or compatible) format. The word length is approximately 4000 to 7000, and no more than 6 figures or tables should be included. Authors are requested to submit mathematical material and graphics in an editable format.

• Abstract and Keywords

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

• References

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

• Copyright and Declaration

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

• Content and Structure

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

• Peer Review and Editing

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

• Biographical Information

All authors are requested to provide a brief biography (approx. 150 words) that includes email address, educational background, career experience, research interests, awards, and publications.

• Acknowledgements and Funding

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

• Address for Submission

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

ZTE COMMUNICATIONS

ZTE Communications has been indexed in the following databases:

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