Jonathan Rodriguez
Christos Verikoukis
John S. Vardakas
Nikos Passas *Editors*

Enabling 6G Mobile Networks

Springer

Enabling 6G Mobile Networks

Jonathan Rodriguez • Christos Verikoukis
John S. Vardakas • Nikos Passas
Editors

Enabling 6G Mobile Networks

Springer

Editors

Jonathan Rodriguez
Instituto de Telecommunicações
Campus Universitário Santiago
Aveiro, Portugal

Faculty of Computing
Engineering and Science
University of South Wales
Pontypridd, UK

John S. Vardakas
Iquadrat Informática SL
Barcelona, Spain

Christos Verikoukis
Centre Tecnològic de Telecomunicacions
de Catalunyal
Parc Mediterrani de la Tecnologia (PMT)
Castelldefels, Spain

Nikos Passas
University of Athens, Panepistimiopolis
Athens, Greece

ISBN 978-3-030-74647-6 ISBN 978-3-030-74648-3 (eBook)
https://doi.org/10.1007/978-3-030-74648-3

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my dearest newborn niece
Sienna Rodriguez-Brian
Born on 27thNovember 2020, 3.45am
Frimley Park, Berkshire

Foreword

As Fifth Generation (5G) mobile networks are being rolled out, the telecom industry and academia are now coordinating the 6G research effort towards defining the requirements and use cases for Beyond 5G (B5G) or so-called Sixth Generation (6G) mobile networks. 6G will be more encompassing in terms of communication requirements in contrast to its predecessor, being more society centric in terms of requirements; in addition to the vertical market requirement, it is widely accepted that the 6G drive will be influenced by global policy on sustainability goals for an ageing and growing population, as well as addressing societal challenges. The aim is to deliver a 6G architecture that promotes digital inclusion and accessibility, as well as unlocking economic value and opportunities in rural communities.

This book edition aims to address the ongoing international effort towards the 6G paradigm, through the lens of international European training networks and early-stage researchers. The authors provide an overview on the drive towards 6G by identifying key enabling technologies and system requirements and highlighting developments in the global B5G/6G arena. This provides the impetus for the subsequent chapters that target enabling 6G technologies, as well as advanced 5G technologies, that may become part of future standards.

These works are the key scientific outcomes of three major international research initiatives (H2020-ETN-SECRET, H2020-ETN-SPOTLIGHT, and H2020-ETN-5GSTEPFWD) that address complementary themes on B5G research, including virtualization, wireless-optical interoperability, and UDNs (Ultra-Dense Networks) to provide a significant step towards solving the 6G riddle.

This book is an excellent example of the key collaborative outputs that can be achieved through European training programs and collaboration, that has the potential to influence international research efforts.

I would like to thank the coordinators of SECRET, SPOTLIGHT, and 5GSTEPFWD and the editors that have notable track records in mobile communication, exceptional experience in international collaborative research, and are well poised to guide the interested reader through their take on the 6G odyssey.

Sincerely,

Qatar University, Doha, Qatar

Mohsen Guizani

Mohsen Guizani (S’85–M’89–SM’99–F’09) received his BS (with distinction) and MS in electrical engineering, MS and PhD in computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1986, 1987, and 1990, respectively. He is currently Professor at the Computer Science and Engineering Department in Qatar University, Qatar. He is currently the Editor-in-Chief of the *IEEE Network Magazine*, serves on the editorial boards of several international technical journals, and is the Founder and Editor-in-Chief of *Wireless Communications and Mobile Computing* journal (Wiley). He is the author of nine books and more than 750 publications in refereed journals and conferences. He is the recipient of the 2017 IEEE Communications Society Wireless Technical Committee (WTC) Recognition Award, the 2018 Ad Hoc Technical Committee Recognition Award for his contribution to outstanding research in wireless communications and ad hoc sensor networks, and the 2019 IEEE Communications and Information Security Technical Recognition (CISTC) Award for outstanding contributions to the technological advancement of security. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the Chair of the TAOS Technical Committee. He is a Fellow of the IEEE and a Senior Member of ACM.

Preface

The future society is heading towards an increasingly digitized world that is connected and data driven, where many services will be dependent on instant and virtually unlimited connectivity. As 5th Generation (5G) research reaches the twilight, the research community must go beyond 5G and look towards the 2030 connectivity landscape, namely 6G (6th Generation Mobile Networks). It is worthy to note that it is not clear exactly what 6G will be, but most certainly will consider immature technologies as part of the drive towards beyond 5G, but more specifically it will be influenced by the way in which data is collected, processed, transmitted, and consumed within the wireless network.

5G technology was driven by the commercial operators to accommodate future capacity requirements for their customer base, as well as complemented by efficient manufacturing demands from industry in the shape of IoT (Internet of Things). The technical success of 5G hinges on enabling technology that will deliver a much wider range of data services to a much broader variety of devices and users. However, 6G will be more encompassing in terms of connectivity requirements, and more society centric in design, i.e., in addition to industry business models, it is widely accepted that the 6G drive will be influenced by global policy on sustainability goals for an ageing and growing population, as well as addressing societal challenges. The aim is to deliver a 6G architecture that promotes digital inclusion and accessibility, as well as unlocking economic value and opportunities in rural communities.

Harnessing on the plethora of services offered by 5G technology, 6G aims to integrate an even richer set of services to its portfolio, which include virtual and augmented reality (and even mixed reality), telepresence, and autonomous vehicles for ecological transport and logistics. This will be based on introducing new enabling technology that can target ambitious KPIs (Key Performance Indicators) that factor in 10–100 times more capability over 5G networks. This will require disruptive architectures that can build on 5G technology to deliver market-relevant solutions.

The first roll-out of 5G (2019-2020) is targeting sub 6 GHz small cells based on a Cloud Radio Access Network (C-RAN) architecture. However, subsequent rollouts will deploy the full 5G vision that addresses the hyperdense deployment of small cells based on millimeter wave frequencies, where larger swathes of spectrum are available. This will be coupled with multiple antenna technology deployed on a massive scale that in synergy will have a multiplier effect on user peak data rates and cell capacity, enabling the 5G system to fulfil the ambitious performance indicators specified by 3GPP.

However, going beyond 5G systems, the so-called B5G paradigm is pushing back further the boundaries on communication systems in a bid to introduce tactile internet applications that combine ultralow latency with extremely high availability, reliability, and security. Speed is also a key design requirement in 6G systems. Whereas 4G was about megabit connectivity, 5G pushed the gigabit barrier, and 6G is eventually expecting to deliver theoretical terabit speeds.

To entertain B5G systems will undoubtedly rely on several enabling technologies that already have their footprints engraved within now legacy 5G systems, that includes softwarization, optical networking, and high-frequency (above 6GHz) communications.

Network softwarization is already a main feature in 5G systems, which is expected to further evolve to encompass a complete overhaul of the underlying infrastructure to software; where once the edge network was confined to the operator's infrastructure, it will now include fixed and mobile devices, and may include infrastructure that may not be under the complete ownership of the operator. Indeed, the edge network will migrate to the users' local area space, to include devices in the very near vicinity. This will be a big step towards reducing the latency in the network by enabling the local caching of popular data, as well as content migration. Moreover, computing tasks such as local task offloading will be possible, enabling further virtualization of user handsets and further enhancing battery lifetime.

Optical infrastructures will provide the necessary backhaul capacity to meet the expected demand in data traffic, that will be also be softwarized to provide operators the muscle to adapt the optical infrastructure to harness the dynamic load in the network in an efficient manner.

Despite the ongoing roll-out of 5G, a key bottleneck to ultra-high speed is still the limitations imposed by the underlying spectrum. The market demands towards higher bit rates to entertain virtual reality applications requires the use of higher frequencies over the terahertz (THz) band (0.1-10 THz), which will be key to ubiquitous 6G networks. In particular, THz frequencies have the potential to deliver ample spectrum, over hundred Gigabits per-second (Gbps) data rates, massive connectivity, denser networks, and highly secure transmissions.

Therefore, the merger of the optical, Terahertz wireless, and virtualization technology domains will result in a fully flexible and efficient high-speed communication platform which the research community is envisaging as beyond 5G/6Gcommunications, to enable the mass deployment of small cells to what is referred to as ultra-dense networks (UDNs). This enables not only enhanced broadband connectivity, but delivers a communication medium for emerging very low latency tactile application. It is worthy to note that when we refer to B5G technologies, these can eventually be part of the 6G standard, or part of subsequent 5G releases.

This book edition aims to be one of the first to tackle the 6G odyssey, providing a concerted technology roadmap towards the 6G vision focused on the interoperability between the wireless and optical domain, including the benefits that are introduced through virtualization and software-defined radio. This edition was motivated by two pieces of the jigsaw coming together. On one hand, the niche that currently exists in the literature towards providing the latest developments on the 6G roadmap and, on the other, the opportunity to harness ongoing international research that is currently addressing the 6G story. The outcome resulted in the publication of *Enabling 6G Mobile Networks*, which aims to be at the forefront of beyond 5G technology by reflecting the integrated works of several major European collaborative projects (H2020-ETN-SECRET, H2020-ETN-5GSTEPFWD, H2020-ETN-SPOTLIGHT).

The focus of this edited book was engineered according to the targeted audiences, and should be used by such. These include: practicing engineers and under/postgraduate students working in this field. The edited book includes simple fundamental concepts as primers, as well as current market applications in order to either provide a base on which the interested reader can acquire new knowledge on optical, wireless, or cloud computing, or provide a platform for developing practical applications for modern-day mobile communications. This includes the development of antenna and filter technologies for millimeter wave applications, cloud computing for pooling of resources and prediction, or mobile fronthaul network design and resource management for seamless wired-wireless communications. The book will serve as a useful tool for researchers (academia and industry) to draw inspiration towards the design of innovative protocols/algorithms targeting cloud computing, optical and mobile communications for next-generation systems, including beyond 5G, and 6G; and finally, it provides the inspiration for stakeholders (academia and vendors) to build new project proposals for this highly evolving field. In particular, this book identifies the scenarios, design requirements, and performance indicators for 6G that will provide a basis for building new scenarios, or identifying research challenges that still need to be solved; it will provide in-depth insight on the current state of the art and help identify the missing gaps in order to develop new innovation.

We sincerely hope that you find this book edition insightful and as providing a basis for inspiring further research on this fast-evolving field.

Aveiro, Portugal Jonathan Rodriguez

Castelldefels, Spain Christos Verikoukis

Barcelona, Spain John S. Vardakas

Athens, Greece Nikos Passas

Acknowledgments

As far as the editors are aware, this is the first book of its kind targeting the evolution of 5G towards 6G, to provide a fully informative guide on the emerging Beyond 5G landscape. We focus our vision on the merger of the optical, wireless and virtualization domains to provide an integrated view on the emerging Beyond 5G architecture, that may ultimately become part of late release 5G standard or become a fundamental pillar of the 6G architecture.

We aim to be at the forefront of Beyond 5G technology by reflecting the integrated works of several major European training network (H2020 ETN) projects (SECRET, 5G STEPFWD, SPOTLIGHT). The editors would like to thank all the supervisory teams from the beneficiaries and associated partners, and early-stage researchers that have contributed to this book as respective authors, and have made this contributed book possible.

I would also like to thank both Xavier Gelabert and Georgios Koudouridis from Huawei Sweden in their roles as disssemination and outreaching coordinators for the SECRET project, and for their valuable supervisory contribution to Chapter 3.

Going beyond the training network contributors, the editors would also like to thank Cláudia Barbosa and the Mobile System Research Group from the Instituto de Telecomunicações – Aveiro for their respective contributions towards providing editorial support, and on the 6G international research effort emanating from their H2020-MCSA (EXPLOR, eBORDER) and FCT (i-Five Project – POCI-01-0145-FEDER-030500) project portfolio.

The editors would also like to thank the Mobile and Satellite Communications Group from the University of South Wales, UK, and Prof. Ifiok Otung for their contributions on Terahertz communications through their MCSA International Fellowship program (5GACE).

Finally, we would like to thank the Springer Team – New York for their support throughout the editorial process.

Introduction

The 6G vision is wide-reaching to cater for many vertical sectors, and it is not the purpose of this book to cover all aspects on 6G, but to provide a primer on the "fundamentals of 6G Mobile networks" that is reliant on UDN technologies, optical-wireless convergence, and cloud-based services through softwarization; as such this book is self-contained and structured accordingly.

This book constitutes four parts and 15 chapters, which include:

Part I, "**5G and Beyond Mobile Landscape**," that constitutes a comprehensive chapter outlining the 5G and beyond landscape, providing the interested reader with insights into emerging 6G technologies and system requirements. The intention is to position the works from the respective authors to highlight better the impact of their contribution within the B5G/6G paradigm; this will be elaborated in Chap. 1.

Part II, "**UDNs for B5G**," will highlight the UDN core, and provide insights into technology challenges and solutions for enabling the hyper-dense deployment of small cells. B5G will harness massive-scale MIMO and distributed architectures to support the notion of cell-free MIMO, and therefore the optimal operating conditions are investigated here, giving valuable insights into their future deployment. Moreover, central to 6G-enabling technologies is RIS (Reconfigurable Intelligent Services), using which the channel propagation conditions are modified through reflective surfaces, opening up new opportunities going beyond typical signal processing enhancing techniques on legacy transceivers. We aim to provide an overview on the latest development in this area. Networking resources represent a significant cost to the operator in terms of capital (CAPEX) and operational (OPEX) expenditure. Therefore, how to attain the best use of the available resources is always on the operator's agenda. This part addresses optimal radio resource management techniques that assume cooperation within distributed antenna systems. Moreover, the security and RF challenges in UDN networks are also addressed here. All these topics will be elaborated in Chaps. 2, 3, 4, 5, and 6.

Part III, "**PON technology for UDNs**," outlines how passive optical networks (PONs) will play a role in integrated optical-wireless fronthauling. Optical access standardization has evolved rapidly, with broadband technologies such as 10G-capable passive optical networks (XG-PON) and the next-generation (NG) of PON

(NGPON2) already widely deployed. Moving forward, towards meeting the ever-growing bandwidth demands and supporting the said mobile access traffic, the optical technology roadmap is heading beyond 100G-PON as a vehicle for enabling higher spectral efficiency and reduced CAPEX/OPEX costs; indeed, the market dictates deploying NG networks as part of a convergence solution. Coexistence with wireless and legacy PON technologies thus constitutes a vital point in the NGPON roadmap, and is widely recognized as the mobile fronthaul enabler and generally referred to as RoF (Radio over Fiber) technology that raises significant challenges in terms of finding the optimal modulation waveforms, and detection due to the presence of optical nonlinearities. In this context, these issues are addressed in this section, which also presents proposals for practical integrated optical-wireless interfaces and new insights into practical fronthaul performance. Moreover, future emerging networks will be softwarized, which also means an integrated optical-wireless fronthaul component. How virtualization is applied to the fronthaul, and the opportunities that arise in terms of network slicing and functional split will be detailed here. This will be covered in Chaps. 7, 8, and 9.

Part IV, "**Cloud based UDNs for beyond 5G**," addresses the cloud-based pillar of the book. One of the core aspects of B5G platforms is enabling network virtualization, a networking paradigm borrowed from the cloud computing world that supports highly configurable and dynamic allocation of resources and overall system flexibility. This will open up new opportunities in terms of lowering the cost of ownership for mobile network operators, as well as enhancing the QoS (Quality of Service) for end-users from lower latency and reliable service provisioning to lowering battery consumption in handset devices. In this context, this part will address resource management in the C-RAN core, where we consider computation as well as radio resources in synergy. Also, the management of baseband computing resources will be addressed in terms of planning with backup resources, what the authors refer to as shared-path shared-compute planning (SPSCP) strategies. The authors will also shed light on the specific challenges associated with so-called network slicing, functional split, and their self-automation through the use of AI (artificial intelligence). Also, cloud computing is employed to the edge network in emerging 5G systems; however, this section investigates a new definition where the edge network now includes mobile devices in the near vicinity. This will take a new step towards introducing further gains from the use of virtualization in emerging B5G networks. This progressive paradigm is addressed in terms of investigating a novel content distribution approach for B5G systems that is reliant on local mobile nodes, and how this can be demonstrated using open-source virtualization tools. Finally, how the cloud can be used for effective content management through caching is also addressed. These topics are elaborated in Chaps. 10, 11, 12, 13, 14, and 15.

The specific organization of each chapter is as follows:

Chapter 1, Drive Towards 6G As 5G mobile networks are being rolled out, the telecom industry and academia are now coordinating the 6G research effort towards defining the requirements and use cases for beyond 5G or so-called 6G mobile networks. 6G envisages an evolutionary communication platform based on complete network softwarization, inclusive communications mediums, and UDNs to cater for the mass market demands, that are envisaging ultrahigh speeds, tactile response time, and lower cost of network ownership by 2030. This chapter provides an overview of the drive towards 6G. In this context, we first remind the reader on where we are now in terms of 5G standardization, which acts a useful baseline to position the 6G efforts. Subsequently, we define the 6G use-case or potential applications that provide the impetus for the consensus on the 6G system requirements. Thereafter, we revise the enabling technologies that are actively being explored as potential suitors for the 6G paradigm, and conclude our perspective by reviewing global initiatives on B5G/6G activities.

Chapter 2, Cell Free MIMO System for UDNs Distributed or cell-free (CF) massive MIMO (MaMIMO) is predominantly becoming a promising key enabling technology for beyond 5G systems, which aims to deliver huge increases in network capacity and quality of service. The CF MIMO architecture aims to deliver localized high coverage hot spots by harnessing the massive number of distributed transmit and receive antennas. The distributed nature allows the user equipment to escape the rigid cell boundaries of legacy mobile network to have access to multiple access points that are spatially distributed. The gain in capacity or reliability will be influenced by the number of antenna elements and exploiting antenna diversity. Chapter 2 aims to explore the mechanics behind CF MIMO systems, which not only includes the performance analysis under favorable radio conditions, but also on the waveform type and detection. In the first instance, favorable propagation properties are studied for a CF MaMIMO system for both line of sight (LoS) and multipath Rayleigh fading channels, where also the advantages of distributed vs. centralized arrays are presented using 3D beamforming, including a low complexity partially centralized zero-forcing technique to cancel interference. To deal with the asynchronous transmission between network nodes in CF architectures, filter bank-based multicarrier (FBMC) waveforms have been proposed as a potential alternative to legacy 5G CP-OFDM (Cyclic Prefix Orthogonal Frequency Division Multiplexing) due to better spectral efficiency and robustness towards synchronization errors. The key drawback of FBMC systems is the high intrinsic interference caused by the loss of complex orthogonality between subcarriers. To address the interference problem, Chap. 2 proposes an iterative interference cancellation (IIC)-based bit-interleaved coded modulation with an iterative decoding (BICM-ID) receiver, which is compared to the baseline CP-OFDM waveform.

Chapter 3, Radio Resource Management and Access Polices for B5G In emerging 5G and beyond wireless technologies, UDNs will be employed to serve a massive number of devices with mobile access. Although UDNs can provide a basis for high-throughput systems, their capability is still largely potential and ideal rather than practical, due to the challenging conditions associated with the networking environment, user mobility, and the stringent 5G design requirements. There are still several challenges to solve in terms of radio resource management and access policies that require new innovations towards enhancing and optimizing the deployed 5G infrastructure, where undoubtedly any future uptake will become part of a future 5G/6G release. One of the major challenges in UDNs is interference due to the close deployment of base stations. Game theory can be used for modeling various Radio Resource Management (RRM) problems that appropriately manages this interference in the networks. In this context, Chap. 3 explores coalitional games for characterizing and solving the user association problem that relies on cooperation among players. This is further extended to demonstrate how they can be utilized to design RRM for 5G and beyond use-cases that rely on CoMP (Cooperative Multipoint). Moreover, mobility requirements are becoming more challenging, both in terms of robustness against handover (HO) failure and reducing energy consumption. Chapter 3 will investigate the mobility problem in mobile networks employing small cell technology such as 5G, and how UL (Uplink)-measurement-based RS (received signal) HO schemes can reduce the HO signaling overheads and power consumption in contrast to legacy downlink-measurement-based approaches. A key use-case for 5G is Machine Type Communications (MTC) or the Internet of Things (IoT). Resource optimization in MTC is gaining increased attention, being responsible for attaching the MTC devices to the network. This is creating huge challenges due to the massive amount of devices contending for resources that will create excessive collisions at the RAN. Therefore, we analyze how the NB-IoT (CIoT technology) RAN parameters "repetition" and "retransmission" can be optimally allocated to reduce the collision rate.

Chapter 4, Energy-Efficient RF for UDNs This chapter provides insights into the design of energy-efficient and multi-standard RF front-end for next generation multi-homing small cell devices. Next-generation UDNs need to be green, or in other words "energy aware," so as to support future emerging smart services that are likely to be bandwidth hungry, as well as support multi-mode operation (5G, LTE, LTE-A, HSDPA, 3G among others). In this context, this chapter targets the RF front-end architecture that considers functional blocks harnessing current RF design standardization and spectrum policy design requirements to provide concrete solutions in terms of reconfigurable filtenna structures, antenna design, and power amplifiers. In particular, we consider MMIC (Monolithic Microwave Integrated Circuit) technology for Power Amplifier (PA) design targeting 5G handsets, where MMIC offers compact circuit configurations and the confinement of electromagnetic fields within the semiconductor materials. Moreover, the PA design is also considered for the base station application, in the form of energy-efficient Load Modulation PAs based on the Doherty technique, that offers benefits such as simplicity and lack of complex circuitry. This is fabricated and tested in the lab to investigate the linearity-output power trade-off. Furthermore, hybrid antenna-filter techniques are considered. The first contribution in this category is referred to as the Differentially Fed Reconfigurable Filtering Antenna for mid-band 5G applications that offers promising characteristics such as multifunctional property, high common-mode suppression, high roll-off skirt selectivity, and low radiated power loss, among others. This is employed here to develop a filtering antenna with reconfigurability/tenability function at the heart of the design. The second contribution is the compact filtering antenna for Phased Arrays targeting 5G mmWave FDD backhaul applications. The new wireless fronthaul/backhaul network in 5G and beyond is expected to be reconfigurable to satisfy the dynamic nature of mobile traffic. Phased Array Antenna (PAA) in the backhaul equipment is a promising solution to respond not only to this design requirement, but also enable the system with the ability to automatically recover the link when misalignments occur, increasing its availability. Finally, the insensitive Phased Array Antenna for 5G smartphone applications is proposed, where the insensitive property provides robust and consistent performance for different antenna substrate materials, including the handheld effect that is pivotal in handset applications.

Chapter 5, Security for UDNs: A Step Towards 6G The next-generation mobile networks are focusing on small cells technology, resulting in the formation of UDNs. These can be considered as a wireless network of mobile small cells (MSCs), where these mobile devices are virtual in nature. They can offer ultra-high speed ubiquitous connectivity on demand and increased energy efficiency, whereas the mobile network infrastructure benefits from a reduction in traffic due to offloading. However, MSCs can potentially face a variety of security and privacy challenges. This chapter covers three important security infrastructures: (i) decentralized key management schemes, (ii) intrusion detection and prevention schemes, and (iii) blockchain-based integrity schemes. Decentralized key management enables the heterogeneous mobile devices to securely exchange cryptographic keys. These cryptographic keys can then be utilized by blockchain-based integrity schemes to establish secure and reliable communication channels between mobile devices, even in the presence of malicious adversaries. The intrusion detection and prevention schemes attempt to identify and remove these malicious adversaries from the network. These security infrastructures can potentially be used as stepping stones towards a security architecture for general MSC architectures and 6G networks.

Chapter 6, Channel Estimation in RIS-Aided Networks Reconfigurable intelligent surface (RIS) is a recently emerging transmission technology for wireless communication applications. Regarded as an emerging solution for the next-generation 6G, RIS is a nearly passive device that realizes a smart radio environment with low hardware cost and energy consumption. In particular, RIS is a two-dimensional surface made of metamaterials that is capable of manipulating the incident electromagnetic waves in arbitrary ways. The main selling point of RIS is its near-passive nature, since it does not require a large power source to redirect the waves, coupled with low cost and complexity of large-scale deployments. However, these benefits are complemented by major challenges that need to be addressed in terms of channel estimation in RIS-aided communication systems. Recently, many protocols and algorithms are proposed to handle this problem. In this chapter, we review the latest development on channel estimation in RIS-aided systems and suggest research questions that still need to be solved.

Chapter 7, Integrated Optical-Wireless Interface and Detection PONs are associated with fiber-to-the-home (FTTH) connections providing broadband and high-speed communications. However, the 5G market has shaped the design requirements for mobile networking towards even higher-capacity networks to cater for the foreseen demand in traffic. This has spurred a new viewpoint in backhaul networking involving the gradual migration from the existing WDM (Wave Division Multiplexing)-PONs to ultra-dense WDM-PONs within urban areas. A dynamic way to achieve these requirements is by employing hybrid photonic-wireless links operating in the lightly licensed millimeter wave (mm-wave) bands that coexist with UDWDM-PONs, raising new challenges in terms of seamless interoperability and signal detection. The remainder of this chapter aims to answer these challenges, where the authors target to provide solutions towards the integration of optoelectronic integrated components within the 5G base stations, as well as to present digital signal processing techniques for the mitigation of distortions caused to FTTH signals by the passive optical network. In the first instance, the physical concepts describing optical heterodyning and the operation of uni-travelling carrier photodiodes (UTC-PDs) are explained. Moreover, the processes on the acquisition of the full-equivalent circuit of these devices are analyzed by using both simulation tools as well as mathematical equations. Furthermore, the properties of transimpedance-amplifiers (TIAs) is elaborated in terms of gain and noise. Finally, the results of the co-simulation between a UTC-PD and multistage TIA operating in V-band are discussed, underlining all the important aspects of high-speed, high-gain optoelectronic co-integration. For FTTH applications, the advanced 25G avalanche photodiodes with high sensitivity are costly in the passive optical networks. Thus, the combination of semiconductor optical amplifiers (SOA) used as pre-amplifiers with photodiodes (PD) is proposed as a low-cost solution to improve system power budget. However, the SOA nonlinearities due to the gain saturation, e.g., the pattern effect degrades system performance significantly, which worsens with increasing system speed. Hence, digital signal processing (DSP) is often used as a powerful tool, and combined with the use neural networks, has the potential to compensate for linear and nonlinear distortions. The performance of NN is investigated within an intensity modulation with direct detection (IM/DD) system with 50G PAM4 signals, showcasing the nonlinearity compensation and receiver dynamic range improvement.

Chapter 8, Modulation and Equalization Techniques for mmWave ARoF The new 5G architecture includes an integrated optical-wireless network architecture, where optical technology will complement radio in order to handle the new capacity demands over both the backhaul and fronthaul network, leading to the notion of Analog radio-over-fiber (ARoF). ARoF implies attractive benefits such as wide area coverage, high spectral efficiency, and reduced power consumption, among others. Nevertheless, the combination of these technologies implies new challenges to solve, where the impact of combined mmWave radio and optical channel impairments (high free-space path loss (FSPL), chromatic dispersion, and phase noise) will affect the radio detection and link performance. This chapter aims to study and analyze techniques to reduce the degradation introduced by the mmWave ARoF channel by revisiting the signal processing in the radio-optical transceiver link. The radio modulation format selection is key to receiver performance; however, there are limited studies available for investigating the impact of legacy and emerging modulation schemes on mmWave ARoF systems. In this context, the authors compare, experimentally, modulation candidates for mmWave ARoF and provide new insights into their performance, which suggests that legacy OFDM might not tick all the boxes as it once did; moreover, in such as a converged system, the analogue radio signal will be subject to chromatic dispersion in the standard single mode fiber (SSMF), spurring the need for equalization to compensate dispersion in the fiber. Therefore, this chapter studies channel equalization at the radio receiver based on a simulated mmWave ARoF scenario, and provides new insights in performance. Optical amplifiers and modulators are crucial devices in mmWave ARoF systems. The REAM (reflective electro-absorption modulator)–SOA (Semiconductor Optical Amplifier) integrated into a single chip is investigated as an alternative to directly modulated lasers (DMLs) in the optical link, where EAM-based transmitters have the potential to provide better transmission performances because of the absence of adiabatic chirp. Moreover, the SOA is sought to increase signal propagation distances to envisage the 5G coverage requirements. In this context, the authors investigate the device and optical link performance in terms of key parameters such as extinction ratio, insertion losses, and gain. In particular, an experimental digital transmission is demonstrated by utilizing this device, achieving a bit rate of 50-Gb/s.

Chapter 9, Optical-Wireless System Performance, Deployment and Optimization As the 5G milestone approaches, there needs to be a concerted effort towards practical performance evaluation, deployment strategies, and optimization in order to fully capitalize on the 5G benefits and KPIs (Key Performance Indicators). This chapter aims to provide just that. The authors provide new insights on the optical-wireless link performance for converged FiWi/mmWave 5G networks, which are characterized by wide steering angle (90° degrees) and multi-user support. The performance is evaluated for enhanced Mobile Broadband (eMBB) and Dense Fixed Wireless Access (FWA) hot spot scenarios. Regarding the deployment strategies, we also consider how integrated optical-wireless networks can be deployed as a small cell network. The authors target small cell deployments for real hot spot scenarios considering the ARoF fronthaul, where a live sports event is considered as a use-case; this provides the impetus for general fronthaul deployment guidelines. Moreover, assessing the mobile network performance is pivotal to ensuring we are attaining the planned QoE (Quality of Experience) targets. We first provide a comparative study between vendor-specific and open-source small cells that is referred to as the OAI (Open Air Interface), which is commonly used for in-house testing. This is used as a basis for studying the holistic view on network performance using the IxChariot application. The objective is not only to assess the viability of IxChariot as an experimental tool, but to compare the OAI with vendor-specific equipment. The IxChariot platform instantly assesses network performance, including wireless performance by using a simple server-client topology. Not only do we investigate the system performance and practical deployment approaches, we also address their optimization based on ML approaches. Virtual resource management and big data represent two prominent 5G paradigms that offer new opportunities in terms of network management and optimization. In this chapter, we consider how ML can play a major role in network optimization by bringing intelligence to the limelight, by investigating how learnt patterns or relationships between network states and QoE can be used towards intelligent prediction and optimization.

Chapter 10, Virtual Networking for Lowering Cost of Ownership 5G and beyond mobile networks generations are heading towards a predominantly softwarized network capitalizing on key technology trends such as virtualization and autonomous management. These technology solutions will undoubtedly offer significant benefits to all stakeholders: users have the potential to experience enhanced QoS; mobile network operators (MNOs) can benefit from a more cost-effective network, while opening up new opportunities for business models and actors. Even though the fundamental virtualized RAN architecture is fairly mature, there is somewhat limited insight into the way we can manage the available resources. To this extent, this chapter is devoted to providing some design recommendation to the planning of virtual networking infrastructures, and to resource allocation (RA) design. In the first instance, we highlight the benefits of virtualization and autonomous technologies when applied to the RAN infrastructure. This provides a basis for the proposal of a cost-efficient strategy referred to as shared-path shared-compute planning (SPSCP) that assigns a primary and a backup RCC (Radio Cloud Centre) node to each RAU (Radio Access Unit). To decrease the overall cost of the network, the SPSCP strategy tries to maximize sharing of the backup connectivity and computing resources. Thereafter, we migrate from resource backup planning to the RA problem, i.e., how we decide on the optimal resource allocation to serve the subscriber group in terms of both computational and radio resources. In this context, a literature review is conducted on existing approaches for RA, highlighting the existing technology gaps and open research challenges. Finally, a hybrid Resource Allocation design aiming at improving energy efficiency (EE) on the C-RAN is proposed.

Chapter 11, Advanced Cloud-Based Network Management for 5G C-RAN C-RAN (also referred to as centralized RAN) is a pivotal part of 5G, where radio functionalities are decoupled into Base Band Units (BBU) and Remote Radio Heads (RRH), where the BBU is centralized from multiple sites into a single geographical point such as a cloud data center. Such a technology comes with minimal cost, high energy efficiency, and centralized signal networking that has the potential to enhance the radio performance through mitigating interference. Although C-RAN appears to be a promising access architecture, the efficient management of the resources in C-RAN to satisfy traffic demand is still a significant challenge due to user mobility and the dynamic nature of the networking environment. In this context, Chap. 11 aims to shed some light on the specific challenges associated with C-RAN management and potential solutions based on network slicing, functional split, and their self-automation through the use of AI. The authors present a joint slice-based C-RAN solution by exploring different functional split configurations between the central and distributed units. Using this suggested approach, the proposed framework was shown to enhance the fronthaul (FH) network infrastructure scalability, as well as providing enhanced QoS. Moreover, a novel real-time RAN network slicing approach is presented. Given the maximum capacity of the radio interface, a novel radio resource management algorithm is proposed where the dimension of each slice is dynamically defined through the joint evaluation of the tenant's Service Level Agreement (SLA) and the real-time traffic performance. The framework is tested on a 5G research experimental testbed using real radio and user equipment. Finally, self-automation is presented through AI applied to zero-touch networks based on C-RAN architecture.

Chapter 12, Resource Management for Cost-Effective Cloud Services Virtualization and SDN technologies play a pivotal role in 5G networks, and will continue to evolve as we head towards 6G networking. Virtualization is envisioned to enable full flexibility in mobile networks by partitioning network functions into virtual instances which can be deployed on general-purpose hardware and moved dynamically in the network (Cloud or Edge of the network) according to requirements. Therefore, resource management is vital to provide optimal allocation of network resources in response to the underlying mobile network demands, and offers huge benefits in terms of savings in Total Cost of Ownership (TCO). Resource management raises significant challenges in terms of scaling, allocation, migration, and optimization. In this context, Chap. 12 aims to tackle these challenges head on by elaborating on recent developments in this area. We address how resources are pooled, forecasted, and migrated between abstract servers to have computing resources on demand. This is demonstrated within a fog-enabled C-V2X architecture for distributed 5G applications, as well in UDNs where the collaboration between MECs enables cloud migration, among other services. Furthermore, network management functions can benefit from the virtualization of radio access networks. The pooling of computational resources for implementing RAN functions according to cell load requirements in 5G RAN can provide cost-effective centralized detection at the MEC (Mobile Edge Computing) node; how this can be done with ML Machine Learning (ML) is explored herein. Furthermore, the problem can be further extended by considering the optimization framework for jointly optimizing and managing MEC resources, which includes factoring in admission control for user requests, resource calendaring (scheduling), and bandwidth constrained routing, as well as the determination of available nodes that provide computing and storage capacity.

Chapter 13, Demonstrating Cloud-Based Services for UDNs: Content Distribution Case Study Going beyond 5G deployment, the market will continue to grow and new delay stringent application will emerge that will require the 5G architecture to evolve. This will drive further the innovation of legacy technology enablers such as softwarization and small cell technology. Softwarization introduces massive flexibility into managing networks effectively, where the future envisages a complete virtualization of the network that includes mobile devices acting as an additional pool of networking resources. While small cell technology will evolve to become ultra-fast hyperdense networks to support ultrahigh speeds, the question arises as to how we can exploit UDNs and virtualization for enhancing the delivery of so-called enhanced broadband services and their experimentation. In this context, Chap. 13 will first review the latest developments in enabling technology tools and data dissemination in cellular networks as a basis for the proposed novelty. Moreover, the notion of energy-efficient content distribution for virtual UDNs is proposed, where we harness virtualization, long-range and local area networking capability, to develop a new networking topology for cost-effective data distribution, i.e., for effective traffic offloading to the small cell network. Moreover, NCC is investigated as an overlay technology for enhancing network resiliency. This concept is evaluated using a new experimental tool that was purposely designed for testing cloud-based services within a virtual MSC (Mobile Small Cell) environment. In this context, we discuss a framework for testing SDN-based services for 5G and beyond, and will showcase how content distribution can be effectively implemented within an SDN-cooperative-based ecosystem as a case study.

Chapter 14, SDN-Based Resource Management for Optical-Wireless Fronthaul Going beyond 5G, the next-generation architecture aims to provide an integrated communication platform as a service in order to handle the different types of devices and varied traffic loads. In this context, many operators are moving to software-defined networking (SDN) and network function virtualization technologies (NFV). These technologies help softwarize and virtualize the network architecture and management plane to create enhanced communication capabilities and resource optimization techniques, such as network slicing and functional split. In addition, network softwarization helps to reduce the huge investments implied by 5G, due to high capacity and low latency requirements. This chapter investigates the applications of softwarization based on fronthaul functional split service through programmable networks, and network slicing either as a service (network slicing use-case) or as a technique (flexible functional split decision use-case). By employing these technologies in synergy through the SDN paradigm, significant gains could be achieved in terms of efficient QoS provisioning.

Chapter 15, Cloud-Based Content Management for B5G The explosive growth in multimedia applications and content over the last decade has placed enormous demands on bandwidth and computationally limited backhaul networks, which is likely to continue as we head towards the 6G era. The ability to manage this content in terms of caching offers multiple benefits such as network offloading, service latency, and cost reduction, which results in enhanced cellular network performance. However, to find the optimal content caching strategy is a challenge within itself, which is often modelled as an optimization problem to increase QoE (Quality of Experience)-related parameters (such as service latency, throughput, cache hit probability, energy, etc.) under resource (such as cache size, computation, bandwidth, etc.)-limited networks. In Chap. 15, we build on the paradigm, by exploring not only the inherent flexibility that the underlying virtual infrastructure can provide in terms of caching services, but also how we use machine learning to self-automate and enhance the caching updating policy. This also includes a perspective on the management of content computation. In this context, the authors first explore learning-based content management and propose a content caching technique using DL (Deep Learning). As a case study, the impact of user mobility and content popularity estimations on performance was investigated to obtain the optimal content update technique. Furthermore, the problem was revisited using the generalized knapsack scheme in synergy with DL. Finally, the authors also propose task offloading harnessing on MEC (mobile edge computing) capability aiming to reduce the power consumption of the system by enabling the notion of truly virtualized handsets. In this scenario, the authors extend the notion of the edge network to include mobile devices/helper nodes in the near vicinity to provide computation services. This is an emerging 6G paradigm that is referred to as “Dew Computing,” which is exploited here towards offloading task content.

Contents

Part I 5G and Beyond Mobile Landscape

Part II UDNs for B5G

Part III PON Technology for UDNs

Part IV Cloud Based UDNs for Beyond 5G

About the Editors

Jonathan Rodriguez received his master's degree in Electronic and Electrical Engineering and PhD from the University of Surrey (UK) in 1998 and 2004, respectively. In 2005, he became a researcher at the Instituto de Telecomunicações, Portugal, and Senior Researcher in 2008, where he established the Mobile Systems Research group. He has served as project coordinator for major international research initiatives (Eureka LOOP, FP7-ICT-C2POWER, H2020-ITN-SECRET, NATO PHYSEC) while serving as technical manager for FP7-ICT-COGEU and FP7-SEC-SALUS. His project portfolio includes participation as Principal Investigator in over 45 international competitive research grants. In 2017, he became a full professor at the University of South Wales (UK). He has trained over 15 doctoral students and 24 post-doctoral researchers. His professional affiliations include: Senior Member of the IEEE (2013), Chartered Engineer (CEng) (2013), Fellow of the IET (2015), and Senior Fellow of the HEA (2020). He has authored over 600 publications that include 170 journals and 10 book editorials. His research interests include: B5G mobile networking architectures, radio resource management, and security.

Christos Verikoukis received the his PhD from the Technical University of Catalonia, Barcelona, Spain, in 2000, in the area of broadband indoor wireless communications. He is currently a Research Director with Telecommunications Technological Centre of Catalonia (CTTC/CERCA), Spain, and an Adjunct Professor with UB. He has authored 138 journal papers and more than 200 conference papers. He has coauthored more than three books, 14 chapters, and four patents. He is serving as the project coordinator of the H2020 MonB5G, MARSAL, 5GMediaBUB and SEMANTIC projects. He and has served as the Principal Investigator of national projects. He has participated in more than 35 competitive projects. He has supervised 15 PhD students and five postdoctoral researchers. Dr. Verikoukis is serving as the IEEE ComSoc EMEA Director, IEEE ComSoc Board of Governors member and a Member-at-Large of the IEEE ComSoc GITC. He received the Best Paper Award at the 2011 and 2020 IEEE International Communications Conference, the IEEE GLOBECOM 2014 and 2015, and the 2016 European Conference on Networks and Communications (EUCNC), and the EURASIP 2013 Best Paper Award of the *Journal on Advances in Signal Processing*.

John S. Vardakas received a Dipl.-Eng. in Electrical Computer Engineering from the Democritus University of Thrace, Greece, in 2004 and his PhD from the Electrical Computer Engineering Dept., University of Patras, Greece, in 2012. He is a Senior Researcher at Iquadrat Informatica, Barcelona, Spain, since 2012. He has authored more than 38 journal articles and 65 conference articles, while he has participated in more than 15 competitive research programs. His research interests include teletraffic engineering, performance evaluation of wireless and optical networks, resource management of communication networks, protocols' design for Radio-over-Fiber/Fiber-Wireless networks, MAC algorithms' development for WLANs and WSNs, and algorithms' design for Demand Response in Smart Grid environments. He is a Senior Member of the IEEE and the Technical Chamber of Greece (TEE).

Nikos Passas received his Diploma (honors) from the Department of Computer Engineering, University of Patras, Greece, and his PhD from the Department of Informatics and Telecommunications, University of Athens, Greece, in 1992 and 1997, respectively. He is currently a member of the teaching staff in the Department of Informatics and Telecommunications of the University of Athens, and a group leader of the Green, Adaptive and Intelligent Networking (GAIN) research group in the department. Dr. Passas has served as a guest editor and technical program committee member in prestigious magazines and conferences, such as the *IEEE Wireless Communications Magazine*, *Wireless Communications and Mobile Computing Journal*, IEEE Vehicular Technology Conference, IEEE PIMRC, IEEE Globecom, etc. He has published more than 120 papers in peer-reviewed journals and international conferences and has also published 1 book and 11 book chapters. His research interests are in the area of mobile network architectures and protocols.

Abbreviations

1D	One-dimensional
2D	Two-dimensional
2DBF	Two-dimensional Beamforming
3D	Three-dimensional
3DBF	Three-dimensional Beamforming
3GPP	3rd Generation Partnership Project
4G	Fourth Generation Mobile Networks
50G-EPON	50 Gb/s Ethernet PON
5G	Fifth Generation Mobile Networks
5G NR	5G New Radio
5GC	5G Core
5GIC	5G Innovation Centre
5G-PPP	5G Public Private Partnership
6G	Sixth Generation Mobile Networks
6GFP	6Genesis Flagship Program
6GIC	6G Innovation Centre
A/D	Analog-to-Digital
AaaSFC	Application-as-a-Service Function Chain
ADC	Analog to Digital Converter
ADSL	Asymmetric Digital Subscriber Line
AI	Artificial Intelligence
AIaaS	AI as a Service
AM/PM	Amplitude Modulation/Phase Modulation
AMF	Access and Mobility Management Function
AoA	Angle of Arrival
AoD	Angle of Departure
AP	Access Point
APD	Avalanche Photodiode
APIs	Application Programming Interfaces
APR	Avalanche Photoreceiver
AR	Augmented Reality

ARF	Anti-Reflection
ARoF	Analog Radio-over-Fiber
ASE	Amplified Spontaneous Emission
ASK	Amplitude Shift Keying
ATIS	Alliance for Telecommunications Industry Solutions
AWG	Arrayed Waveguide Grating
AWVG	Arbitrary Waveform Generator
AWGN	Additive white Gaussian noise
B5G	Beyond 5G
BBU	Baseband Unit
BER	Bit Error Ratio
BH	Backhaul
BICM-ID	Bit-Interleaved Coded Modulation with Iterative Decoding
BPF	Band-Pass Filter
BPM	Bin Packing Minimization
BS	Base Station
BT	Beam Training
BtB	Back-to-Back
BW	Back-off Window
CAPEX	Capital Expenditure
CCO	Capacity and Coverage Optimization
CCP	Content Classifier Policy
CD	Chromatic Dispersion
CDF	Cumulative Distribution Function
CE	Coverage Enhancements
CF	Cell-Free
CH	Coalition Head
CHR	Cache Hit Ratio
CIoT	Cellular IoT
CL-PKC	Certificateless PKC
CM	Coalition Member
CMN	Content Management Node
CMRI	China Mobile Research Institute
CN	Core Network
CND	Check Node Decoder
CNN	Convolutional Neural Network
CO	Central office
CoMP	Coordinated Multipoint
CP	Cyclic Prefix
CP	Content Providers
CP-OFDM	Cyclic-Prefix Orthogonal Frequency Division Multiplexing
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
CR	Coding Ratio
CRAN	Cloud/Centralised Radio Access Network

C-RAN	Cloud/Centralised Radio Access Network
CS	Compressed Sensing
CSF	Cost Scaling Factor
CSI	Channel State Information
CSMF	Communication Service Management Function
CST	Computer Simulation Technology
CU	Central Unit
CUs	Central Units
C-V2X	Cellular-V2X
D/A	Digital-to-Analogue
D2D	Device to Device
DAC	Digital-to-Analog Converters
DAS	Distributed Antenna System
DC	Data Center
DCF	Dispersion Compensating Fiber
DeMUX	De-Multiplexer
DER	Dynamic Extinction Ratio
DFE	Decision Feedback Equalizer
DFT	Discrete Fourier Transform
DI	Delay Interferometer
DL	Deep Learning
DL/UL	Downlink/Uplink
DML	Directly Modulated Laser
DMM	Distributed Mobility Management
DMT	Discrete Multi-Tone
DN	Data Networks
DNN	Deep Neural Network
DoS	Denial-of-Service
DP	Dynamic Programming
DPA	Doherty Power Amplifier
DPO	Digital Phosphor Oscilloscope
DP-PSK	Dual Polarization PSK
DP-QPSK	Dual Polarization QPSK
DQN	Deep Q-network
D-RAN	Distributed RAN
DRL	Deep Reinforcement Learning
DRoF	Digital Radio-over-Fiber
dRoF	Digitalised Radio over Fibre
DSP	Digital Signal Processing
DU	Distributed Unit
DWDM	Dense Wavelength Division Multiplexing
E/O	Electro-Optic
E2E	End-to-End
EAM	Electro-Absorption Modulator
EC	Edge Cloud

EC-GSM	Extended Coverage – GSM
ECL	External Cavity Laser
eCPRI	Enhanced Common Public Radio Interface
EDFA	Erbium Doped Fiber Amplifier
EE	Energy Efficiency
EER	Envelope Elimination and Restoration
eLAA	Enhanced Licensed-Assisted Access
EM	Electromagnetic
eMBB	Enhanced Mobile Broadband
eMTC	Enhanced MTC
EPA	Extended Pedestrian A
EPC	Evolved Packet Core
EPON	Ethernet Passive Optical Network
ET	Envelope Tracking
ETSI	European Telecommunication Standards Institute
ETU	Extended Typical Urban
EU	European Union
EVA	Extended Vehicular A
FBGs	Fiber Bragg Gratings
FBMC	Filter Bank-based multicarrier
FBS	Femto Base Station
FDD	Frequency Division Duplexing
FD-TTP	Fully Distributed TTP
FFE	Feed Forward Equalizer
FFT	Fast Fourier Transform
FH	Fronthaul
Filtenna	Filtering Antenna
FiWi	Fiber-Wireless
FPGA	Field-Programmable Gate Array
FR	Frequency Range
F-RAN	Flexible RAN
FRMCS	Future Railway Mobile Communication System
FSaaS	Functional Split as a Service
FSPL	Free-Space Path Loss
FTTH	Fiber-to-the-Home
FTTx	Fibre-to-the-Everything
FWA	Fixed Wireless Access
GaAs	Gallium Arsenide Semiconductor
GaN	Gallium Nitride
Gbps	Gigabits/Gigabytes Per Second
GEO	Geostationary Orbit
GEPON	Gigabit Ethernet PON
GFDM	Generalized Frequency Division Multiplexing
GHz	Gigahertz
gNB	5G base station

GNSS	Global Navigation Satellite System
GPON	Gigabit Passive Optical Access Network
GPP	General Purpose Processing
GPS	Global Positioning System
GRNN	General Regression Neural Networks
GRU	Gated Recurrent Unit
GSG	Ground-Signal-Ground
GTP-U	GRE Tunnelling Protocol for User Plane
HAL	Hardware Abstraction Layer
HAPS	High-Altitude Platform Stations
HARQ	Hybrid Automatic Repeat Request
HBT	Heterojunction Bipolar Transistor
HCC	Harmonic Control Circuit
H-CRAN	Hybrid-Cloud Radio Access Network
HCS	Harmonized Communication and Sensing
HD	High Definition
HEMT	High-Electron-Mobility Transistor
HF	High Frequency
HO	Handover
HOF	Handover Failures
HPPP	Homogeneous Poisson Point Process
HR	High Reflection
HS	Hot Spot
HSP	High-Speed PON
HTC	Human Type Communications
HWLN	Hybrid WiFi-LiFi Network
i.i.d.	Independent and Identically Distributed
IAB	Integrated Access and Backhaul
IC	Integrated Circuit
ICI	Inter-Carrier Interference
IDLP	Intrusion Detection and Location-Aware Prevention
ID-PKC	Identity-based PKC
IDPS	Intrusion Detection and Prevention Scheme
IEEE	The Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IF	Intermediate Frequency
IFFT	Inverse Fast Fourier Transform
IFoF	Intermediate Frequency over Fiber
IIC	Iterative Interference Cancellation
IIoT	Industrial Internet-of-Things
ILP	Instruction Level Parallelism
IM	Intensity Modulation
IMD	Inter-Modulation Distortion
IMDD	Intensity Modulation Direct Detection
IMN	Input Matching Network

InP	Infrastructure Provider
IoE	Internet of Everything
IoT	Internet of Things
IP	Internet Protocol
IP3	Third-Order Intercept Point
ISI	Inter-Symbol Interference
ISMN	Inter-Stage Matching Networks
ITS	Intelligent Transportation System
ITU	International Telecommunications Union
ITU-R	International Telecommunications Union-Radio Communication
JT	Joint Transmission
JT-CoMP	Joint Transmission Coordinated Multi-Point
KDC	Key Distribution Center
KPI	Key Performance Indicator
Lagg	Low Aggregation
LD	Last Day
LDPC	Low-Density Parity Check
LED	Light Emitting Diode
LEO	Low Earth Orbit
LHD	Lidar and HD
LiFi	Light Fidelity
LINC	Linear Amplification with Non-linear Components
LLR	Log-Likelihood Ratio
LMS	Least Mean Square
LNA	Low Noise Amplifier
LO	Local Oscillator
LoS	Line-of-Sight
LPF	Low-Pass Filter
LS	Least Square
LSTM	Long Short-Term Memory
LTE	Long-Term Evolution
LTSA	Linear Tapered Slot Antenna
LV	Last Value
M2M	Machine-to-Machine
MAC	Medium Access Control
MaMIMO	Massive Multiple-Input Multiple-Output
MANET	Mobile Ad hoc Network
MANO	Management and Orchestration
MBS	Main Base Station
MC	Mobile Cloud
MCC	Mobile Cloud Computing
MCF	Multi-Core Fibre
MCL	Maximum Coupling Loss
MCS	Modulation and Coding Scheme
MD	Map Data

MDM	Mode Division Multiplexing
MDN	Mixture Density Network
MDP	Markov Decision Process
MDUA	Minimum-Distance User Association
MEC	Multi(-access) Edge Computing
MESFET	Metal Semiconductor Field Effect Transistor
MH	Mobile Helper
MHN	Mobile Heterogeneous Network
MI	Mutual Information
MIC	Microwave Integrated Circuit
MIM	Metal Insulator Metal
MIMO	Multi Input Multi Output
MINLP	Mixed-Integer Nonlinear Programming
MIoT	Massive IoT
MISO	Multi-Input Single-Output
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MLB	Mobility Load Balancing
MLD	Maximum-Likelihood Detection
MLPR	Multi-Layer Perceptron Regressor
MLR	Multiple Linear Regression
MMC	Mobility Markov Chains
MMF	Multi-Mode Fibers
MMIC	Monolithic Microwave Integrated Circuit
mMIMO	Massive MIMIO
MMSE	Minimum-Mean-Squared-Error
mMTC	Massive Machine-Type Communications
mmWave	Millimeter wave
MNO	Mobile Network Operator
MON	Monitors
MPA	Medium Power Amplifier
MQW	Multiple Quantum Well
MRO	Mobility Robustness Optimization
MSC	Mobile Small Cell
MSE	Mean Square Error
M-SMF	Multiple Single Mode Fibre
MTC	Machine Type Communications
Multi-CAP	Multiband Carrierless Amplitude Phase
MUX	Multiplexer
MVNO	Mobile Virtual Networks Operators
MZM	Mach Zendher Modulator
NAF	Normalized Advantage Functions
NBI	Northbound Interface
NB-IoT	Narrow Band IoT
NC	Network Coding

NCC	Network Coded Cooperation
NC-MSC	Network Coding-Enabled Mobile Small Cell
NE	Nash Equilibrium
NF	Network Function
NFV	Network Function Virtualization
NFVI	NFV Infrastructure
NFVO	NFV Orchestrator
NGFI	Next Generation Fronthaul Interface
NGMN	Next Generation Mobile Networks
NG-PON2	Next Generation PON2
NLoS	Non Line-of-Sight
NN	Neural Network
NOMA	Non-Orthogonal Multiple Access
NPRACH	NB-Physical Random-Access Channel
NPUSCH	Narrowband Physical UplinkShared CHannel
NR	New Radio
NR-U	New Radio-Unlicensed
NRZ	Non-Return-to-Zero
NSA	Non-Stand-Alone
NSI	Network Slice Instance
NSMF	Network Slice Management Function
NSSMF	Network Slice Subnet Management Function
NTE	Network Terminating Equipment
NTU	Non-Transferable Utility
OAI	Open Air Interface
OBPF	Optical Bandpass Filter
OBSAI	Open Base Station Architecture Initiative
OCU	Optimal Content Update
OF	OpenFlow
OFC	OpenFlow controller
OFDM	Orthogonal Frequency-Division Multiplexing
OFH	Optical Fronthaul
OH FEC	Overhead Forward Error Correction
OLT	Optical Line Terminal
OMN	Output Matching Network
OMP	Orthogonal Matching Pursuit
ONF	Open Network Foundation
ONU	Optical Network Unit
OOB	Out-of-band
OOK	On Off Keying
OPEX	Operating Expenditure
OQAM	Offset Quadrature Amplitude Modulation
OTF	Tunable Optical Fibre
OTO	One-to-one
OXC	Optical Cross Connect

P	Protected
P2P	Peer-to-Peer
PA	Power Amplifier
PAA	Phased Array Antenna
PAE	Power Added Efficiency
PAM4	Four-level Pulse Amplitude Modulation
PAPR	Peak-to-Average Power Ratio
PARAFAC	PARAllel FACtor
PBS	Polarization Beam Splitter
PC	Polarization Controller
PCB	Printed Circuit Board
PC-ZF	Partially Centralised Zero-Forcing
PD	Photodiode
PDCP	Packet Data Convergence Protocol
pdf	Probability Density Function
PDN	Packet Data Network
PD-TTP	Partially Distributed TTP
PDU	Protocol Data Unit
PGW	PDN Gateway
pHEMT	Pseudomorphic High-Electron-Mobility Transistor
PHY	Physical Layer
PIC	Photonic Integrated Circuit
PIFA	Planar Inverted-F Antenna
PKC	Public Key Cryptography
PKI	Public Key Infrastructure
PoI	Points of Interest
PON	Passive Optical Network
PoS	Proof of Stake
PoW	Proof of Work
PP	Point Process
PPB	Photons per Bit
PPDI	Polarization and Phase Diverse Intradyne
PPM	Pulse Position Modulation
PPN	Poly-Phase Network
PRB	Physical Resource Block
PRBS	Pseudo Random Binary Sequence
PRS	Preliminary Resource Sharing
PS	Polarization Scrambling
PS QPSK	Polarization Switch QPSK
PSA	PDU Session Anchor
PSD	Power Spectral Density
PSK	Phase Shift Keying
PSS	Proactive Secret Sharing
PtP	Point-to-Point
QAM	Quadrature Amplitude Modulation

QCSE	Quantum Confined Start Effect
QKD	Quantum Key Distribution
QoE	Quality of Experience
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RA	Resource Allocation
RAN	Radio Access Network
RANaaS	RAN as a Service
RAR	Re-Auth-Request
RAT	Radio Access Technology
RAU	Radio Aggregation Unit
RB	Resource Block
RCC	Radio Cloud Center
RD	Resource Duplication
REAM-SOAs	Reflective EAM-SOAs
RF	Radio Frequency
RFID	Radio Frequency Identification
RGBM	RRH Group-Based Mapping
RIRS	Reconfiguration and Improved Resource
RIS	Reconfigurable Intelligent Surface
RL	Reinforcement Learning
RLC	Radio Link Control
RLNC	Random Linear Network Coding
RMSE	Root Mean Square Error
RN	Relay Node
RNC	Radio Network Controller
RNN	Recurrent NN
RoF	Radio-over-Fibre
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RRU	Remote Radio Unit
RS	Reference Signal
RSRP	Reference Signal Received Power
RSS	Resource Scheduling Strategy
RT	Route
RTBC	Real-Time Broadband Communication
RTD	Real-time Traffic Data
RU	Radio Unit
RUS	Reflecting Units
Rx	Receiver
RZF	Regularized ZF
SA	Signal Analyzer
SBA	Service-Based Architecture
SBI	Southbound Interface

SBS	Subbase Station
SBS	Small Base Station
SC	Single Carrier
SC-FDM	Single Carrier Frequency Division Multiplexing
SD	Software Defined
SDA	Software-Defined Access
SDM	Spatial Division Multiplexing
SDMN	Software-Defined Mobile Networks
SDN	Software-Defined Network(ing)
SDR	Software-Defined Radio
SD-RAN	Software-Defined RAN
SD-UCP	Software-Defined Unified Control Plane
SE	Spectral Efficiency
SECRET	SEcure network Coding for Reduced Energy nexT generation mobile small cells
SF	Sequential Fixing
SFB	Synthesis Filter Bank
SFC	Service Function Chaining
SFS	SF and Scheduling
SGW	Serving Gateway
Si	Silicon
SI-BH	Semi-Insulating Buried Heterostructure
SiC	Silicon Carbide
SIC	Successive Interference Cancellation
SiGe	Silicon-Germanium
SINR	Signal-to-Interference-Plus-Noise Ratio
SIW	Substrate Integrated Waveguide
SL	Service Layer
SLA	Service Level Agreement
SLO	Service Level Objective
SMF	Session Management Function
SMF	Single Mode Fiber
SN	Source Node
SNR	Signal-to-Noise Ratio
SO	Self-Organization
SOA	Semiconductor Optical Amplifier
SoA/SoTA	State-of-the-Art
SON	Self-Organizing Network
SP	Service Provider
SPP	Set the Partitioning Problem
SPSCP	Shared-Path Shared-Compute Planning
SRE	Smart Radio Environment
SRS	Sounding Reference Signal
SSC	Spot-Size Converter
SSMF	Standard Single-Mode Fiber

STBC	Space Time Block Codes
SU	Storage Unit
TaLP	Task Level Parallelism
TB	Transport Block
Tbps	Terabits per Second
TCO	Total Cost of Ownership
TCP	Transmission Control Protocol
TDM	Time Division Multiplexer
TDMA	Time-Division Multiple Access
TE	Transverse Electric
ThLP	Thread Level Parallelism
THz	TeraHertz
TIA	Transimpedance Amplifier
TL	Tunable Laser
TML	Transmission Line
TNC	Transport Network Controller
TNL	Transport Network Layer
TP	Transmission Point
TP-Selection	Transmission-Point Selection
TSP	Telecom Service Providers
TTI	Transmission Time Interval
TT-ID-PKC	Threshold Tolerant-ID-PKC
TTP	Trusted Third Party
TTT	Time-to-Trigger
TU	Transferable Utility
TWDM	Time and Wavelength-Division Multiplex(ing)
TWDM-PON	Time Wavelength Division Multiplexing PON
Tx	Transmitter
UAM	Urban Air Mobility
UAV	Unmanned Aerial Vehicles
UCBC	Uplink Centric Broadband Communication
UDN	Ultra-Dense Network
UDWDM	Ultra-Dense Wavelength Division Multiplexing
UE	User Equipment
UFMC	Universal-Filtered Multi-Carrier
UL	Uplink
UP	Unprotected
UP CL	Uplink Classifier
UPF	User Plane Function
uRLLC	Ultra-Reliable and Low Latency Communications
USRP	Universal Peripheral Radio Software
UT	Utility Function
UTC	Uni-Travelling Carrier
UTC-PD	UTC Photodiode
V2I	Vehicle-to-Infrastructure

V2X	Vehicle-to-Everything
vBBU	Virtual BBU
VIM	Virtualized Infrastructure Manager
VLC	Visible Light Communication
VND	Variable Node Decoder
VNF	Virtual Network Function
VOA	Variable Optical Attenuator
vOLT	Virtual OLT
VR	Virtual Reality
vRAN	Virtualized RAN
vSDN	Virtual SDN
VSG	Vector Signal Generator
VSS	Verifiable Secret Sharing
WDM	Wavelength Division Multiplexer
WDM-PON	Wavelength Division Multiplexed Passive Optical Network
WG	Working Group
WiMAX	Worldwide Interoperability for Microwave Access
WNV	Wireless Network Virtualization
XaaS	Anything as a Service
xDSL	Digital Subscriber Line
XR	Extended Reality
ZF	Zero-Forcing
ZFBF	Zero-Forcing Beamforming
ZSM	Zero-touch Network and Service Management
δ	Loss Tangent
ε_r	Permittivity

Part I
5G and Beyond Mobile Landscape

Chapter 1
Drive Towards 6G

Firooz B. Saghezchi, Jonathan Rodriguez, Zoran Vujicic, Alberto Nascimento, Kazi Mohammed Saidul Huq, and Felipe Gil-Castiñeira

Abstract As fifth-generation (5G) mobile networks are being rolled out, the telecom industry and academia are now coordinating the 6G research effort towards defining the requirements and use cases for beyond 5G (B5G) or so-called sixth-generation (6G) mobile networks. 6G envisages an evolutionary communication platform based on complete network softwarisation, inclusive communications mediums including satellite, and ultra-dense networks to cater for the market demands that requires ultra-high speeds, tactile response time, and lower cost of network ownership by 2030. This chapter provides an overview of the use cases for B5G/6G systems, including holographic telepresence, digital twin, connected robotics, distributed artificial intelligence, and blockchain technologies. It further reviews the current standardisation and deployment status of 5G technology as a baseline and the drive towards 6G by identifying key enabling technologies, system requirements, and an overview on global B5G/6G activities.

F. B. Saghezchi (✉) · Z. Vujicic
Instituto de Telecomunicações, Campus Universidade de Aveiro, Aveiro, Portugal
e-mail: firooz@av.it.pt

J. Rodriguez
Instituto de Telecomunicações, Campus Universitário Santiago, Aveiro, Portugal

Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, UK

A. Nascimento
Departamento de Matemática e Engenharias, Universidade da Madeira, Funchal, Portugal
e-mail: ajn@uma.pt

K. M. S. Huq
Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, UK
e-mail: kazi.huq@southwales.ac.uk

F. Gil-Castiñeira
atlanTTic Research Center for Telecommunication Technologies, Universidade de Vigo,
Information Technologies Group, Vigo, Spain
e-mail: xil@gti.uvigo.es

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_1

1.1 Introduction

The specification of fifth-generation (5G) mobile networks has already passed the standardisation phase, and 5G networks are being rolled out around the world. In November 2020, the International Telecommunication Union (ITU) approved that the 5G New Radio (NR) and 5G Core technologies developed by the 3rd Generation Partnership Project (3GPP) conform with the IMT-2020 performance requirements. Having achieved this important milestone, the stakeholders and academic researchers are now shifting their attention towards beyond 5G (B5G) paradigm.

The first rollout of 5G (2019–2020) is targeting sub-6 GHz small cells based on a cloud radio access network (C-RAN) architecture. However, subsequent rollouts will deploy the full 5G vision addressing the hyperdense deployment of small cells based on millimetre-wave frequencies, where larger swathes of spectrum are available. This will be coupled with multiple antenna technology deployed on a massive scale that, in synergy, will have a multiplier effect on user peak data rates and cell capacity, enabling the 5G system to fulfil the ambitious performance indicators specified by 3GPP.

However, going beyond 5G systems (what is referred to as the B5G paradigm), this aims to push back further the boundaries on communication systems in a bid to introduce tactile Internet applications that combines ultra-low latency with extremely high availability, reliability, and security. Speed is also a key design requirement in 6G systems. Whereas in 4G we spoke about megabit terms, 5G pushed this to the gigabit barrier, and 6G is now expected to deliver theoretical terabit speeds.

To entertain, B5G systems will undoubtedly rely on virtualisation and software-defined network that will provide operators with a fully dynamic and efficient networking architecture that is able to match the networking resources to the actual load in the network, whereas new enabling technologies based on terahertz (THz) communications will unlock large swathes of available spectrum and potential speed. In synergy, these technologies will provide a communication platform for hosting a plethora of new application such as digital replicas (twins), distributed artificial intelligence (AI), augmented reality (AR), and autonomous vehicles, among others.

The heart of B5G will be reliant on network softwarisation that was already a main feature in 5G systems. However, this technology will further evolve to encompass a complete overhaul of the underlying infrastructure where once the edge network was confined to the operator's infrastructure, it will now include fixed and mobile devices and may include infrastructure that may not be under the complete ownership of the operator.

Indeed, the edge network will migrate to the users' local area space, to include devices in the very near vicinity. This will take a large step towards reducing the latency in the network by enabling the local caching of popular data, as well as content migration. Moreover, computing tasks such as local task offloading will be possible enabling further virtualisation of user handset and improving battery lifetime.

Optical infrastructures will provide the necessary backhaul capacity to meet the expected demand in data traffic that will also be softwarised to provide operators the muscle to adapt the optical infrastructure to harness the dynamic load in the network in an efficient manner.

Despite the ongoing rollout of 5G, a key bottleneck to ultra-high speed is the underlying spectrum. However, driven the market demand towards higher bit rates to entertain virtual reality (VR) applications requires higher frequencies over the THz band (0.1–10 THz) that will be key to ubiquitous 6G networks. In particular, THz frequencies have the potential to deliver ample spectrum, over hundred gigabits-per-second (Gbps) data rates, massive connectivity, denser networks, and highly secure transmissions.

Therefore, the merger of the optical, THz wireless, and virtualisation technology domains will result in a fully flexible and efficient high-speed communication platform to what the research community is envisaging as B5G communications, to enable the mass deployment of small cells, or what is referred to as the ultra-dense networks (UDNs). This will enable not only enhanced broadband connectivity but a communication medium to deliver future emerging very-low-latency tactile application. It is worthy to note that when we refer to B5G technologies in this book, these can eventually be part of the 6G standard or part of subsequent 5G releases.

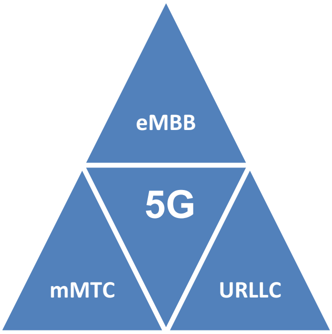
This chapter addresses abovementioned technologies, use cases, and challenges for B5G/6G systems. Starting with an overview of 5G systems, standardisation, and deployment status, this provides the launch pad for the current developments on 6G. We introduce the B5G/6G use cases along with a definition of the network architecture that will motivate the radio protocols/algorithm challenges addressed in the subsequent chapters. Last but not the least, this chapter concludes with a review on the current 6G research activities and initiatives around the world, as we head towards the 6G era.

The rest of this chapter is structured as follows: Sect. 1.2 reviews the current status of 5G systems including the standardisation and network deployment activities as baseline; Sect. 1.3 identifies the drivers to advance mobile networks towards B5G/6G in terms of potential use cases; Sect. 1.4 discusses system requirements for 6G networks; Sect. 1.5 presents key enabling technologies that advance 6G and telecom market in the coming years; and Sect. 1.6 reviews the ongoing research and development activities around the globe towards 6G. Finally, Sect. 1.7 concludes this chapter.

1.2 5G Systems Overview

Three main use cases have been identified for 5G towards a fully connected society and accelerating the digital transformation in different verticals, namely, enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and

Fig. 1.1 Key use cases driving 5G



ultra-reliable low-latency communications (URLLC) [1–3]. Meeting the communication requirements of these use cases has mostly been driving the main innovations in the mobile industry over the past decade or so, as portrayed by Fig. 1.1.

The first use case (eMBB) targets enhancing mobile networks in terms of speed, capacity, and coverage to provide broadband mobile connectivity to cater for ever-increasing mobile data traffic generated by mobile smartphones (e.g. high-quality images and videos) and offer emerging applications such as AR/VR. The target is to offer up to 10–20 gigabits-per-second (Gbps) mobile data speeds providing connectivity in densely crowded areas such as sport events, concerts, and so forth.

The second 5G use case (mMTC) aims to address challenges such as connecting tens of billions of machines or resource-constrained Internet of Things (IoT) nodes (e.g. sensors, actuators, etc.) to the network as new clients. This will bring new applications such as connected homes, connected cars, etc. According to the Cisco Annual Internet Report (2018–2023), there will be 29.3 billion networked devices by 2023, on average 3.6 networked devices per capita; half of this will be machine-to-machine (M2M) connections. This would require an unprecedented shift in the industry from serving traditional human clients in fourth-generation (4G) systems towards an infrastructure that is well prepared to support M2M communications. The network would indeed need a huge capacity addition to accommodate these devices, as well as a new level of optimisations to efficiently use scarce radio resources towards transporting the heterogeneous types of traffic profiles generated by these nodes. It would also need to take into account the energy constraints of these devices which are often battery powered, in the development of their communication protocols.

The third use case (URLLC) targets network innovations to support time-critical applications such as autonomous driving, vehicle-to-everything (V2X) communications (i.e. vehicle-to-vehicle (V2V) or vehicle-to-infrastructure (V2I)), unmanned aerial vehicles (UAV), intelligent transportation system (ITS) [4], and critical infrastructures (e.g. smart grid [5], smart manufacturing [6], precision farming, railways, etc.). This use case targets shrinking the radio access network latency from 20 ms in 4G down to 5 ms and eventually to below 1 ms. The use case would potentially bring new real-time control applications requiring the communication of

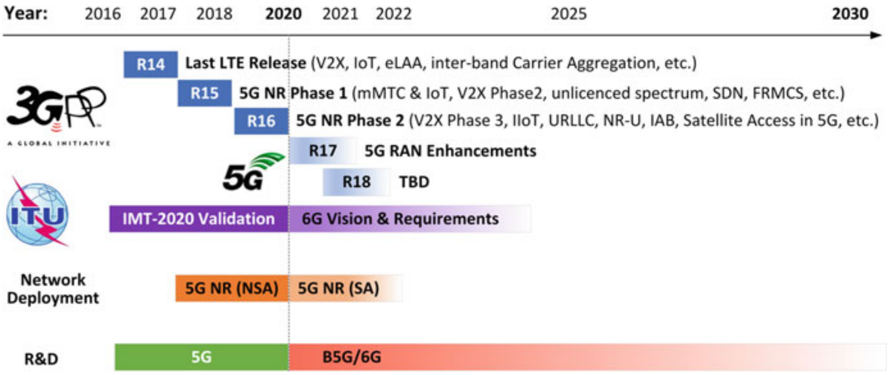

Year:
2016
2017
2018
2020
2021
2022
2025
2030
R14
Last LTE Release (V2X, IoT, eLAA, inter-band Carrier Aggregation, etc.)
R15
5G NR Phase 1 (mMTC & IoT, V2X Phase2, unlicenced spectrum, SDN, FRMCS, etc.)
R16
5G NR Phase 2 (V2X Phase 3, IIoT, URLLC, NR-U, IAB, Satellite Access in 5G, etc.)
R17
5G RAN Enhancements
R18
TBD
3GPP
A GLOBAL INITIATIVE
5G
ITU
IMT-2020 Validation
6G Vision & Requirements
Network Deployment
5G NR (NSA)
5G NR (SA)
R&D
5G
B5G/6G

Fig. 1.2 Timeline for 5G and beyond 5G (B5G) network development

human touch sense (e.g. for remote driving, among others), leading to the era of so-called tactile Internet [7]. A major advancement towards meeting the requirements of this use case is the improvement in mobile edge computing that brings cloud computing benefits to the radio access network, therefore releasing a tremendous amount of capacity in the backhaul links while considerably reducing network latency.

Figure 1.2 illustrates the timeline for 5G and B5G network development. In particular, the 3GPP, as a leading standardisation body, specifies radio access networks through different releases. Release 14 (R14) was the last release of the 4G Long-Term Evolution (LTE) standard, introducing further advancements such as V2X, narrowband IoT (NB-IoT), enhanced Licensed-Assisted Access (eLAA), etc. Any release after R14 is considered as a 5G release.

R15 was the first phase of 5G standards where 5G NR was introduced for the first time. It contained 5G specifications for both non-standalone (NSA) and standalone (SA) architectures. In the former, 5G radio access and its NR interface are connected through the existing 4G infrastructure, Evolved Packet Core (EPC), therefore making NR technology available without replacing the existing infrastructure. In this configuration, only the 4G services are supported, but enjoying the capacities offered by the 5G New Radio (lower latency, etc.). In contrast, in the SA architecture, the whole network is 5G, including the radio access, transport, and core networks.

R16 covered the second phase of 5G specifications introducing further advancements to 5G NR, e.g. V2X communications, industrial IoT (IIoT), URLLC, unlicensed access, satellite access, and so on. R16 is the first release meeting IMT-2020 requirements for 5G systems that is defined by the ITU; thus, the 3GPP releases succeeding R16 are considered as B5G specifications. The ITU is now speculating the requirements for 6G to address the communication needs of our connected society in the year 2030.

1.3 Drivers Towards Beyond 5G

Having achieved IMT-2020 requirements for 5G, the research community have already started envisioning B5G/6G era. Figure 1.3 illustrates five promising use cases for B5G/6G on which a general consensus is beginning to emerge.

The first use case, *holographic telepresence*, permits real-time communications of realistic, full motion, three-dimensional (3D) images of distant people and objects for projection in a room (e.g. meeting room, classroom, surgery room, etc.) accompanied by real-time audio communications [8]. This can, e.g. reduce the necessity of travel for attending business meetings and facilitate applications such as remote surgery or distant learning.

The second key use case, *digital twin*, allows to make a real-time, complete, and executable digital back-up of an asset, system, or subsystem (e.g. Industry 4.0, smart grid, etc.) [9]. The virtual model can be exploited, among others, for simulation, diagnostics, fault prediction, and overhaul. Once the updates passed all the tests over the real-time virtual model, it can replace the existing physical model without any downtime, thus pushing the boundaries of reliability in these systems.

The third use case, *connected robotics and autonomous vehicles*, permits the mobile infrastructure to be used to connect different components of a control system (i.e. physical process, controller, sensors, and actuators) distributed across a wide geographic region [10]. The application requires stringent delay and jitter performance. An emerging application of this use case is tactile Internet, which requires to communicate the human touch sense for remote control purposes. In such applications, the maximum tolerable network latency can be less than 1 ms.

The fourth use case, *IoT, distributed AI, and big data*, aims to empower mobile networks with AI, machine learning (ML), and big data technologies to automate network operation. For instance, big data and deep learning techniques can be used to predict network traffic and allocate radio resources (spectrum, time, power, spatial beam, etc.) to different cells or groups of cells or to different (virtual) network slices. This can remarkably improve the resource utilisation efficiency in mobile networks.

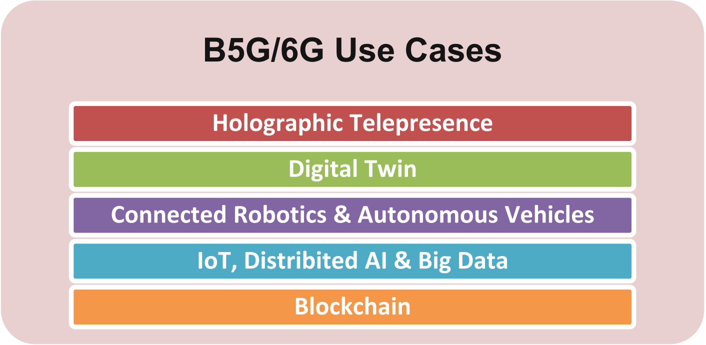

B5G/6G Use Cases
Holographic Telepresence
Digital Twin
Connected Robotics & Autonomous Vehicles
IoT, Distribited AI & Big Data
Blockchain

Fig. 1.3 Driving use cases for B5G/6G

Moreover, distributed AI and federated learning algorithms can be performed on Multi-Access Edge Computing (MEC) servers at the access network of 5G for third parties to run time-critical IoT applications such as ITS, smart manufacturing, smart agriculture, etc. The use of distributed AI over edge computing infrastructure not only can save considerable amount of bandwidth on backhaul links, since there is no need to communicate tremendous amounts of raw data to backend servers, but can considerably reduce the latency as the data is processed at edge servers in close proximity of smart sensors where the data is generated.

Last, but not least, *blockchain* is a promising technology for B5G/6G to store data in a decentralised, persistent, anonymous, and auditable manner [11]. It has a great potential to address security and privacy concerns caused primarily due to the collection of large volumes of sensitive data, including user's personal data or industrial data (containing intellectual property), for running distributed AI and ML algorithms.

1.4 6G Requirements

In this section, we present envisioned requirements for 6G, including both performance and architectural requirements – i.e. fog computing to lower latency, connection to satellite, and the exploitation of distributed AI across all layers.

1.4.1 6G Performance Requirements

In terms of network performance requirements, 6G is expected to achieve the following advancements; see also the radar chart illustration in Fig. 1.4 [12].

- Increasing the peak mobile data rate to 1 terabits per second (Tbps), 50 times enhancement over 5G (20 Gbps)
- Increasing the experienced rate for highly mobile users to 1Gbps, 100 times increase over 5G (100 Mbps)
- Increasing the connection density to 10^7 devices/Km^2, 10 times increase over 5G (10^6 devices/Km^2)
- Decreasing the air latency down to 0.1 ms, 10 times improvement over 5G (1 ms)
- Boosting both the spectral and energy efficiency of the air interface at least twice in comparison to 5G
- Improving the network reliability to 10^{-7}, 100 times improvement over 5G (10^{-5})

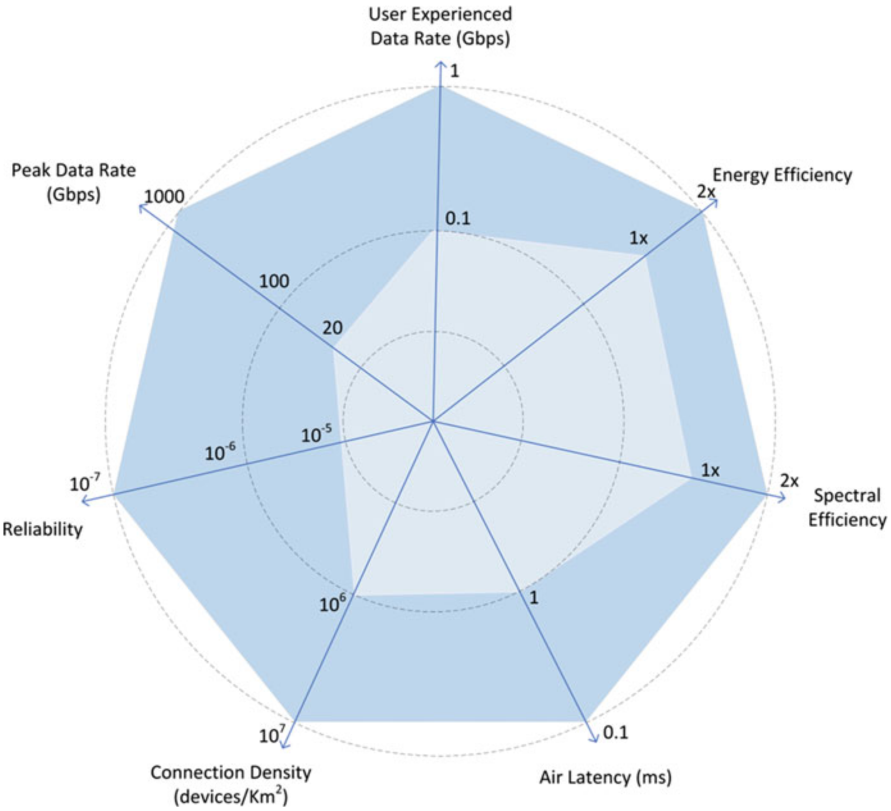


Fig. 1.4 Performance requirements for 6G

1.4.2 6G Network Architecture Requirements

In terms of network architecture, 6G requires a decentralised federated architecture relying on fog/edge computing architecture. The data processing is performed in a distributed fashion on nearby edge/fog nodes without needing the (bandwidth-consuming) data to travel all the way up to the cloud servers. This not only can help reduce the air latency, because of data being processed at nearby edge/fog servers, but also can considerably alleviate the strain on backhaul links, pushing the network capacity towards a new limit. Furthermore, it improves network reliability and opens the network platform for scalable processing to tackle the big data challenges in 6G.

It is expected that distributed AI and data processing techniques will play pivotal role in most 6G use cases. Therefore, transparent and reliable execution of distributed AI/ML algorithms at MEC servers is of crucial importance for 6G network architecture. The algorithms must be trustworthy complying to, for example, the ethical rules published by the EC's High-Level Expert Group on AI.

According to these guidelines, trustworthy AI should be (1) *lawful*, respecting all applicable laws and regulations; (2) *ethical*, respecting ethical principles and values; and (3) *robust*, both from a technical perspective while taking into account its social environment [13].

Effective computation load balancing is needed to hand over mobile users in fog computing environment from one MEC node to another. This is especially challenging when the user is frequently crossing the boundary of small cells in ultra-dense heterogeneous networks operating in THz or millimetre-wave (mm-wave) bands. Additionally, secure data management techniques are needed to authenticate edge/fog nodes, preserve the user's data privacy, and ensure the integrity of, e.g. the control commands for the use cases that are involved with remote control (e.g. cyber-physical systems, networked robotics, etc.).

Last but not the least, efficient satellite connectivity is another important aspect of 6G network architecture. Various use cases (e.g. autonomous cars) require continuous reception of the geolocation information from the satellite using, e.g. Global Positioning System (GPS), Global Navigation Satellite System (GNSS), etc. Apart from that, sea travellers or users located at remote locations (e.g. machines or IoT nodes deployed in the field) where there is no mobile infrastructure can rely on satellite communications to relay their data. In this regard, new network management architecture is needed to jointly operate terrestrial and satellite networks, e.g. to hand over the user from one to another, or share the scarce radio spectrum between the two networks.

1.5 Enabling Technologies for B5G/6G

Figure 1.5 illustrates key envisaged technologies to advance mobile industry towards B5G/6G, namely, quantum and THz communications, visible light communication (VLC), virtualised RAN, distributed AI and big data, energy harvesting and fog computing, as well as enhanced optical-wireless convergence. Here, we briefly highlight their significance for brevity, where we recommend the interested reader to the respective subsections for more detailed insights.

Ultra-dense Small Cell Networks: network hyperdensification is a fundamental approach to increase network capacity by increasing bit/sec/Hz/Km^2 spectral efficiency [14]. The recent mm-wave, THz, and VLC technologies have a considerably short radio coverage, due to their line-of-sight propagation characteristics, but they can provide tens of Gbps wireless speeds since they have access to huge amounts of intact spectrum available above 30 GHz. As the demand for additional network capacity and mobile speed rises, future networks will become even denser. Several challenges need to be solved, e.g. handover between different small cell base stations (BSs), inter-cell interference management, and self-organisation of small cell BSs since it will be intractable to manually configure this vast number of BSs.

Virtualised Network: network virtualisation enables setting up different logical network slices over the same physical infrastructure, thus allowing multitenant

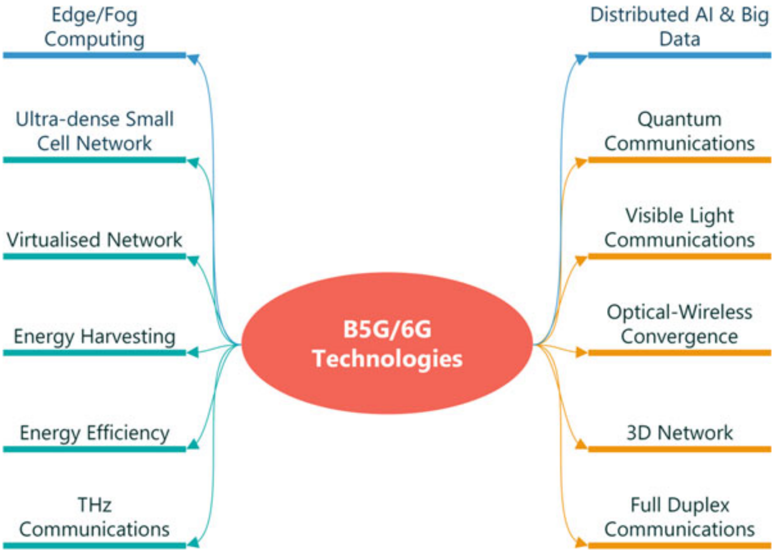


Fig. 1.5 Enabling technologies for B5G/6G

network operation and opening up the network to vertical industries, with different capacity and quality-of-service (QoS) requirements. The slices can be set up in demand expanded or shrunk dynamically according to their capacity need [15]. Nevertheless, the management and orchestration of network slices and allocating them resources (on the fly) require scalable and tractable solutions. Furthermore, the open interfaces that allow the programmability of the network introduce new attack vectors. Therefore, a consistent multi-level security framework is needed for ensuring software integrity, remote attestation, and dynamic detection and mitigation of threats [16]. Further discussion on this technology will follow later on in Sect. 1.5.4.

AI and Big Data: as the complexity of mobile networks grows rapidly, it becomes intractable to manage them using traditional explicit programming techniques. Therefore, new level of automation or self-configuration is needed so as the network can learn from explorative actions or from training examples on how to adapt to the varying traffic volume or channel quality, among others. In this context, AI and ML techniques will play a more prominent role in 6G. Furthermore, as the data is generated everywhere by a myriad of small cell BSs, user equipment (UE), or IoT sensors, scalable big data technology is highly relevant to rip the benefit of mining this data in real time. Further elaboration on this topic is provided in Sect. 1.5.5.

Energy Efficiency/Harvesting: another important aspect for sustainable development of mobile industry is to curb its carbon footprint, which is currently comparable to the greenhouse gas emission of the airline industry [17, 18]. Smart antenna arrays and beam forming techniques can be used to pinpoint sharp radiation beams towards the receiver and avoid power waste (and interference) by radiation in unnecessary directions. This requires practical solutions for adaptive beam alignment as any small movement of mobile users can result in misalignment and hence link failure. Furthermore, on the UE and IoT sensors' side (for the vertical use cases), energy-efficient communications techniques are crucial to cope with the battery limitation of these nodes [19, 20]. The IoT nodes may also rely on *harvesting* energy from electromagnetic radiations to operate either battery-less or over a considerably extended battery lifetime. More discussions on energy harvesting will follow later on in Sect. 1.5.6.

THz and VLCs: mm-wave communications is a 5G technology unlashing tens of GHz of spectrum in mm-wave band (30 GHz to 300 GHz) for mobile communications [21]. However, as this technology is getting mature, the industry is now looking for new breakthroughs in THz band (300 GHz to 10 THz) [22], which will enable Tbps wireless communications for mobile users. In quest for new spectrum, the VLC communication opens up even higher amounts of spectrum for wireless communications, compared to THz communications; it operates in much higher-frequency band of 430 THz to 790 THz where 360 THz of spectrum can be exploited for wireless communications. The VLC technology modulates the light intensity of light-emitting diodes (LEDs) and can use the already existing lighting infrastructure (e.g. indoors, street lampposts, vehicle headlights, etc.) for wireless data transmission while they are also serving for the illumination purposes at the same time. The technology is promising as it has already demonstrated 10 Gbps wireless speed. The network deployment is much cheaper than the radio-frequency (RF) technology. However, it faces some challenges such as the reliability of the connectivity and its offered QoS since any light blockage can result in the failure of the communication link. Moreover, seamless handover from one VLC cell to another or from the VLC technology to the mobile network needs further investigation. For further discussion on these topics, please refer to Sects. 1.5.2 and 1.5.3.

Quantum Communications: quantum communications is another disruptive technology that will revolutionise the Internet. It takes advantage of quantum physics to transmit data. The technology uses quantum bits (qubits) to securely transmit data over optical fibres. The qubit has several interesting properties that are fundamentally different from classical information *bits*. Unlike information bits that can be in either 0 or 1 state at a given time, qubit can be in a *superposition* of 0 and 1, that is, a combination of 0 and 1. Furthermore, if one tries to measure the state of this superposed qubit, he/she will end up a probabilistic outcome of 0 or 1 depending on the "amount" of 0 and 1 in the qubit. However, after this *quantum measurement*, the original qubit state will collapse into the measured state. For instance, if the outcome of measuring a superposed qubit corresponds to the state 0, the qubit will collapse into the state 0, and any further measurement will give 0 as the result, independent of the original amount of 1 in the superposed state. Quantum communications exploits the *quantum entanglement* property to transfer data between two distant nodes. This property allows two quantum particles (photons) to be entangled, meaning that their quantum state is completely correlated. Therefore, by transmitting one of these particles, we can measure the quantum state of its source peer at the destination node. Another distinctive property of qubit is *no-cloning*. That is, unlike classical communication system where making a copy of the information bit is possible, in quantum communications, it is not possible to make such a copy of the qubit. This has profound consequences on the *quantum Internet* design as it, for example, prevents the quantum information from being retransmitted or transmitted to more than one destination (multicasting or broadcasting). This is a radical difference with respect to classical networks, where packet retransmission or broadcasting is widely exploited for implementing several layer-2, layer-3, and layer-4 functionalities, such as medium access control, automatic repeat request, route discovery, and packet retransmission employed by the Transmission Control Protocol (TCP). As a consequence, the link layer must be carefully re-thought and re-designed [23]. The abovementioned properties however make quantum communications appealing for establishing secure communication channels through quantum key distribution (QKD); any attempt to read the data on transit will distort the state of qubits and thus will become detectable for the receiver [24, 25]. For further discussion on this topic, we refer the interested reader to Sect. 1.5.1.

3D Networks: traditionally, terrestrial networks rely on BSs that are deployed in two dimensions (latitude and longitude) on land surfaces. However, as the incorporation of UAVs is becoming prevalent in different industries (e.g. smart agriculture, wildfire monitoring, inventory monitoring, etc.), mobile flying small cell BSs can be deployed on demand in 3D for better coverage and QoS [26–28]. Several challenges are relevant, including security and the prevention of the hijack of these flying BSs, cooperative relaying for mesh networking, 3D multiple-input-multiple-output (MIMO) beam forming incorporating a set of moving transmitters mounted on UAVs, etc.

Optical-Wireless Convergence: 6G systems will offer Tbps speeds for mobile users. They will also add massive capacity to accommodate new users and offer bandwidth-hungry services by continuously increasing the number of deployed BSs per Km^2. This puts an unprecedented pressure on the backhaul links and the transport network, primarily relying on fibre-optic communications to carry the tremendous volume of user traffic. On the other hand, apart from access networks, wireless communications can also be exploited in the transport network to complement fibre links to release their load [29]. As such, optical and wireless networks are migrating from two independent networks that are configured and optimised separately towards a converged unified network that can pool their resource and cooperate tightly for joint optimisation of their shared resources [30, 31]. However, scalable solutions are needed to make reasonable compromise between centralisation of the management of optical and wireless domains and the tractability of the optimisation. Extended discussion on this topic will follow later in Sect. 1.5.7.

1.5.1 Quantum Communication

The field of quantum communication refers to the techniques capable of exploiting quantum mechanics towards gaining increased control over generation, detection, and transport of optical signalling. Some of the fields that fall under the umbrella of quantum communication offer exciting features for B5G networks, including quantum switching and computing, QKD, and quantum structure-based high bandwidth optical transceivers for THz applications [31–37]. These concepts may collectively greatly contribute to key aims of the next-generation networks, in enabling secure and *entanglement*-based communications, operation over a massive amount of unlicensed spectrum, and high data rates [34, 35, 38, 39]. However, quantum communication, particularly on the level of integration with classical systems, is still in relatively early stages of research [34].

Quantum technology innovation is emerging in the access network scenarios at various levels, where schemes leveraging qubits, QKD schemes, and optical to THz conversion have been demonstrated [35, 36, 40]. The advances made in the field of photonics allow broadband tunability of the THz carrier frequency and direct optical to THz conversion of wavelength-division multiplexing (WDM) signals using ultrafast uni-travelling-carrier (UTC) photodiodes. Various photonics-enabled wireless transmission schemes on carrier frequencies beyond 300 GHz and over 10 Gbps data rates have been demonstrated, while some ventured beyond 100 Gbps aggregate data rate transmission [41–43]. It is, however, of note that the highest reported single-channel error-free transmission data rate is currently 50 Gbps [43] and that most of the aforementioned implementations are limited to a very modest reach of up to 1 m.

Multiuser QKD using quantum transceivers and architecture compliant with point-to-multipoint passive optical network (PON) has been demonstrated in 2013, offering a proof of concept for high cryptographic security in PONs [44]. Secret key exchange may be performed among 64 users sharing a single high-speed single-photon detector for reduced cost. In 2015, the same group proposed a quantum-secured gigabit-capable PON (GPON), leveraging QKD while investigating the impairments of Raman scattering stemming from the co-propagation of classical and quantum signals [45]. In 2019, this research was extended to specifically address the peer-to-peer (P2P) traffic, expected to become dominant in the future, enabling P2P communication in secure quantum access without it being established through the central office via using an upstream signal [35]. This enables relaxed fibre spectral occupation and significantly reduced transmission latency. Such developments, however, are still prohibitively costly in terms of practical deployments, especially when considering PON-based optical fronthaul as a network segment. Additionally, having in mind the standardised PON wavelength plan, it is also of note that Raman scattering-induced impairment will likely make QKD implementations even more challenging. Generally, the challenges relating to the convergence of classical and quantum-based systems are expected to be of scientific research interest as we head towards the next-generation networks.

1.5.2 Terahertz Communications

It is widely agreed that 6G networks should achieve greater system capacity (>100 times compared to 5G networks), higher data rate (in the range of Tbps), and greater user density (to adequately support the IoT and Nano-Things paradigms) [46, 47, references therein]. These enhancements are based on the predicted monthly smartphone traffic that should reach 136 exabytes by 2024, i.e. about four times the amount of traffic registered in today's networks (33 exabytes in 2019) [48]. As a result, there is significant interest in the development of innovative solutions for B5G ultra-fast, ultra-dense, and heterogeneous networks. It is generally accepted that there are three major ways to obtain several orders of increase in throughput gain, extreme densification of the communication infrastructure, large quantities of newly available spectrum, and massive antenna systems, allowing a throughput gain in the spatial dimension. One of the solutions to fulfil such demanding requirements is to use bandwidth beyond the microwave and mm-wave [49] spectra, towards higher frequencies in the THz frequency range.

THz band communication is envisioned as a key wireless technology to satisfy real-time traffic demand for mobile heterogeneous network (MHN) [50] systems, addressing the spectrum scarcity and capacity limitations of current wireless systems. Although the frequency regions below the THz bands have been considerably investigated (i.e. the microwave and the far infrared), this is still not the case for the THz bands, mainly due to the lack of mature and cost-effective THz technology. However, recent advancements are enabling practical THz communications systems, and thus it is time for the wireless research community to consider the THz region.

Among the candidate technologies for higher-frequency communication, the THz spectrum offers more exciting potentials than the mm-wave spectrum, enabling the realisation of Tbps wireless links [51]. In addition, THz frequencies enable more directionality than mm-wave due to reduced antenna aperture. The shorter wavelength of THz, when compared to mm-wave, makes it more directional and less prone to free-space diffraction. Moreover, distances between the transmitter and receiver in THz will be much less than in mm-wave, reducing the power consumption on both BS and the UE sides, which can consequently result in disruptive reduction in the carbon footprint of the mobile industry. Due to these promising features, among others, the research community has recently started to explore the THz bands for wireless communication vigorously.

1.5.3 Visible Light Communication

The shift towards mm-wave frequencies in the attempt to unlock additional spectral availability has led to new considerations on how to tackle the resulting increase in path loss, leading the research community to focus on the "enhancement of the probability of line-of-sight (LoS) techniques".

Compared to architectures relying on femtocells and beamforming techniques, VLC is a technology that aims at providing increased LoS probability and high-frequency operation at lower infrastructure cost [52, 53]. The established link is based on direct intensity modulation of LEDs, at frequencies unperceivable to the human eye. As VLC entails point-to-point communication, it has eventually evolved into a full wireless access networking system, referred to as the Light Fidelity (LiFi). LiFi relies on LED infrastructure for natural beamforming in providing bidirectional multiuser communication, including handover, with the added benefit of multi-purpose operation and energy efficiency [54]. Besides, the wireless capacity bottlenecks are closely related to those of indoor access networks, as up to 80% of wireless traffic is consumed by indoor users [55].

The most obvious benefit of LiFi technology is the vast potential of using up to 300 THz of the visible light spectrum [56]. Data rates as high as 14 Gbps have been demonstrated using LiFi technology paired with WDM [57]. Furthermore, it also enables further reduction of cell size, beyond mm-wave cells [53], as well as intrinsically secure networking since the indoor scenario provides entirely opaque boundaries. Apart from the broadband access, a number of other compelling use cases, such as indoor localisation and V2V communication, are feasible. Clearly, since in providing the aforementioned benefits LiFi relies heavily on LoS communication, when it comes to providing broadband access, it will likely be largely limited to indoors.

Typical LiFi technology limitations and challenges lie in the mitigation of inter-cell interference and link blockage. One promising approach towards meeting those challenges follows the technology convergence paradigm in pursuing a hybrid Wi-Fi-LiFi network (HLWN) architecture [58–60]. As LiFi and Wi-Fi utilise different transmission spectra, the lack of inter-technology interference may enable full exploitation of their individual complementary advantages and result in high data rates and ubiquitous coverage that no single technology is able to provide on its own. However, considering the typical indoor scenario of highly overlapping Wi-Fi and LiFi coverage areas, as well as generally larger Wi-Fi coverage and higher LiFi data rates, signal strength-based strategies typically applied in homogeneous systems are no longer viable and may lead to Wi-Fi network overload [59]. Novel handover and load balancing techniques and impairment mitigation scenarios are thus of interest towards the next-generation, heterogeneous networks.

1.5.4 Virtualised Networks: Network Slicing, Functional Split, and Management

One of the core aspects of B5G/6G platforms is enabling network virtualisation, a networking paradigm that supports highly configurable and dynamic allocation of resources. The result system flexibility supports legacy and backward compatibility on the one hand and seamless evolution and convergence of technologies on the other. This will provide a medium for effectively managing network resources, to ensure enhanced QoS provisioning to the end user, while also taking a step towards reducing the cost of network ownership for the operators. The virtualised infrastructure will detach the network operator from ownership of networks, providing new opportunities in terms of virtual network operators and third-party network infrastructure owners. To understand the benefits of virtualised RAN, it is worthwhile to revisit the enabling technology paradigms that make this possible and considered at the very heart of B5G/6G system design that includes software-defined network (SDN) and network function virtualisation (NFV).

- *Software-Defined Networking*: the scale of the network evolution imposes challenges on its management and configuration in traditional networks. SDN-centralised architecture enables easier management of the large-scale networks by decoupling the control plane from data plane. The logical control plane has the ability of easy configuration along with effective management functions. In software-defined networking, the job of forwarding packets is performed by the centralised network controller through programmable interfaces. The rise in scale of the network imposes stress on the centralised control in SDN. To minimise load on the centralised controller, it is preferable to design an architecture that enables the processing of frequent events near the switches. One possibility brought by the concept of SDN is to implement a hierarchical architecture to reduce the load on the centralised controller. A local controller can be defined to connect to one or more switches, which is further controlled by the root controller. The root controller other than controlling the local controller performs functions that need the network-wide view of the system.
- *Network Function Virtualisation (NFV)*: it is a concept that leverages virtualisation feature for transforming the network node functions into virtual functions that can be further chained together to enable different communication services. One or more virtual machines that run on different network nodes such as network switches and servers can be used to run a virtual network function. NFV uses commodity hardware to run software virtualisation techniques for implementation of network functions. The virtual appliances have the advantage of instantiating network functions without installation of new hardware.

The key benefits of virtualisation include network slicing, functional split design, and effective network management.

Network Slicing

Network slicing enables one physical network to be sliced into multiple, virtual, end-to-end (E2E) networks, each logically isolated, including the network devices across the access, transport, and core networks. A slice in this new network architecture is allocated to each service type, with distinct QoS characteristics and requirements. This permits operators to provide network as a service and can enforce strong isolation between different slices such that any actions in one slice does not affect other operating slices.

Even though there is technology development on network slicing, the majority of them focused on slicing at the core network due to massive advancements in SDN, virtualisation, and NVF technologies. On the other hand, there is limited research on the radio access network (RAN) slicing, in which most of them focused on slicing a single RAN resource, e.g. either spectrum or BS. As a matter of fact, network capacity is inarguably a critical resource of RAN; however, other resources of RAN such as cache space, backhaul capacity, and computing at the RAN also need to be considered for RAN network slicing.

Functional Split

C-RAN is a network architecture in which the BS is disaggregated into two parts, the remote radio heads or units (RRHs or RRUs) and the baseband unit (BBU). By this separation, it is possible to combine more BBUs into one shared pool of BBUs; in other words, numerous RRUs can use one centralised BBU. Sharing and virtualising BBU functions leads to several advantages, such as improved hardware utilisation as well as reduction in deployment costs and operational costs, due to saving in BS's power consumption.

In BSs with RRHs (or distributed BSs), the baseband processing is separated from the RF frontend, forming the RRH where the physical transceiver is located, and connected through an optical fibre link to the BBU or data unit where the baseband processing takes place. Enabling the sharing of one BBU among several RRHs, the architecture results in reduced deployment costs. Moreover, the distance between the RRH and BBU can be up to 40Km, enabling the placement of BBU in more convenient locations, reducing deployment and maintenance costs even further. In general, the BBUs are responsible for the baseband digital signal processing, while the RRUs handle all radio functions, such as DA/AD (digital-to-analogue/analogue-to-digital) signal conversion, amplification, and transmission/reception of signals.

Virtualised Network Management

The main design objective of RAN slicing is to flexibly and adaptively manage RAN resources among slice owners (or tenants), so that the RAN infrastructure can be more efficiently utilised. In the meantime, it is necessary to maintain a certain degree of independence among slices (i.e. performance isolation and functional isolation), so that the operators can maintain full control of their slices to meet their service requirements. Without appropriate slice isolation, service interruptions may happen, leading to poor performance in the multi-service RAN slicing environment.

There are several open challenges with the management of C-RANs. For example, designing a radio resource scheduling strategy (RSS) for virtual RANs is much more complicated than the traditional RANs, and the legacy RSS approaches cannot be applied. Therefore, novel RSSs dedicated to RAN slicing need to be developed, with the aim to maximise resource utilisation subject to slice isolation requirements.

Moreover, another key challenge is to provide robust service connectivity to end users in case of high mobility. However, the current approaches for network slicing are not designed to handle mobility in the network. Indeed, handling and orchestrating the radio access and core network will be very challenging in case of mobility which would require migration of services from one point to another across the network. Moreover, a strong coordination among multiple cells would also be required to handle such cases. In addition, inter-slice coordination might also be required to handle mobility as a single slice operator might not have sufficient resources in a specific area to support its mobile users.

Computing-as-a-Service
Future networks will offer computing services, e.g. for IoT applications. This will play a vital role in realising a number of context-aware real-time services. One option, e.g. would be to adopt cloud computing for such services. Cloud computing enhances the user's overall quality of experience (QoE) by providing the shared computing and storage resources online as a service in an elastic, sustainable, and reliable manner within a virtual container. Nevertheless, sensitive applications incur performance degradation because of the distance between the cloud and the end user. Therefore, edge computing is a solution for delay-sensitive applications that pushes the computing resources to the edge of the network. In fact, MEC, as standardised by 3GPP, allows the placement of storage and computing resources at the BS. However, the computing paradigm is currently limited to the edge network, which includes the BS, the BBU, and the RRH, that is commonly referred to as fog computing. However, 6G aims to push back the boundaries on the edge network to include mobile devices that can be referred to as dew computing. This will represent the next step in softwarisation to reduce further the computing latency while also offering new opportunities for task offloading and reducing further the energy consumption in mobile devices.

1.5.5 Artificial Intelligence and Big Data

AI is generally a computing system that is able to interact with its surrounding environment. It can *sense* the outside world, by collecting data, *mine* it to *predict* potential variations, and *reason* to best respond to those changes in the environment [61].

ML is a subset of AI. It is defined as the science of making computers know how to perform a certain task without explicitly being programmed, rather by letting them generalise from a training dataset. Learning algorithms are classified into three main categories: supervised, unsupervised, and reinforcement learning (RL). The first two types rely on mining data to build a predictive model. The training dataset can be either labelled or unlabelled. Consequently, the corresponding learning algorithm can be supervised or unsupervised, respectively. In supervised learning, the algorithm is provided with the true output values (labels) of training examples, whereas in unsupervised learning, it is up to the learning algorithm to figure out the true output labels. The output variable (label) can take on either continuous (numeric) or discrete (nominal) values. Based on that, the learning algorithm performs either a regression or a classification/clustering task.

Unlike supervised or unsupervised learning that fundamentally relies on a training dataset, RL learns from taking explorative actions and its gained experience from interacting with the environment. After taking each action, the algorithm evaluates how well this action helps the system approach its objective. Over time, it learns an optimal policy that returns the best action maximising its long-term cumulative reward [61].

Big data is the science of developing scalable algorithms to analyse tremendously large volumes of data, e.g. generated by smart sensor, IoT nodes, etc., that cannot be handled by traditional techniques [64, 70–72].

In general, AI is a promising tool to cope with continuously growing complexity of mobile networks by making them self-adaptive, self-healing, and self-reconfigurable based on, e.g. the traffic load, QoS requirement, interference/noise level, etc. [73]. For example, it is argued that 80 per cent of power consumption of mobile network occurs at RANs, especially at BSs, since the planning of BSs is currently based on the peak traffic and disregards the huge variations in the traffic load during different hours of a day [73]. To that end, an AI-empowered BS can help save a considerable deal of energy by intelligently adjusting the transmit power according to instantaneous variations in the traffic load.

Regression techniques can be used, e.g. for channel estimation [62] or spectrum sensing [63], whereas classification and clustering can be employed for anomaly detection [64, 65], fault detection [66], or intrusion detection [67, 68]. RL may also be employed for power control, channel selection, or self-configuration/optimisation of femtocells [69]. For example, Calabrese et al., in [61], apply RL for radio resource management, proposing an architecture splitting the learner and the actor. Every BS has an independent actor that takes actions and shares its experience with a central learner module sitting in the core network, and the learner is shared among the actors. This gives different local agents to share their explorative experience to find an optimal resource management policy more effectively.

However, despite these efforts, the exploitation of AI, ML, and big data techniques in mobile networks is still in early research stages, and the technology needs to get mature for real-life implementation and integration into the upcoming 6G networks.

1.5.6 Energy Harvesting and Fog Computing

Sustainability and Energy Efficiency

Sustainable network operation in 5G is imperative to reduce the overall operation cost and take steps towards achieving carbon footprint targets by 2030. Over the last decade, there has been a significant interest in climate change and energy sustainability for information and communications technologies [77–80]. As a matter of fact, telecommunication networks are one of the leading sources of global carbon dioxide emissions [81]. In addition, the unavailability of a reliable energy supply from electricity grids in some areas is forcing mobile network operators (MNOs) to use sources like diesel generators for power, which not only increase operating costs but also contribute to pollution. Energy harvesting is a technology that allows infrastructure nodes to be powered by the energy converted from the environment, such as sunlight, wind power, and tides, and has the potential to provide a sustainable energy source with zero carbon emission [82–87].

Sustainability in B5G/6G networks can be achieved adopting either energy-efficient design or renewable energy sources. Mobile nodes may operate on harvested energy either from renewable energy sources (e.g. solar, wind, etc.) or from RF signals for battery-less operation or for extending their battery lifetime. However, designing energy-efficient systems with energy harvesting is a challenging task as there are significant random variations in the harvesting of energy from the natural sources or the electromagnetic radiations. Therefore, it is practically more feasible to use hybrid energy sources that utilise both harvested and grid or battery energy to enable the continuous operation of the devices.

Fog Computing

To provide low-latency services to end users, a new framework referred to as fog computing was developed [74], in which a large number of closely located and often decentralised devices, fog nodes, can communicate and cooperate with each other to perform certain computational tasks. Fog computing complements existing cloud services by distributing computation, communication, and control tasks closer to end users. According to the Next-Generation Mobile Networks (NGMN) Alliance [75], fog computing will be an important component of 5G systems and beyond, providing support for computation-intensive applications that require low latency, high reliability, and secure services. The success of fog computing heavily relies on the ubiquity and intelligence of low-cost fog nodes to reduce the latency and relieve network congestion [74–76].

Allowing fog nodes to utilise the energy harvested from nature can provide ubiquitous computational resources anywhere at any time. Fog nodes can rely on renewable energy sources to support low-latency, real-time computation. Incorporating energy harvesting into the design of the fog computing infrastructure is still relatively unexplored, in contrast to data centres that can be supported by massive photovoltaic solar panels or wind turbines. In fact, fog nodes are often limited in size and location, and it is generally difficult to have a global resource manager that coordinates resource distribution among fog nodes in a centralised fashion. Developing a simple and effective method for fog nodes to optimise their energy and computational resources, enabling autonomous resource management according to the time varying energy availability and user demands, is still an open problem.

1.5.7 Enhanced Optical-Wireless Convergence

One of the key requirements for B5G/6G networks is supporting on-demand, high-rate traffic flow with seamless and spectrally efficient coexistence of services [88]. This will be made possible by pushing further the boundaries on small cell technology and their densification, the advent of massive MIMO (mMIMO) and mm-wave technologies, as well as the application of broadband optical access technology as part of a convergence solution towards meeting high bandwidth demands [88–91]. The mMIMO technology may enable dramatic increase of network capacity at no spectral cost. Streamlined with beamforming and space-division multiplexing (SDM) technologies, it has the potential to alleviate the performance bottleneck, particularly in densely populated areas. Moreover, moving on to higher RF carrier frequencies by introducing mm-wave transmission schemes further enables the densification of antenna elements in an mMIMO system, down to the centimetre scale, as well as the densification of network cells by moving towards the femtocell technology. On top of these key enabling technologies, network centralisation is envisioned as the underlying scenario towards cost-effective, hierarchical, and dynamic asset, radio resources, and interference management. The re-location of processing functionalities from the RRHs to the BBU pools allows simplified remote-site maintenance as well as orchestration techniques such as coordinated interference cancellation and dynamic load balancing. Along this path, the emergence of C-RAN architectures is highly reliant on broadband high-speed and low-latency fibre-optic connectivity. Point-to-multipoint (e.g. power splitter-based) topologies of PON, with already considerable worldwide deployments, are well positioned to offer a potentially cost-effective fronthaul solution [92]. Moreover, supporting diverse mobile traffic is already one of the main market drivers towards future emerging PON standards.

The apparent synergy of optical and wireless technologies is highly promising; however, their integration will have profound repercussions on the network segment between the BBU pool and the RRH, referred to as the optical fronthaul (OFH). Its design and optimisation are non-trivial both technically and economically. Namely, in tandem, the aforementioned technologies potentially impose prohibitive OFH bandwidth requirements threatening to offset the centralisation cost benefits. As such, typical 4G fronthaul schemes relying on digital radio over fibre (RoF) transmission and protocols like the Common Public Radio Interface (CPRI) have become increasingly more impractical with higher scalability demands. As an indicative example, a 100 MHz channel bandwidth would require the fronthaul rate of over 6 Gbps [93]. An 8x8 MIMO operation would further increase the said bandwidth requirement to over 40 Gbps per sector, while high-end mm-wave frequency operation would place additional strain on the bandwidth of opto-electrical system components. Moreover, the transmitted signal in such systems is constantly digitised with bandwidth overhead regardless of user activity, thus exhibiting poor adaptability to network traffic. It is clear that such a fronthaul design applied to a scenario of 100+ antennas per site would not only make for a spectrally inefficient solution even if supported by ultra-dense WDM technology; it would be entirely cost-prohibitive. Any added processing that may help deflate bandwidth overhead, such as compression, may in turn lead to signal degradation and a prohibitive increase in latency. The limitation of state-of-the-art optical technology is also to be accounted for, with electro-optical component bandwidth and associated vendor cost exhibiting a nearly linear relationship [94].

The high bandwidth lab demonstrations certainly go beyond the market limitations; however, their current technological maturity affects the standardisation paths. PON standardisation has come a long way since the already deployed next-generation (NG) PON2, enabling 10 Gbps per wavelength transmission, with aggregate 25 Gbps residential and 40 Gbps business service rates [95]. Since then, PON standardisation has been driven not only by the market demands for higher capacity but also maximising current market presence of existing technologies as well as the reduced cost of operation.

In 2015, the IEEE 802.3ca Task Force has initiated the specification of 25G/50G/100G Ethernet Passive Optical Network (EPON) systems, settling on 50 Gbps system motivated by the techno-economic factors [94, 96]. Recently, IEEE 802.3cs Super-PON Task Force has initiated the specification of 10 Gbps PON focusing on high reach and user count, while higher-speed PON is focusing on specifying 50 Gbps single-wavelength PON [97, 98]. The ability to support future networks over standardised PON technologies for cost-effectiveness and coexistence is thus largely dependent on their capacity requirements. The converged optical-wireless networks are thus rapidly evolving towards a fully centralised solution based on CPRI and digitised RoF (dRoF).

A logical next step is to consider the re-distribution of baseband processing, so that the OFH bandwidth requirement relaxation does not prohibitively offset the benefits of centralisation. Various functional splits have been proposed with this aim [99]. The solutions based on dRoF with optimised functional split, whereby a part of the baseband processing is moved to the RRH such that symbols are transmitted over the OFH rather than the digital samples, enable reduced bandwidth requirement along with statistical multiplexing. Such solution generally outperforms analogue RoF (ARoF) counterparts in terms of nonlinear impairment mitigation [100] and is more mature in terms of standardisation support [99]. However, this comes at a cost of partially reduced benefit of centralisation; it requires potentially costly and power-hungry digital-to-analogue (DAC) conversion at the remote site, as well as the up-conversion to a high-end mm-wave frequency.

A way to maintain higher level of centralisation while potentially coping with high-end mm-wave frequency transmission is to rely on ARoF schemes assisted with optical heterodyning [101]. In such systems, multiple baseband radio signals are up-converted to an intermediate frequency (IF) and electrically multiplexed before being modulated onto an optical carrier, transmitted over fibre, and finally optically up-converted to intended mm-wave frequency. Here, the choice of the particular IF generation scheme will have a great impact on the complexity and cost of required optical components. For instance, weather a single- or a multi-laser technique is employed will have a significant impact on the system power efficiency, robustness to nonlinearity, and/or phase noise impairments and cost (component count). Compared to 4G systems, this is a sound step forward for sub-6 GHz LTE transmission over typical access PON fibre lengths of up to 20Km [102], as the dispersion-induced carrier fading tolerance is increased compared to the typical CPRI-based mobile fronthaul. Depending on the spectral distribution of the chosen IFs for baseband radio signals, however, a scaled-up variant may still be limited in transmission distance due to dispersion-induced power fading. Even more importantly, depending on the optical distribution network topology and the amplifier placement, high signal launch power will induce potentially high-power penalties due to fibre nonlinearity and more so over spectrally efficient WDM schemes. In any case, scalability ought to be carefully addressed in considering increased data rates and mm-wave frequencies for B5G/6G networks.

Recently, direct THz to optical conversion of a wireless THz signal has been implemented as well, by integrating an ultra-wideband silicon-plasmonic Mach-Zehnder modulator at the wireless receiver side [40]. Beyond these transmission frequencies, innovation enabling increased baseband bandwidth of optical modulators and receivers is crucial. In the context of photonics-enabled THz communications, beam steering paired with beamforming technology, the transmitter output power, as well as the bandwidth of system amplifiers are recognised as some of the main challenges and limitations towards future photonics-based systems [103].

The optimised re-distribution trade-off also greatly involves the effort to simplify the potentially costly optical infrastructure, particularly at the remote site. It also means that optical infrastructure ought to be virtualised to the highest possible degree, towards supporting the C-RAN-based heterogeneous, on-demand operation. Virtualisation of optical infrastructure would also bring benefits of more flexible and resilient topologies along with legacy coexistence and re-use of existing fibre deployments. Flexible, SDN-based solutions are expected to greatly contribute to centralised, re-configurable, and vendor-agnostic operation. Moreover, as opposed to committing to a definitive version of PON best suited in this scenario, the preferred aim may be in increasing the overall OFH flexibility in an SDN environment. Although the majority of the aforementioned approaches have complementary merits, it is possible that enabling their hybrid coexistence may be the best option to support heterogeneous operation.

1.6 Ongoing Activities Towards 6G

In this section, we review ongoing research and development activities towards 6G around the globe, e.g. major international initiatives, research projects, forums, and pre-standardisation activities.

1.6.1 6G in Europe

After successful delivery of 5G standard by 3GPP in mid-2020, several initiatives have already been set up in Europe to lead research and development activities towards 6G.

For instance, the pre-standardisation work group of 5G Infrastructure Association (5G IA) – the private side of the 5G Infrastructure Public Private Partnership (5G-PPP) where the European Commission represents the public side – is running a consultation to identify the potential to impact standardisation from the expected timeline, phases, and key areas of work for B5G and 6G research towards 2030. The idea is to collect feedback to help consolidate a B5G and 6G research to support activities related to the European Union (EU) research ecosystem with the final aim of maximising impact on standardisation.

Another major European initiative is the 6Genesis Flagship Program (6GFP), led by University of Oulu. It is an 8-year large-scale research initiative set to ultimately develop, implement, and test key enabling technologies for 6G. As one of the first 6G initiatives in the world, it started in 2018 in collaboration with Nokia, VTT Technical Research Centre of Finland, Aalto University, BusinessOulu, and Oulu University of Applied Sciences. 6GFP aims to develop key enabling technologies for 6G. The team has published a series of 6G white papers and organised two editions of 6G Wireless Summit, in 2019 and 2020.

On 12 November 2020, University of Surrey, UK, announced the launch of its 6G Innovation Centre (6GIC). This centre will be a leading global research hub for 6G focused on advanced telecommunications engineering to bring together the physical and virtual worlds, towards the realisation of the *Internet of senses*. It involves governments, regulators, mobile operators, vendors, enterprises, and leading research and development centres. Capitalising on the wealth of experience gained from the development and validation of 5G technologies at 5G Innovation Centre (5GIC), 6GIC has set its strategic vision for 6G in a recent white paper [104].

Additionally, Hexa-X is a European 6G flagship research project that has been awarded funding from the European Commission under Horizon 2020 (H2020) programme. The project, which is led by Nokia, brings together a competitive consortium of major ICT, industry, and academic stakeholders to lay 6G groundwork and set the direction for future research and standardisation focus areas. The project has begun on 1 January 2021, with a focus on developing the vision for future 6G systems and developing key technology enablers to connect human, physical, and digital worlds. It addresses 6G challenges such as connecting intelligence, network of networks, sustainability, global service coverage, extreme experience, and trustworthiness.

1.6.2 6G in North America

On 21 February 2019, the US president, Donald Trump, announced on Tweeter that he wanted 5G and even 6G in the USA as soon as possible, urging that American companies had to step up their efforts or get left behind.

In May 2020, the Alliance for Telecommunications Industry Solutions (ATIS) – a standard organisation responsible for delivering standards and solutions to advance ICT industry transformation in the USA – issued a call to action to promote US 6G Leadership and shared its vision for collaboration across government, academia, and industry to promote US leadership on the path to 6G.

In Oct 2020, ATIS launched Next G Alliance – an initiative to advance North American mobile technology leadership over the next decade through private sector-led efforts. With a strong emphasis on technology commercialisation, the work will encompass the full lifecycle of research and development, manufacturing, standardisation, and market readiness.

Top US and Canadian operators (e.g. AT&T, Bell, Verizon, T-Mobile, and Telus); tech giants such as Apple, Microsoft, Google, and Facebook; as well as leading telecom and information companies, including Ericsson, Nokia, Qualcomm, Intel, Cisco, Hewlett-Packard, and Keysight Technologies, have already joined the Next G Alliance to create the roadmap to the next decade of strong global mobile technology leadership for the US companies and influence the US government's funding priorities and actions to incentivise the industry.

1.6.3 6G in Asia

On 6 November 2020, China launched the so-called "world's first 6G" test satellite into orbit, which will verify THz communications performance in space. Furthermore, Huawei has already begun research on 6G, according to its chief executive Ren Zhengfei. At a CNBC-hosted panel in September 2019, he said that the company had begun to carry out research on 6G "a long time ago" and had parallel work being done on 5G and 6G – but being in an "early phase". In November 2020, Huawei's executive director gave a keynote speech: "Defining 5.5G for a Better, Intelligent World". Wang said that 5.5G will be an evolution of 5G envisaging real-time interaction experience for individual users, enhancing cellular IoT capabilities, and exploring new scenarios, including Uplink Centric Broadband Communication (UCBC), Real-Time Broadband Communication (RTBC), and Harmonized Communication and Sensing (HCS), in synergy to realise a better, intelligent world. He said that "Going beyond the original three application scenarios to six, 5.5G will take us beyond the Internet of Everything (IoE), enabling the intelligent IoE . . . " In particular:

- UCBC will accelerate the intelligent upgrade of industries that will enable a tenfold increase in uplink bandwidth. This service is targeting manufacturers who need to upload videos in machine vision and massive broadband IoT, enabling the acceleration of intelligent upgrade to industry manufacturing; UCBC can also enhance user experience in indoor scenarios, with larger area coverage and uplink capacity.
- RTBC will deliver an immersive, real-life experience to the end user. RTBC supports large bandwidth and low communication latency (tenfold increase in bandwidth for a given latency and level of reliability).
- HCS enables autonomous driving. By applying the beam scanning technology to mMIMO technology, sensing will be an added service as well as communication. In indoor scenarios, HCS is capable of providing location services.
- Sub-100 GHz usage pattern needs to be restructured. To realise the industry's vision, 5.5G needs to use more sub-100 GHz spectrum for full-band uplink and downlink decoupling and full-band carrier aggregation on demand.
- AI with 5G networks can count on limitless intelligence for autonomous decision-making.

Wang concluded his speech by stating, "Unified standards and industry collaboration are the core DNA that shapes the success of the global wireless communications industry. The development of 5.5G requires collaboration between all parties up and down the value chain".

Japan reportedly intends to dedicate 220 billion yen ($2.03 billion/€1.81 billion) to encourage private sector research and development on 6G. The Japanese government is drawing up a 6G strategy and setting up a dedicated panel on the technology, reports say. The panel will focus on technological development, potential use cases, and policy, according to the Internal Affairs and Communications Ministry. The government hopes to boost public-private cooperation in 6G research and development to gain a lead over other countries.

On 14 July 2020, Samsung released a white paper entitled "The Next Hyper-Connected Experience for All" [12], outlining the company's vision for 6G. Samsung envisions AI as a main foster to 6G. Their vision is that AI will be detrimental to handle the massive amounts of data – associated with hundreds of billions of connected machines and humans – that needs to be collected and utilised in 6G systems. AI will, among other issues:

- Improve performance of handover operation taking into account network deployments and geographical environments
- Optimise network planning involving BS location determination
- Reduce network energy consumption
- Predict, detect, and enable self-healing of network anomalies

Regarding new services, Samsung's vision is that 6G will be driven by mainly three key services:

- Truly immersive extended reality (XR)
- High-fidelity mobile hologram

- Digital replica

These technologies require advanced device form factors to support mobile and active software content, but current mobile devices lack sufficient standalone computing capability. Besides, progress on mobile computing power and battery capacity cannot keep pace with the requirements of these applications. Samsung's vision is that these challenges can be overcome by offloading computing to more powerful devices or servers. The services also have greater demands on wireless capacity, for which the current user-experienced data rate of 5G is not sufficient.

Finally, Samsung's view for 6G networks reflects terrestrial components, e.g. fixed BSs or moving BSs, as well as non-terrestrial components, e.g. air-planes, Urban Air Mobility (UAM) systems, Low Earth Orbit (LEO) and Geostationary Orbit (GEO) satellites, and High-Altitude Platform Stations (HAPS).

1.7 Conclusion

As 5G networks are being deployed, the mobile industry is now focusing on enhancing these systems and envisioning the requirements and specifications for B5G/6G systems. In this context, this chapter opened the discussion for several promising use cases for B5G/6G such as holographic telepresence, digital twin, autonomous vehicles, distributed AI, big data, and blockchain, which can address communications challenges such as digital transformation and fully connected society by the year 2030. To this end, several enabling technologies were discussed for B5G/6G, noticeably network virtualisation, THz and VLCs, quantum communications, small cell hyperdense networks, fog computing, distributed AI and big data, as well as the convergence of wireless and optical networks. The chapter reviewed 6G requirements and foreseen architecture along with ongoing 6G research activities and initiatives around the globe, highlighting future research opportunities and challenges.

Acknowledgments The reach leading to these results was partially funded from i) European Union's Horizon 2020 research and innovation programme under 5GENESIS project with Grant Agreement No. 815178; ii) European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie project EXPLOR with grant agreement No 872897; iii) European Regional Development Fund (FEDER), through COMPETE 2020, POR ALGARVE 2020, Fundação para a Ciência e a Tecnologia (FCT) under i-Five Project (POCI-01-0145-FEDER-030500); iv) ECSEL joint undertaking which is co-funded by the EU H2020 programme under grant agreement 876487 (NEXTPERCEPTION) and national funding agencies in Belgium, the Czech Republic, Finland, Germany, Italy, the Netherlands, and Spain; and iv) the "European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 839573".

References

1. ITU. (2020). ITU completes evaluation for global affirmation of IMT-2020 technologies. *Press Release*. https://www.itu.int/en/mediacentre/Pages/pr26-2020-evaluation-global-affirmation-imt-2020-5g.aspx. Accessed 15 Dec 2020.
2. Saghezchi, F. B. et al. (2015, May 8). Drivers for 5G. *Fundamentals of 5G Mobile Networks*, 1–27. https://doi.org/10.1002/9781118867464.ch1.
3. Morgado, K. M., Huq, S., Mumtaz, S., & Rodriguez, J. (2018). A survey of 5G technologies: regulatory, standardization and industrial perspectives. *Digital Communications and Networks, 4*(2), 87–97. https://doi.org/10.1016/j.dcan.2017.09.010
4. Sucasas, V., et al. (2015). *Efficient privacy preserving security protocol for VANETs with sparse infrastructure deployment* (pp. 7047–7052). 2015 IEEE International Conference on Communications (ICC). https://doi.org/10.1109/ICC.2015.7249450
5. Saghezchi, F. B., Saghezchi, F. B., Nascimento, A., & Rodriguez, J. (2015). *Game-theoretic based scheduling for demand-side management in 5G smart grids* (pp. 8–12). 2015 IEEE Symposium on Computers and Communication (ISCC). https://doi.org/10.1109/ISCC.2015.7405446
6. Saghezchi, F. B., et al. (2019). Machine learning to automate network segregation for enhanced security in industry 4.0. In V. Sucasas, G. Mantas, & S. Althunibat (Eds.), *Broadband Communications, Networks, and Systems (BROADNETS 2018), Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* (Vol. 263, pp. 149–158). Springer. https://doi.org/10.1007/978-3-030-05195-2_15
7. Fettweis, G. P. (2014). The tactile internet: Applications and challenges. *IEEE Vehicular Technology Magazine, 9*(1), 64–70.
8. Viswanathan, H., & Mogensen, P. E. (2020). Communications in the 6G era. *IEEE Access, 8*, 57063–57074. https://doi.org/10.1109/ACCESS.2020.2981745
9. Liu, G., et al. (2020). Vision, requirements and network architecture of 6G mobile network beyond 2030. *China Communications, 17*(9), 92–104. https://doi.org/10.23919/JCC.2020.09.008
10. Saad, W., Bennis, M., & Chen, M. (2020). A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. *IEEE Network, 34*(3), 134–142. https://doi.org/10.1109/MNET.001.1900287
11. Zheng, Z., Xie, S., Dai, H.-N., Chen, X., & Wang, H. (2018). Blockchain challenges and opportunities: A survey. *International Journal of Web and Grid Services, 14*(4), 352–375. https://doi.org/10.1504/IJWGS.2018.095647
12. Samsung Research. (2020). *6G the next hyper-connected experience for all* (White Paper). Accessed 24 Jan 2021 [Online]. Available: https://research.samsung.com/next-generation-communications
13. Ethics guidelines for trustworthy AI | Shaping Europe's digital future. *The EC's High-Level Expert Group on AI* (2019). https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Accessed 25 Jan 2021.
14. Busari, S. A., Saghezchi, F. B., Mumtaz, S., & Rodriguez, J. (2020, September). Multi-objective hybrid scheduler enabling efficient resource management for 5G UDN. In *IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks, CAMAD*, Vol. 2020. https://doi.org/10.1109/CAMAD50429.2020.9209298.
15. Barakabitze, A., Ahmad, R., Mijumbi, A., & Hines, A. (2020). 5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges. *Computer Networks, 167*, 106984. https://doi.org/10.1016/j.comnet.2019.106984
16. Ordonez-Lucena, J., Ameigeiras, P., Lopez, D., Ramos-Munoz, J. J., Lorca, J., & Folgueira, J. (2017). Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges. *IEEE Communications Magazine, 55*(5), 80–87. https://doi.org/10.1109/MCOM.2017.1600935

17. Saghezchi, F. B., Radwan, A., Rodriguez, J., & Dagiuklas, T. (2013). Coalition formation game toward green mobile terminals in heterogeneous wireless networks. *IEEE Wireless Communications, 20*(5), 85–91. https://doi.org/10.1109/MWC.2013.6664478
18. Saghezchi, F. B., Radwan, A., & Rodriguez, J. (2017). Energy-aware relay selection in cooperative wireless networks: An assignment game approach. *Ad Hoc Networks, 56*. https://doi.org/10.1016/j.adhoc.2016.12.001
19. Alam, M., Yang, D., Huq, K., Saghezchi, F., Mumtaz, S., & Rodriguez, J. (2016). Towards 5G: Context aware resource allocation for energy saving. *Journal of Signal Processing Systems, 83*(2). https://doi.org/10.1007/s11265-015-1061-x
20. Saghezchi, F. B., Radwan, A., Rodriguez, J., & Taha, A.-E. M. (2014). *Coalitional relay selection game to extend battery lifetime of multi-standard mobile terminals*. https://doi.org/10.1109/ICC.2014.6883369
21. Busari, S. A., Huq, K. M. S., Mumtaz, S., Dai, L., & Rodriguez, J. (2018). Millimeter-wave massive MIMO communication for future wireless systems: A survey. *IEEE Communications Surveys & Tutorials, 20*(2), 836–869. https://doi.org/10.1109/COMST.2017.2787460
22. Mumtaz, S., Jornet, J. M., Aulin, J., Gerstacker, W. H., Dong, X., & Ai, B. (2017). Terahertz communication for vehicular networks. *IEEE Transactions on Vehicular Technology, 66*(7), 5617–5625. https://doi.org/10.1109/TVT.2017.2712878
23. Cacciapuoti, A. S., Caleffi, M., Tafuri, F., Cataliotti, F. S., Gherardini, S., & Bianchi, G. (2020). Quantum internet: Networking challenges in distributed quantum computing. *IEEE Network, 34*(1), 137–143. https://doi.org/10.1109/MNET.001.1900092
24. Gisin, N., & Thew, R. (2007). Quantum communication. *Nature Photonics, 1*(3), 165–171.
25. Zhang, W., Ding, D.-S., Sheng, Y.-B., Zhou, L., Shi, B.-S., & Guo, G.-C. (2017). Quantum secure direct communication with quantum memory. *Physical Review Letters, 118*(22), 220501.
26. Mozaffari, M., Kasgari, A. T. Z., Saad, W., Bennis, M., & Debbah, M. (2018). Beyond 5G with UAVs: Foundations of a 3D wireless cellular network. *IEEE Transactions on Wireless Communications, 18*(1), 357–372.
27. Sharma, P. K., & Kim, D. I. (2018). Coverage probability of 3-D mobile UAV networks. *IEEE Wireless Communications Letters, 8*(1), 97–100.
28. Sharma, P. K., & Kim, D. I. (2019). Random 3D mobile UAV networks: Mobility modeling and coverage probability. *IEEE Transactions on Wireless Communications, 18*(5), 2527–2538.
29. Siddique, U., Tabassum, H., Hossain, E., & Kim, D. I. (2015). Wireless backhauling of 5G small cells: Challenges and solution approaches. *IEEE Wireless Communications, 22*(5), 22–31. https://doi.org/10.1109/MWC.2015.7306534
30. Abdalla, M., Rodriguez, J., Elfergani, I., & Teixeira, A. (2019). Towards a converged optical-wireless Fronthaul/Backhaul solution for 5G networks and beyond. *Optical and wireless convergence for 5G networks*, IEEE, pp. 1–29.
31. Tzanakaki, A., et al. (2017). Wireless-optical network convergence: Enabling the 5G architecture to support operational and end-user services. *IEEE Communications Magazine, 55*(10), 184–192. https://doi.org/10.1109/MCOM.2017.1600643
32. Khalif, B. N. A., Hasan, J. A. K., Alhumaima, R. S., & Al-Raweshidy, H. S. (2020). Performance analysis of quantum based cloud radio access networks. *IEEE Access, 8*, 18123–18133.
33. Flamini, F., Spagnolo, N., & Sciarrino, F. (2018). Photonic quantum information processing: A review. *Reports on Progress in Physics, 82*, 016001.
34. Di Renzo, M., Debbah, M., Phan-Huy, D. T., Zappone, A., Alouini, M. S., Yuen, C., Sciancalepore, C., Alexandropoulos, G. C., Hoydis, J., De Rosny, J., et al. (2019). Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come. *EURASIP Journal on Wireless Communications and Networking, 2019*, 1–20.
35. Cai, C., Sun, Y., Niu, J., & Ji, Y. (2019). A quantum access network suitable for internetworking optical network units. *IEEE Access, 7*, 92091–92099.

36. Nagatsuma, T., Ducournau, G., & Renaud, C. (2016). Advances in terahertz communications accelerated by photonics. *Nature Photon, 10*, 371–379.
37. Cale, M., & Cacciapuoti, A. S. (2019). Quantum switch for the quantum internet: Noiseless communications through noisy channels. *IEEE Journal on Selected Areas in Communications* arXiv:1907.07432.
38. Welkie, A., Shangguan, L., Gummeson, J., Hu, W., & Jamieson, K. (2017). *Programmable radio environments for smart spaces* (ACM workshop on hot topics in networks). Palo Alto, CA, USA.
39. Bartlett, S. D., Rudolph, T., & Spekkens, R. W. (2003). Classical and quantum communication without a shared reference frame. *Physical Review Letters, 91*(2).
40. Ummethala, S., Harter, T., Koehnle, K., et al. (2019). THz-to-optical conversion in wireless communications using an ultra-broadband plasmonic modulator. *Nature Photonics, 13*, 519–524.
41. Yu, X., et al. (2016). 160 Gbit/s photonics wireless transmission in the 300–500 GHz band. *APL Photon., 1*, 081301.
42. Pang, X. et al. (2016). 260 Gbit/s photonic–wireless link in the THz band. In *Proceedings of 2016 IEEE Photonics Conference (IPC)*, pp. 9–10.
43. Nagatsuma, T., et al. (2016). 300-GHz-band wireless transmission at 50 Gbit/s over 100 meters. In *2016 41st international conference on infrared, Millimeter, and terahertz waves (IRMMW-THz), 2016* (pp. 1–2) https://doi.org/10.1109/IRMMW-THz.2016.7758356
44. Fröhlich, B., Dynes, J. F., Lucamarini, M., Sharpe, Q. W., Yuan, Z., & Shields, A. J. (2013). A quantum access network. *Nature, 501*, 69–72.
45. Fröhlich, J., Dynes, F., Lucamarini, M., Sharpe, A. W., Tam, S. W.-B., Yuan, Z., & Shields, A. J. (2015). Quantum secured gigabit optical access networks. *Scientific Reports, 5*.
46. Fraunhofer. *Beyond 5G -after the next generation*. Fraunhofer Press release. https://www.fraunhofer.de/en/press/research-news/2017/november/beyond-5g-_-after-the-next-generation.html. Accessed 3 July 2019.
47. Akyildiz, F., Jornet, J. M., & Han, C. (2014). Terahertz band: Next frontier for wireless communications. *Physical Communication, 12*, 16–32.
48. Ericsson, A. B. *Traffic exploration tool*. http://www.ericsson.com/TET/trafficView/loadBasicEditor.ericsson. Accessed 3 July 2019.
49. Xiao, M., et al. (2017). Millimeter wave communications for future mobile networks. *IEEE Journal on Selected Areas in Communications, 35*(9, September), 1909–1935.
50. Huq, K. M. S., Jornet, J. M., Gerstacker, W. H., Al-Dulaimi, A., Zhou, Z., & Aulin, J. (2018). THz communications for mobile heterogeneous networks. *IEEE Communications Magazine, 56*(6, June), 94–95.
51. Singh, R., Sicker, D., & Saidul Huq, K. M. (2020). MOTH-Mobility-induced Outages in THz: A Beyond 5G (B5G) application. *2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, USA, pp. 1–9. https://doi.org/10.1109/CCNC46108.2020.9045401.
52. Haas, H. (2013, April). High-speed wireless networking using visible light. *SPIE Newsroom*.
53. Haas, H. (2011, August). *Wireless data from every light bulb*. TED Website.
54. Light Communication. *IEEE 802.11 Task Group*. Available: http://www.ieee802.org/11/Reports/tgbb_update.htm
55. *Cisco Service Provider Wi-Fi: A Platform for Business Innovation and Revenue Generation* (CISCO White paper) (2015).
56. Wu, W., Shen, Q., Wang, M., & Shen, X. S. (2017, May). Performance analysis of IEEE 802.11.ad downlink hybrid beamforming. In *2017 IEEE International Conference on Communications (ICC)*.
57. Tsonev, D., Videv, S., & Haas, H. (2015). Towards a 100 Gb/s visible light wireless access network. *Optics Express, 23*, 1627–1637.
58. Zeng, Z., Dehghani Soltani, M., Wang, Y., Wu, X., & Haas, H. (2020). Realistic indoor hybrid WiFi and OFDMA-based LiFi networks. *IEEE Transactions on Communications, 68*(5), 2978–2991.

59. Wang, Y., & Haas, H. (2015). Dynamic load balancing with handover in hybrid Li-Fi and Wi-Fi networks. *Journal of Lightwave Technology, 33*(22), 4671–4682.
60. Wu, X., & Haas, H. (2020). Load balancing for hybrid LiFi and WiFi networks: To tackle user mobility and light-path blockage. *IEEE Transactions on Communications, 68*(3, March), 1675–1683.
61. Calabrese, F. D., Wang, L., Ghadimi, E., Peters, G., Hanzo, L., & Soldati, P. (2018). Learning radio resource management in RANs: Framework, opportunities, and challenges. *IEEE Communications Magazine, 56*(9), 138–145. https://doi.org/10.1109/mcom.2018.1701031
62. Motade, S. N., & Kulkarni, A. V. (2018). Channel estimation and data detection using machine learning for MIMO 5G communication systems in fading channel. *Technologies, 6*(3, September) Article no. 72. https://doi.org/10.3390/technologies6030072
63. Thilina, K. M., Choi, K. W., Saquib, N., & Hossain, E. (2013). Machine learning techniques for cooperative spectrum sensing in cognitive radio networks. *IEEE Journal on Selected Areas in Communications, 31*(11), 2209–2221. https://doi.org/10.1109/jsac.2013.131120
64. Parwez, M. S., Rawat, D. B., & Garuba, M. (2017). Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network (in English). *Ieee Transactions on Industrial Informatics, 13*(4), 2058–2065. https://doi.org/10.1109/tii.2017.2650206
65. Maimo, L. F., Gomez, A. L. P., Clemente, F. J. G., Perez, M. G., & Perez, G. M. (2018). A self-adaptive deep learning-based system for anomaly detection in 5G networks. *Ieee Access, 6*, 7700–7712. https://doi.org/10.1109/access.2018.2803446
66. Jiang, C. X., Zhang, H. J., Ren, Y., Han, Z., Chen, K. C., & Hanzo, L. (2017). Machine learning paradigms for next-generation wireless networks. *IEEE Wireless Communications, 24*(2), 98–105. https://doi.org/10.1109/mwc.2016.1500356wc
67. Devi, R., Jha, R. K., Gupta, A., Jain, S., & Kumar, P. (2017). Implementation of intrusion detection system using adaptive neuro-fuzzy inference system for 5G wireless communication network. *AEU-International Journal of Electronics and Communications, 74*, 94–106. https://doi.org/10.1016/j.aeue.2017.01.025
68. Li, J. Q., Zhao, Z. F., & Li, R. P. (2018). Machine learning-based IDS for software-defined 5G network. *Iet Networks, 7*(2), 53–60. https://doi.org/10.1049/iet-net.2017.0212
69. Jiang, C., Zhang, H., Ren, Y., Han, Z., Chen, K.-C., & Hanzo, L. (2017). Machine learning paradigms for next-generation wireless networks. *IEEE Wireless Communications, 24*(2), 98–105.
70. Kibria, M. G., Nguyen, K., Villardi, G. P., Zhao, O., Ishizu, K., & Kojima, F. (2018). Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. *Ieee Access, 6*, 32328–32338. https://doi.org/10.1109/access.2018.2837692
71. Zhang, N., Yang, P., Ren, J., Chen, D. J., Yu, L., & Shen, X. M. (2018). Synergy of big data and 5G wireless networks: Opportunities, approaches, and challenges. *IEEE Wireless Communications, 25*(1, February), 12–18. https://doi.org/10.1109/mwc.2018.1700193
72. Zheng, K., Yang, Z., Zhang, K., Chatzimisios, P., Yang, K., & Xiang, W. (2016). Big data-driven optimization for mobile networks toward 5G. *IEEE Network, 30*(1, January–February), 44–51. https://doi.org/10.1109/mnet.2016.7389830
73. Li, R. P., et al. (2017). Intelligent 5G: When cellular networks meet artificial intelligence (in English). *IEEE Wireless Communications, 24*(5), 175–183. https://doi.org/10.1109/mwc.2017.1600304wc
74. Vaquero, L., & Rodero-Merino, L. (2014). Finding your way in the fog: Towards a comprehensive definition of fog computing. *Proceedings of the ACM SIGCOMM Computer Communication Review, 44*(5), 27–32.
75. NGMN Alliance. (2015, February). *5G white paper* [Online]. Available: https://www.ngmn.org/uploads/media/NGMN5GWhite PaperV10.pdf
76. Dastjerdi, V., Gupta, H., Calheiros, R. N., Ghosh, S. K., & Buyya, R. (2016, January). *Fog computing: Principals, architectures, and applications*. ArXiv e-prints.
77. Yi, S., Li, C., & Li, Q. (2015, June). A survey of fog computing: Concepts, applications and issues. In *Proceedings of the ACM Workshop on Mobile Big Data*, Hangzhou, China, pp. 37–42.

78. Yannuzzi, M., Milito, R., Serral-Gracia, R., Montero, D., & Nemirovsky, M. (2014, December). Key ingredients in an iot recipe: Fog computing, cloud computing, and more fog computing. In *Proceedings of the IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks*, Athens, pp. 325–329.
79. *Google cloud and the environment*. Google [Online]. Available: https://cloud.google.com/environment/
80. Apple becomes a green energy supplier, with itself as customer. (2016, August). *New York Times*. [Online]. Available: https://www.nytimes.com/2016/08/24/business/energy-environment/as-energy-use-rises-corporations-turn-to-their-own-green-utility-sources.Html
81. Microsoft environment: Enabling a sustainable future. Microsoft. [Online]. Available: https://www.microsoft.com/en-us/environment/default.aspx [12]. Apple, Facebook, and Google top Greenpeace energy report card. Fortune.com. [Online]. Available: http://fortune.com/2017/01/10/greenpeace-energy-report-apple-facebook-google/
82. Chamola, V., & Sikdar, B. (2016). Solar powered cellular base stations: Current scenario, issues and proposed solutions. *IEEE Communications Magazine, 54*(5, May), 108–114.
83. Ulukus, S., Yener, A., Erkip, E., Simeone, O., Zorzi, M., Grover, P., & Huang, K. (2015). Energy harvesting wireless communications: A review of recent advances. *IEEE Journal on Selected Areas in Communications, 33*(3, March), 360–381.
84. Xiao, Y., Niyato, D., Han, Z., & DaSilva, L. (2015). Dynamic energy trading for energy harvesting communication networks: A stochastic energy trading game. *IEEE Journal on Selected Areas in Communications, 33*(12, December), 2718–2734.
85. Lu, X., Wang, P., Niyato, D., Kim, D. I., & Han, Z. (2015). Wireless networks with RF energy harvesting: A contemporary survey. *IEEE Communications Surveys Tutorials, 17*(2), 757–789.
86. Xiao, Y., Han, Z., Niyato, D., & Yuen, C. (2015, June). Bayesian reinforcement learning for energy harvesting communication systems with uncertainty. In *Proceedings of the IEEE ICC Conference*, London, UK.
87. Ge, X., Yang, B., Ye, J., Mao, G., Wang, C., & Han, T. (2015). Spatial spectrum and energy efficiency of random cellular networks. *IEEE Transactions on Communications, 63*(3, March), 1019–1030.
88. Hossain, E., & Hasan, M. (2015). 5G cellular: Key enabling technologies and research challenges. *IEEE Instrumentation and Measurement Magazine, 18*(3, June), 11–21.
89. Andrews, J. G., Buzzi, S., Choi, W., Hanly, S. V., et al. (2014). What will 5G be? *IEEE Journal on Selected Areas in Communications, 32*(6, June), 1065–1082.
90. Jayachandran, J., Biswas, K., Mohammed, S. K., & Larsson, E. G. (2020). Efficient techniques for in-band system information broadcast in multi-cell massive MIMO. IEEE Transactions on Communications, 68(10, Oct.), pp. 6157–6173. https://doi.org/10.1109/TCOMM.2020.3007497
91. Al-Dulaimi et al. (2018). Emerging technologies in software, hardware, and management aspects toward the 5G era: Trends and challenges. In *5G networks: Fundamental requirements, enabling technologies, and operations management*, IEEE, ch 1, pp. 13–50.
92. Houtsma, V., van Veen, D., & Harstead, E. (2017). Recent progress on standardization of next-generation 25, 50, and 100G EPON. *Journal of Lightwave Technology, 35*, 1228–1234.
93. *Common Public Radio Interface (CPRI)* [Online]. Available: http://www.cpri.info
94. Vujicic, Z. et al. (2016). Considerations on performance, cost and power consumption of candidate 100G EPON architectures. 2016 18th international conference on transparent optical networks (ICTON), IEEE, pp. 1–6, Trento. https://doi.org/10.1109/ICTON.2016.7550683
95. *40-Gigabit-Capable Passive Optical Network (NG-PON2)*. ITU-T G989.x Series of Recommendations.
96. *Physical layer specifications and management parameters for 25 Gb/s and 50 Gb/s passive optical networks*. IEEE 802.3ca Task Force. http://www.ieee802.org/3/ca/index.shtml
97. IEEE P802.3cs Increased-reach Ethernet Optical Subscriber Access Task Force. http://www.ieee802.org/3/cs/index.html

98. *Higher speed passive optical networks*. ITU-T G.9804.x Series of Recommendations. G.9804.1. Consented in July 2019.
99. Larsen, L. M. P., Checko, A., & Christiansen, H. L. (2019). A survey of the functional splits proposed for 5G mobile Crosshaul networks. *IEEE Communications Surveys & Tutorials, 21*(1), 146–172.
100. Jung, H.-D., Lee, K. W., Kim, J. H., Kwon, Y.-H., & Park, J. H. (2016). Performance comparison of analog and digitized rof systems with nonlinear channel condition. *IEEE Photonics Technology Letters, 28*(6, March), 661–664.
101. Rommel, S., et al. (2020). Towards a Scaleable 5G Fronthaul: Analog radio-over-Fiber and space division multiplexing. *Journal of Lightwave Technology, OSA Publishing, 38*(19), 5412–5422.
102. Zhang, J. et al. (2016) Memory-polynomial digital pre-distortion for linearity improvement of directly-modulated multi-IF-over-fiber LTE mobile fronthaul. In *2016 optical Fiber communications conference and exhibition (OFC), IEEE*, pp. 1–3, Anaheim.
103. Nagatsuma, T. (2019). Advances in Terahertz communications accelerated by photonics technologies. *OptoElectronics and Communications Conference (OECC) and 2019 International Conference on Photonics in Switching and Computing (PSC)*, Fukuoka, Japan, pp. 1–3.
104. Tafazolli, R. (2020). *6G wireless: A new strategic vision* (White Paper). 5GIC Strategy Advisory Board. https://www.surrey.ac.uk/sites/default/files/2020-11/6g-wireless-a-new-strategic-vision-paper.pdf

Part II
UDNs for B5G

Chapter 2
Cell-Free MIMO Systems for UDNs

Roya Gholami, Shammi Farhana Islam, Sumaila Mahama, Dirk Slock, Laura Cottatellucci, Alister Burr, and David Grace

Abstract Distributed or cell-free (CF) massive MIMO (MaMIMO) is a key technology for beyond 5G networks, which promises larger capacity density and quality of service, overcoming the shortcomings of mmWave such as reduced antenna size and path loss. In this chapter, favorable propagation properties are studied for a cell-free MaMIMO system for both line-of-sight (LoS) and multipath Rayleigh fading channels. The advantages of distributed arrays over centralized arrays are presented using 3D beamforming, and a low-complexity partially centralized zero-forcing technique is performed to cancel the interference. To deal with the asynchronous transmission between network nodes in CF architecture, filter bank-based multicarrier (FBMC) has been proposed as a potential option to CP-OFDM in distributed and asynchronous network scenarios since it has better spectral efficiency and robustness to synchronization errors. The performance of a bit-interleaving and iterative decoding receiver is studied to remove the intrinsic interference in FBMC systems.

2.1 Introduction

Cell-free (CF) massive MIMO is considered as a promising technology for satisfying the increasing number of users and high-rate expectations in beyond-5G networks. A CF massive MIMO system comprises a large number of geographically

R. Gholami · D. Slock
EURECOM, Sophia Antipolis, France
e-mail: roya.gholami@eurecom.fr; dirk.slock@eurecom.fr

S. F. Islam (✉) · S. Mahama · A. Burr · D. Grace
University of York, York, UK
e-mail: shammi.islam@york.ac.uk; sumaila.mahama@york.ac.uk; alister.burr@york.ac.uk; david.grace@york.ac.uk

L. Cottatellucci
Friedrich-Alexander University, Erlangen, Germany
e-mail: laura.cottatellucci@fau.de

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_2

distributed single-antenna access points (APs) connected to a central processing unit (CPU). The number of APs is significantly larger than the number of users. The system is not partitioned into cells, and each user is served by all APs simultaneously. Since CF massive MIMO combines the distributed MIMO and massive MIMO concepts, it is expected to reap all the benefits of these two systems. In massive MIMO systems, as the number of antennas at the BS increases and the number of users remains constant, the channels between users and BS tend to become jointly orthogonal, resulting in a phenomenon known as favorable propagation. In Sect. 2.2, we study whether the favorable propagation properties, which enable almost-optimal low-complexity detection via matched filtering in massive MIMO systems, hold for CF massive MIMO with two kinds of channels, namely, channels with path loss and transmit and receive antennas in LoS or in multipath Rayleigh fading. Massive MIMO can be implemented by using a large number of antennas at the network end that can serve multiple users at the same time over the same frequency resource either in a centralized or a distributed approach. In centralized massive MIMO, all the transmitting antennas are collocated at a central position, whereas in the distributed or cell-free approach the antennas are distributed among multiple APs over a wide area of service. In this context, Sect. 2.3 investigates a low-complexity partially centralized zero-forcing (ZF) can be implemented for interference cancellation between multiusers.

CP-OFDM has been standardized as the physical layer waveform for 5G and mmWave applications. However, as we head toward 6G, the network architecture will become highly dense and distributed, especially for MTC or IoT applications. The near CF nature of future wireless networks will require a physical layer waveform that is robust against asynchronous transmission between network nodes. As a result, the strict synchronization requirements of CP-OFDM will render it ineffective in a network that considers such a massive deployment of sensor devices. The synchronization overhead (cyclic prefix and guard band) associated with CP-OFDM may be unmanageable. Filter bank-based multicarrier (FBMC) has been proposed as a potential option to CP-OFDM in distributed and asynchronous network scenarios since it has better spectral efficiency and robustness to synchronization errors. However, FBMC shows performance loss due to the loss of complex orthogonality compared to CP-OFDM, which results in high levels of intrinsic interference in FBMC. In Sect. 2.4, we investigate the performance of a bit-interleaving and interactive decoding receiver, with iterative interference cancellation, to remove the intrinsic interference in FBMC systems.

2.2 Favorable Propagation and Detection for Cell-Free Massive MIMO

In recent years, distributed antenna systems (DASs) have emerged as a promising candidate for future wireless communications thanks to their open architecture and flexible resource management. In DASs, APs are distributed over a wide area and connected to a CPU. DASs present great potentials to enhance spectral and power efficiency compared to traditional cellular systems with centralized base stations (BSs). Users' energy consumption is reduced, and transmission quality is improved by reducing the access distance between users and geographically distributed APs. DASs have been extensively studied in downlink, see e.g. [1, 2] and references therein. Fundamental limits of DASs in uplink have been studied in [3–5]. The capacity per unit area of DASs in uplink has been analyzed in [4, 5] leveraging on a mathematical framework based on Euclidean random matrices and assuming that the network dimensions tend to infinity. To reap the benefits promised by this analysis completely, the use of a centralized optimal joint processing is crucial. However, an optimal maximum-likelihood detector has an unaffordable complexity for large systems. Interestingly, in centralized MIMO systems, the high complexity of centralized joint detectors has been successfully addressed by massive MIMO systems [6].

In massive MIMO systems, as the number of antennas at the BS increases and the number of users remains constant, the channels between users and BS tend to become jointly orthogonal determining a phenomenon known as ***favorable propagation*** [7]. Under orthogonality conditions, low-complexity matched filters are optimum and matched filters attain the same performance of maximum-likelihood detectors, asymptotically. To leverage simultaneously on the advantages of DASs and massive MIMO systems, the concept of cell-free massive MIMO was introduced in [8, 9]. A CF massive MIMO system consists of a massive number of geographically distributed single-antenna APs, which jointly serve a much smaller number of users. CF massive MIMO should combine the mentioned benefits of DAS with the advantages of massive MIMO. If favorable propagation held, matched filtering could be again utilized as a low-complexity and almost-optimal detection method.

In massive MIMO systems with centralized BSs, the assumption of Rayleigh fading provides realistic guidelines for system design. However, in DASs where the APs are massively distributed and several of them could be very close to users and in direct LoS, it becomes relevant to investigate the effects of LoS and path loss on the property of favorable propagation. An initial numerical analysis for CF massive MIMO in Rayleigh fading was presented in [10]. In the following, we study favorable propagation conditions for 2D-DASs and consider two extreme cases of DASs, with all antennas in LoS or in NLoS, and Rayleigh fading, and we study analytically their properties. Conditions for favorable propagation are expressed in terms of eigenvalue moments of the channel covariance matrix. We model APs and users as two independent uniform point processes (PPs) over a regular grid [11]. Under this assumption, the inclusion of path loss and LoS or Rayleigh fading leads to classes of random matrices similar to the Euclidean random matrices proposed in [4, 5]. We show analytically that the favorable propagation conditions are not satisfied in CF massive MIMO systems with APs and users in LoS. On the contrary, they hold in the case of path loss plus multipath Rayleigh fading. When matched filtering is not almost optimum, the use of linear multiuser detectors capable to combat multiuser interference at an affordable computational cost becomes really appealing in practical systems. Then, by extending the unified analytical framework proposed in [12], we analyze the performance of both polynomial expansion detectors in [13] and multistage Wiener filters in [14] and show their equivalence in large-scale DAS. Their performance analysis confirms the expectations of the favorable propagation analysis and the substantial benefits of these detectors compared to matched filters when the favorable propagation conditions are not satisfied.

Notation In the following, $\mathbf{i} = \sqrt{-1}$, and the superscripts ^{T}, *, and ^{H} denote the transpose, conjugate, and conjugate transpose operators, respectively. Uppercase and lowercase bold symbols are utilized to denote matrices and vectors, respectively. The expectation and Euclidean norm operators are denoted by $\mathbb{E}(\cdot)$ and $\|\cdot\|$, respectively. $\mathrm{tr}(\cdot)$ and $\mathrm{diag}(\cdot)$ denote the trace and square diagonal matrices whose diagonal elements are given by the elements of the vector argument, respectively. The Kronecker operator is denoted by $\otimes$. Finally, $\mathcal{CN}(\nu, \sigma^2)$ denotes a complex Gaussian distribution with mean ν and variance σ^2.

2.2.1 System Model and Channel Model

We consider a DAS in uplink with users and APs equipped with a single antenna and independently and uniformly distributed over a squared box of side L and area $A = L^2$ in $\mathbb{R}^2$, denoted by $\mathcal{A}_L = \left[-\frac{L}{2}, +\frac{L}{2}\right) \times \left[-\frac{L}{2}, +\frac{L}{2}\right)$. For the sake of analytical tractability, we assume that users and APs are located on a grid in $\mathcal{A}_L$. Let $\tau > 0$ be an arbitrary small real such that $L = \theta\,\tau$ with θ positive, even integer. Let $\mathbf{w} = \big((-\theta+2w_x)\tau/2, (-\theta+2w_y)\tau/2\big)$ with $w_x, w_y \in \mathbb{Z}$, and we denote by $\mathcal{A}_L^{\#}$ the set of points regularly spaced in $\mathcal{A}_L$ by τ, i.e., $\mathcal{A}_L^{\#} \equiv \big\{\mathbf{w}|\mathbf{w} \in \mathcal{A}_L,\ w_x, w_y = 0,\ 1, \ldots \theta - 1\big\}$. We model the distributed users and APs as homogeneous PPs $\Phi_{\mathcal{T}}$ and $\Phi_{\mathcal{R}}$ in $\mathcal{A}_L^{\#}$ characterized by parameters $\beta_T = \rho_T \tau^2$ and $\beta_R = \rho_R \tau^2$, where ρ_T and ρ_R are the intensities, i.e., the number per unit area, of users and APs, respectively. Then, $N_T = \rho_T L^2 = \beta_T\,\theta^2$ and $N_R = \rho_R L^2 = \beta_R\,\theta^2$ are the number of users and APs, respectively.

All the APs are connected to a central processing unit via a back-haul network such that detection is performed jointly. Users transmit at the same power P. At the central processing unit, the discrete-time N_R-dimensional received signal vector is given by

$$\mathbf{y} = \sqrt{P}\mathbf{G}\mathbf{x} + \mathbf{n}, \tag{2.1}$$

where $\mathbf{x}$ is the N_T-dimensional column vector of independent and identically distributed (i.i.d.) transmitted symbols with $\mathbb{E}\{|x_j|^2\} = 1$; $\mathbf{G}$ is the $N_R \times N_T$ matrix of channel coefficients whose (i, j) element $g_{ij} = g(\mathbf{r}_i, \mathbf{t}_j)$ denotes the channel coefficient between transmitter j and receiver i with Euclidean coordinates $\mathbf{t}_j = (t_{x,j}, t_{y,j})$ and $\mathbf{r}_i = (r_{x,i}, r_{y,i})$, respectively. The N_R-dimensional vector $\mathbf{n}$

denotes the complex AWGN vector with i.i.d. components having zero mean and variance σ^2.

In order to define the channel coefficients, we introduce the path loss matrix $\hat{\mathbf{G}}$ with (i, j) element given by

$$\hat{g}_{ij} = \hat{g}(\mathbf{r}_i, \mathbf{t}_j) = \begin{cases} \dfrac{d_0^{\alpha}}{\|\mathbf{r}_i - \mathbf{t}_j\|_2^{\alpha}} & \text{if } \|\mathbf{r}_i - \mathbf{t}_j\|_2 > d_0 \\ 1 & \text{otherwise,} \end{cases} \tag{2.2}$$

where d_0 is a reference distance and α is the path loss exponent. At short distances between APs and users, i.e., for $\|\mathbf{r}_i - \mathbf{t}_j\|_2 \leq d_0$, the loss is assumed to be negligible and $\hat{g}_{ij} = 1$ in order to remove model artifacts. In the case of antennas in LoS, the channel coefficients impaired by path loss are given by $g_{ij} = g(\mathbf{r}_i, \mathbf{t}_j) = \hat{g}(\mathbf{r}_i, \mathbf{t}_j)\exp(-\mathbf{i}2\pi\lambda^{-1}\|\mathbf{r}_i - \mathbf{t}_j\|_2)$, being λ the radio signal wavelength. In the case of NLoS channels with Rayleigh fading, they are given by $g_{ij} = \hat{g}(\mathbf{r}_i, \mathbf{t}_j)h_{ij}$, where $h_{i,j} \sim \mathcal{CN}(0, 1)$ are i.i.d. complex Gaussian variables modeling small-scale fading.

2.2.2 Mathematical Results for DAS and Cell-Free Massive MIMO Analysis

In this section, we introduce mathematical tools for the analysis of DASs and CF massive MIMO systems. Communication systems modeled by random channel matrices can be efficiently studied via their covariance eigenvalue spectrum [15]. Then, we characterize the spectrum of the channel covariance matrix $\mathbf{C} = \mathbf{G}^H\mathbf{G}$ in terms of its eigenvalue moments defined as follows:

$$m_{\mathbf{C}}^{(n)} = \int \mu^n \mathrm{d}F_{\mathbf{C}}(\mu) = \frac{1}{N_T}\mathbb{E}\{\mathrm{tr}(\mathbf{C}^n)\} \qquad n \in \mathbb{N}, \tag{2.3}$$

where μ and $F_{\mathbf{C}}(\mu)$ denote the eigenvalue and empirical eigenvalue distribution of matrix $\mathbf{C}$, respectively.

Following the approach in [4, 5, 16], we decompose the path loss matrix, $\hat{\mathbf{G}}$, as follows:

$$\hat{\mathbf{G}} = \mathbf{\Psi}_R \hat{\mathbf{T}} \mathbf{\Psi}_T^H, \tag{2.4}$$

where $\hat{\mathbf{T}}$ is a $\theta^2 \times \theta^2$ matrix depending only on the function $\hat{g}(\mathbf{r}_i, \mathbf{t}_j)$, and $\mathbf{\Psi}_R$ and $\mathbf{\Psi}_T$ are $N_R \times \theta^2$ and $N_T \times \theta^2$ random matrices depending only on random APs' and users' locations, respectively. In order to define the matrices $\mathbf{\Psi}_R$, $\mathbf{\Psi}_T$, and $\hat{\mathbf{T}}$, we consider the $\theta^2 \times \theta^2$ path loss matrix $\hat{\boldsymbol{\mathcal{G}}}$ of a system with θ^2 transmit and receive antennas regularly spaced in $\mathcal{A}_L^{\#}$. It can be shown [17] that $\hat{\boldsymbol{\mathcal{G}}}$ is a symmetric block Toeplitz matrix of $\theta \times \theta$ Toeplitz blocks, and asymptotically, for

$\theta^2 \to \infty$, it admits an eigenvalue decomposition based on a $\theta^2 \times \theta^2$ 2D discrete Fourier transform[1] (DFT) matrix $\mathbf{F}$ [18]. Then, we consider the decomposition $\hat{\boldsymbol{\mathcal{G}}} = \mathbf{F}\hat{\mathbf{T}}\mathbf{F}^H$, where the matrix $\hat{\mathbf{T}}$ is a deterministic, asymptotically diagonal matrix whose diagonal elements are the discrete Fourier series of the first row of $\hat{\boldsymbol{\mathcal{G}}}$. The random matrices $\boldsymbol{\Psi}_R$ and $\boldsymbol{\Psi}_T$ are obtained by extracting independently and uniformly at random N_R and N_T rows of matrix $\mathbf{F}$.

In the following, we derive a tight approximation of the eigenvalue moments of the channel covariance matrix $\mathbf{C} = \mathbf{G}^H\mathbf{G}$ for two channel models, which can be efficiently applied to the analysis of favorable propagation properties in CF massive MIMO and the design and analysis of multistage linear detectors.

2.2.2.1 Eigenvalue Moments for Antennas in LoS

We derive the eigenvalue moments for DASs with transmitters and receivers in LoS and channel attenuation given by path loss. We extend the results for 1D-DASs presented in [19] to 2D-DASs. The matrix $\mathbf{G}$ for transmitters and receivers in LoS admits a decomposition similar to $\hat{\mathbf{G}}$, i.e., $\mathbf{G} = \boldsymbol{\Psi}_R\mathbf{T}\boldsymbol{\Psi}_T^H$. As in [5, 16], the eigenvalue moments of the channel covariance matrix are obtained by approximating the random matrices $\boldsymbol{\Psi}_R$ and $\boldsymbol{\Psi}_T$ by the independent matrices $\boldsymbol{\Phi}_R$ and $\boldsymbol{\Phi}_T$, respectively, consisting of i.i.d. zero-mean complex Gaussian elements with variance θ^{-2} to obtain matrix $\widetilde{\mathbf{G}} = \boldsymbol{\Phi}_R\mathbf{T}\boldsymbol{\Phi}_T^H$. This approximation enables the application of classical techniques from random matrix theory and free probability. The derivation of the eigenvalue moments follows the techniques proposed in [12, 20].

The results of *Eigenvalue Moments for Antennas in LoS* are summarized in the following proposition.

Proposition 1 *Let $g(\mathbf{r}_i, \mathbf{t}_j)$ be the function of channel coefficients in LoS, $T(f_1, f_2)$ with $(f_1, f_2) \in [-1/2, +1/2]^2$ be the 2D Fourier series of the sequence obtained by sampling $g(\mathbf{r}_i, \mathbf{t}_j)$ over the regular grid $\mathcal{A}_\infty^\#$, and $m_{\mathbf{T}}^{(2\ell)} = \int_{-1/2}^{+1/2}\int_{-1/2}^{+1/2} |T(f_1, f_2)|^{2\ell} \mathrm{d}f_1 \mathrm{d}f_2$. Consider the matrix $\widetilde{\mathbf{C}} = \widetilde{\mathbf{G}}^H\widetilde{\mathbf{G}}$ with $\widetilde{\mathbf{G}} = \boldsymbol{\Phi}_R\mathbf{T}\boldsymbol{\Phi}_T^H$. For $\theta^2, N_R, N_T \to +\infty$ with $N_T/\theta^2 \to \beta_T$ and $N_R/\theta^2 \to \beta_R$, $\widetilde{C}_{kk}^{(\ell)}$, the k-th diagonal element of matrix $\widetilde{\mathbf{C}}^\ell$, and $m_{\widetilde{\mathbf{C}}}^{(\ell)}$, the eigenvalue moment of order ℓ of the matrix $\widetilde{\mathbf{C}}$ converge to a deterministic value given by*

$$\widetilde{C}_{kk}^{(\ell)} = m_{\widetilde{\mathbf{C}}}^{(\ell)} = \sum_{n=0}^{\ell-1} \sigma^{(\ell-n)} m_{\widetilde{\mathbf{C}}}^{(n)} \qquad \textit{for any } k \textit{ and } \ell \geq 2, \tag{2.5}$$

[1] The 1D DFT matrix over N points is the $N \times N$ matrix with element in row i and column j given by $(\mathbf{F}_1)_{ij} = \frac{1}{\sqrt{N}} \mathrm{e}^{-2\pi\mathbf{i}(i-1)(j-1)/N}$. The definition can be extended to 2D, and the 2D DFT matrix is given by $\mathbf{F} = \mathbf{F}_1 \otimes \mathbf{F}_1$.

with

$$\sigma^{(\ell)} = \int\int \mathbb{P}^{(\ell)}(|T(f_1, f_2)|^2)\mathrm{d}f_1\mathrm{d}f_2,$$

and $\mathbb{P}^{(\ell)}(|T(f_1, f_2)|^2)$ *polynomial in* $|T(f_1, f_2)|^2$ *recursively given by*

$$\begin{aligned}\mathbb{P}^{(\ell)}(|T(f_1, f_2)|^2) = \beta_T m_{\widetilde{\mathbf{C}}}^{(\ell-1)}|T(f_1, f_2)|^2 \\ + \beta_R\beta_T|T(f_1, f_2)|^2\sum_{s=0}^{\ell-2} m_{\widetilde{\mathbf{C}}}^{(s)}\mathbb{P}^{(\ell-s-1)}(|T(f_1, f_2)|^2) \\ + \beta_T^2|T(f_1, f_2)|^2\sum_{s=0}^{\ell-2}\sum_{r=1}^{\ell-2-s} m_{\widetilde{\mathbf{C}}}^{(s)} m_{\widetilde{\mathbf{C}}}^{(r)}\mathbb{P}^{(\ell-s-r-1)}(|T(f_1, f_2)|^2).\end{aligned} \tag{2.6}$$

The initial values of the recursion are $\widetilde{C}_{kk}^{(0)} = m_{\widetilde{\mathbf{C}}}^{(0)} = 1$, $\mathbb{P}^{(1)}(|T(f_1, f_2)|^2) = \beta_R|T(f_1, f_2)|^2$*, and* $\widetilde{C}_{kk}^{(1)} = m_{\widetilde{\mathbf{C}}}^{(1)} = \sigma^{(1)} = \beta_R m_{\mathbf{T}}^{(2)}$*. Proposition 1 suggests a simple algorithm 1 to determine* $m_{\widetilde{\mathbf{C}}}^{(\ell)}$ *and* $\widetilde{C}_{kk}^{(\ell)}$ *detailed in the following.*

Algorithm 1

Initial step: Let $\mu_0 = \rho_0(x) = 1, \sigma^{(1)} = \mu_1 = \beta_R m_{\mathbf{T}}^{(2)}$, $\rho_1(x) = \beta_R x$, and $\ell = 2$.

Step ℓ:
- Define polynomial in x
  $\rho_\ell(x) = \beta_T\mu_{\ell-1}x + \beta_R\beta_T x\sum_{s=0}^{\ell-2}\mu_s\rho_{\ell-s-1}(x)$
  $+ \beta_T^2 x\sum_{s=0}^{\ell-2}\sum_{r=1}^{\ell-2-s}\mu_s\mu_r\rho_{\ell-s-r-1}(x)$ and write it as a polynomial in x.
- In $\rho_\ell(x)$, replace the monomial $x, x^2, \ldots, x^\ell$ by the moments $m_{\mathbf{T}}^{(2)}, m_{\mathbf{T}}^{(4)}, \ldots, m_{\mathbf{T}}^{(2\ell)}$, respectively, and assign the result to $\sigma^{(\ell)}$.
- Compute $\mu_\ell = \sum_{n=0}^{\ell-1}\sigma^{(\ell-n)}\mu_n$.
- Assign μ_ℓ to $m_{\widetilde{\mathbf{C}}}^{(\ell)}$ and $\widetilde{C}_{kk}^{(\ell)}$.
- Increase ℓ by a unit.

By applying the previous algorithm, we obtain the following eigenvalue moments:

$$m_{\widetilde{\mathbf{C}}}^{(1)} = \beta_R m_{\mathbf{T}}^{(2)}, \tag{2.7}$$

$$m_{\widetilde{\mathbf{C}}}^{(2)} = \beta_R^2\beta_T m_{\mathbf{T}}^{(4)} + \beta_R(\beta_R + \beta_T)(m_{\mathbf{T}}^{(2)})^2, \tag{2.8}$$

$$m_{\widetilde{\mathbf{C}}}^{(3)} = \beta_R^3\beta_T^2 m_{\mathbf{T}}^{(6)} + 3\beta_R^2\beta_T(\beta_R + \beta_T)m_{\mathbf{T}}^{(2)}m_{\mathbf{T}}^{(4)} + \left[\beta_R\beta_T(3\beta_R + \beta_T) + \beta_R^3\right](m_{\mathbf{T}}^{(2)})^3. \tag{2.9}$$

2.2.2.2 Eigenvalue Moments for Rayleigh Fading Channel

We consider the channel matrix for Rayleigh fading given by $\mathbf{G} = (\hat{g}_{ij}h_{ij})_{i=1,\ldots N_R}^{j=1,\ldots N_T}$, and we determine an asymptotic approximation of its eigenvalue moments by approximating $\hat{\mathbf{G}}$, the path loss matrix, by $\breve{\mathbf{G}} = (\breve{g}_{ij})_{i=1,\ldots N_R}^{j=1,\ldots N_T} = \mathbf{\Phi}_R \hat{\mathbf{T}} \mathbf{\Phi}_T^H$. Then, the following result holds.

Proposition 2 *Let $\hat{g}(\mathbf{r}_i, \mathbf{t}_j)$ be the path loss function, $\hat{T}(f_1, f_2)$, with $(f_1, f_2) \in [-1/2, 1/2]^2$, be the 2D discrete Fourier series of the sequence obtained by sampling $\hat{g}(\mathbf{r}_i, \mathbf{t}_j)$ over a regularly spaced grid $\mathcal{A}_\infty^\#$, and $m_{\hat{\mathbf{T}}}^{(2\ell)} = \int_{-1/2}^{+1/2}\int_{-1/2}^{+1/2} |\hat{T}(f_1, f_2)|^{2\ell}\,\mathrm{d}f_1\,\mathrm{d}f_2$. Consider the matrix $\widetilde{\mathbf{G}} = (\breve{g}_{ij}h_{ij})_{i=1,\ldots N_R}^{j=1,\ldots N_T}$. As $L \to +\infty$, the eigenvalue moment of order ℓ of the matrix $\widetilde{\mathbf{C}} = \widetilde{\mathbf{G}}^H\widetilde{\mathbf{G}}$ converges to the deterministic value given by*

$$m_{\widetilde{\mathbf{C}}}^{(\ell)} = (m_{\hat{\mathbf{T}}}^{(2)})^\ell \sum_{k=0}^{\ell-1} \frac{1}{k+1} \binom{\ell-1}{k}\binom{\ell}{k} \beta_T^k\,\beta_R^{\ell-k}. \tag{2.10}$$

The first three eigenvalue moments of the covariance matrix for Rayleigh fading channel converge to the following values:

$$m_{\widetilde{\mathbf{C}}}^{(1)} = \beta_R m_{\hat{\mathbf{T}}}^{(2)}, \tag{2.11}$$

$$m_{\widetilde{\mathbf{C}}}^{(2)} = \beta_R(\beta_R + \beta_T)(m_{\hat{\mathbf{T}}}^{(2)})^2, \tag{2.12}$$

$$m_{\widetilde{\mathbf{C}}}^{(3)} = \left[\beta_R\beta_T(3\beta_R + \beta_T) + \beta_R^3\right](m_{\hat{\mathbf{T}}}^{(2)})^3. \tag{2.13}$$

2.2.3 *Favorable Propagation in Cell-Free Massive MIMO*

In this section, we analyze the favorable propagation conditions in DASs through the characteristics of the channel eigenvalue moments. In a favorable propagation environment, when the users have almost orthogonal channels, the channel covariance matrix $\mathbf{R}$ is almost diagonal and satisfies the following properties:

$$\frac{m_{\mathbf{R}}^{(\ell)}}{\mathrm{tr}[(\mathrm{diag}(\mathbf{R}))^\ell]} \approx 1 \qquad \forall \ell \in \mathbb{N}^+, \tag{2.14}$$

where $m_{\mathbf{R}}^{(\ell)}$ denotes the ℓ-order eigenvalue moment of matrix $\mathbf{R}$. These properties are asymptotically satisfied for centralized massive MIMO systems, in rich scattering environments, when the number of users stays finite while the number of antennas at the central base station tends to infinity. By making use of the observation that in large DAS, as $L \to \infty$, $\widetilde{C}_{kk} = \beta_R m_{\mathbf{T}}^{(2)}$, we obtain that $\mathrm{tr}[\big(\mathrm{diag}(\widetilde{\mathbf{C}})\big)^{\ell}] = \beta_R^{\ell} (m_{\mathbf{T}}^{(2)})^{\ell}$ such that (2.14) specializes for DAS with LoS channel and $\ell = 2, 3$ as follows:

$$\frac{m_{\widetilde{\mathbf{C}}}^{(2)}}{\beta_R^2 (m_{\mathbf{T}}^{(2)})^2} = 1 + \frac{\beta_T}{\beta_R} + \beta_T \frac{m_{\mathbf{T}}^{(4)}}{(m_{\mathbf{T}}^{(2)})^2} \tag{2.15}$$

$$\frac{m_{\widetilde{\mathbf{C}}}^{(3)}}{\beta_R^3 (m_{\mathbf{T}}^{(2)})^3} = 1 + 3\frac{\beta_T}{\beta_R} + \frac{\beta_T^2}{\beta_R^2} + 3\beta_T \left(1 + \frac{\beta_T}{\beta_R}\right) \frac{m_{\mathbf{T}}^{(4)}}{(m_{\mathbf{T}}^{(2)})^2} + \beta_T^2 \frac{m_{\mathbf{T}}^{(6)}}{(m_{\mathbf{T}}^{(2)})^3}. \tag{2.16}$$

As $\beta_R \to \infty$, while β_T is kept constant, i.e., for $\beta_T/\beta_R \to 0$ and $\beta_T > 0$, the ratios (2.15) and (2.16) converge to the following limiting values:

$$\frac{m_{\widetilde{\mathbf{C}}}^{(2)}}{\beta_R^2 (m_{\mathbf{T}}^{(2)})^2} \to 1 + \beta_T \frac{m_{\mathbf{T}}^{(4)}}{(m_{\mathbf{T}}^{(2)})^2} \tag{2.17}$$

$$\frac{m_{\widetilde{\mathbf{C}}}^{(3)}}{\beta_R^3 (m_{\mathbf{T}}^{(2)})^3} \to 1 + 3\beta_T \frac{m_{\mathbf{T}}^{(4)}}{(m_{\mathbf{T}}^{(2)})^2} + \beta_T^2 \frac{m_{\mathbf{T}}^{(6)}}{(m_{\mathbf{T}}^{(2)})^3}, \tag{2.18}$$

and conditions (2.14) are not satisfied, so CF massive MIMO systems with APs and users in LoS do not offer favorable propagation.

2.2.4 *Favorable Propagation Condition in an Uplink DAS with Rayleigh Fading Channel*

For DASs with path loss and Rayleigh fading channel, the moment ratios in (2.14) converge to one for all $\ell \geq 1$, as $\beta_T/\beta_R \to 0$ and $\beta_R \to \infty$ as shown in the following:

$$\begin{aligned}\frac{m_{\widetilde{\mathbf{C}}}^{(\ell)}}{\mathrm{tr}\left[\left(\mathrm{diag}(\widetilde{\mathbf{C}})\right)^{\ell}\right]} &= \frac{\beta_R^{\ell} (m_{\hat{\mathbf{T}}}^{(2)})^{\ell} \sum_{k=0}^{\ell-1} \frac{1}{k+1} \binom{\ell-1}{k} \binom{\ell}{k} \left(\frac{\beta_T}{\beta_R}\right)^k}{\beta_R^{\ell} (m_{\hat{\mathbf{T}}}^{(2)})^{\ell}} \\ &= 1 + \sum_{k=1}^{\ell-1} \frac{1}{k+1} \binom{\ell-1}{k} \binom{\ell}{k} \left(\frac{\beta_T}{\beta_R}\right)^k \to 1.\end{aligned} \tag{2.19}$$

Then, conditions (2.14) are satisfied and Rayleigh fading channel offers favorable propagation.

2.2.5 Performance Analysis of Multistage Detectors

Systems with favorable propagation can efficiently utilize the low-complexity matched filter at the central processing unit since it achieves almost-optimal performance in such environments. However, when conditions (2.14) are not satisfied and matched filtering is not almost optimum, the use of low-complexity linear multiuser detection becomes very appealing in practical systems, and linear multiuser detectors are expected to provide substantial gains compared to the matched filter. In the following, we consider low-complexity multistage detectors including both polynomial expansion detectors, e.g., [13], and multistage Wiener filters [14], and we analyze their performance in terms of their SINR by applying the unified framework proposed in [12, 21]. In [12], it is shown that both design and analysis of multistage detectors with M stages can be described by a matrix $\mathbf{S}(X)$ defined as

$$\mathbf{S}(X) = \begin{pmatrix} X^{(2)} + \sigma^2 X^{(1)} & \cdots & X^{(M+1)} + \sigma^2 X^{(M)} \\ X^{(3)} + \sigma^2 X^{(2)} & \cdots & X^{(M+2)} + \sigma^2 X^{(M+1)} \\ \vdots & \ddots & \vdots \\ X^{(M+1)} + \sigma^2 X^{(M)} & \cdots & X^{(2M)} + \sigma^2 X^{(2M-1)}, \end{pmatrix} \tag{2.20}$$

and a vector $\mathbf{s}(X) = (X^{(1)}, X^{(2)}, \ldots, X^{(M)})^T$, where $X = m_{\widetilde{\mathbf{C}}}$ for polynomial expansion detectors and $X = \widetilde{C}_{kk}$ for multistage Wiener filters. From the asymptotic property that $\widetilde{C}_{kk}^{(l)} = m_{\widetilde{\mathbf{C}}}^{(l)}$ for any k and l, we can conclude that multistage Wiener filters and polynomial expansion detectors are equivalent in DAS. Additionally, we can determine the performance of a centralized processor implementing M-stage detectors by applying the following expression [12]:

$$\mathrm{SINR}_M = \frac{\mathbf{s}^T(m_{\widetilde{\mathbf{C}}})\mathbf{S}^{-1}(m_{\widetilde{\mathbf{C}}})\mathbf{s}(m_{\widetilde{\mathbf{C}}})}{1 - \mathbf{s}^T(m_{\widetilde{\mathbf{C}}})\mathbf{S}^{-1}(m_{\widetilde{\mathbf{C}}})\mathbf{s}(m_{\widetilde{\mathbf{C}}})}. \tag{2.21}$$

For $M = 1$, a multistage detector reduces to a matched filter, and (2.21) can be applied to determine its performance and SINR_1 yields the SINR at the output of a matched filter.

2.2.6 Simulation Results

In this section, we validate the analytical asymptotic results by simulations and analyze the performance of multistage detectors in large-scale systems. Throughout this section, the channel is characterized by $\alpha = 2$ and reference distance $d_0 = 1$. In Figs. 2.1 and 2.2, we consider a system with transmitters homogeneously distributed with intensity $\rho_T = 5$ over a finite network of area $A = L^2 = 100$, while the

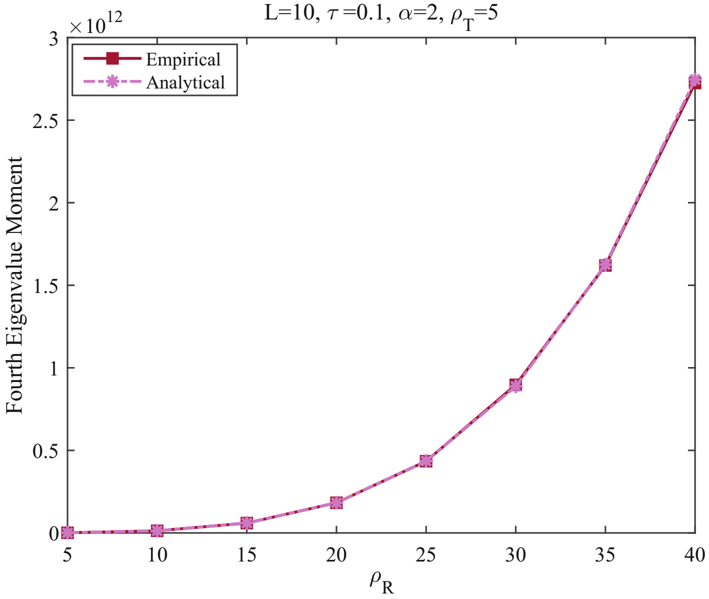


Fig. 2.1 Fourth eigenvalue moment of LoS channel versus ρ_R

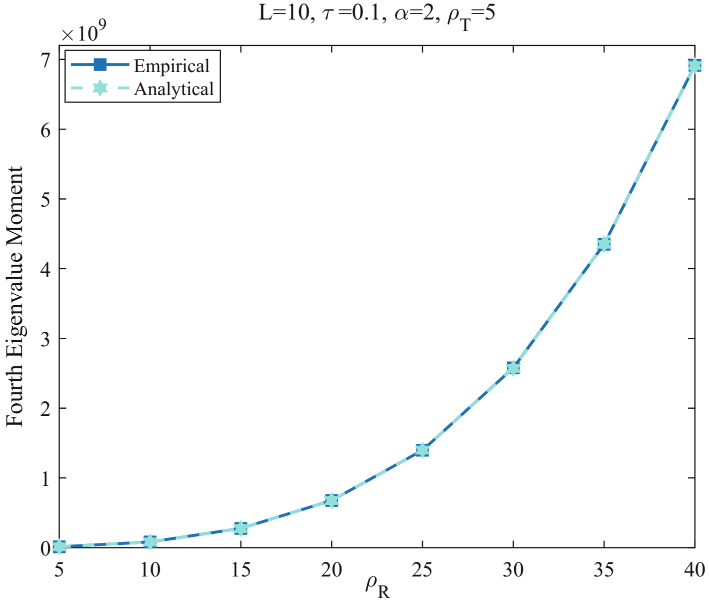


Fig. 2.2 Fourth eigenvalue moment of Rayleigh fading channel versus ρ_R

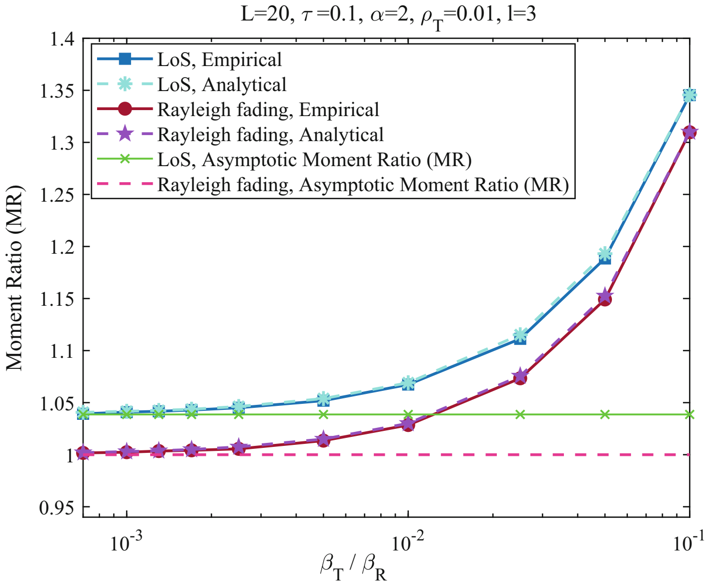


Fig. 2.3 Favorable propagation conditions MR= $m_{\widetilde{\mathbf{C}}}^{(\ell)}/\mathrm{tr}[\big(\mathrm{diag}(\widetilde{\mathbf{C}})\big)^{\ell}]$ versus β_T/β_R

receivers' intensity varies in the range $\rho_R = [5, 40]$. Figure 2.1 compares the fourth eigenvalue moments of LoS channels obtained analytically for $L \to \infty$ by the algorithm with the fourth eigenvalue moments of systems with L finite. Figure 2.2 also compares the fourth eigenvalue moments of NLoS channels with Rayleigh fading obtained analytically as $L \to \infty$ by (2.10) with the fourth eigenvalue moments of systems with L finite. The comparison shows that the asymptotic approximation matches very well practical systems. Figure 2.3 shows the moment ratio $m_{\widetilde{\mathbf{C}}}^{(\ell)}/\mathrm{tr}[\big(\mathrm{diag}(\widetilde{\mathbf{C}})\big)^{\ell}]$ versus $\beta_T/\beta_R = \rho_T/\rho_R$ for $\ell = 3$, $\rho_T = 0.01$, and $\rho_R = [0.1, 15]$. The x-axis is plotted in logarithmic scale. The analytical moment ratios match almost perfectly the ratios for the simulated finite systems of area $A = L^2 = 400$. As predicted analytically, the favorable propagation conditions are not satisfied for the LoS channel, while they hold in the case of the Rayleigh fading. For small ratios β_T/β_R, the curves of LoS and Rayleigh fading converge to the asymptotic moment ratios in (2.18) and (2.19), respectively.

In Figs. 2.4 and 2.5, we consider a system with average SNR at the transmitters equal to 20 dB and analyze the gain of a multistage Wiener filter, or equivalently a polynomial expansion detector over a matched filter in terms of its normalized increase in SINR defined as follows:

$$G = \frac{\mathrm{SINR}_M - \mathrm{SINR}_1}{\mathrm{SINR}_1}. \tag{2.22}$$

Figure 2.4 compares the performance of the two channel models and presents gain G versus ρ_R, the intensity of receivers for M-stage Wiener filters with

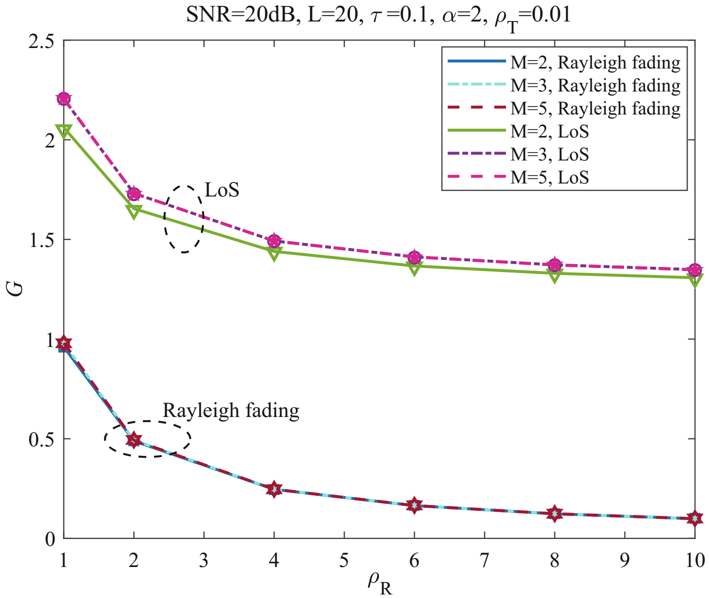


Fig. 2.4 Asymptotic (solid lines) and empirical (markers) gains G versus ρ_R of multistage detectors, with path loss plus LoS or plus Rayleigh fading

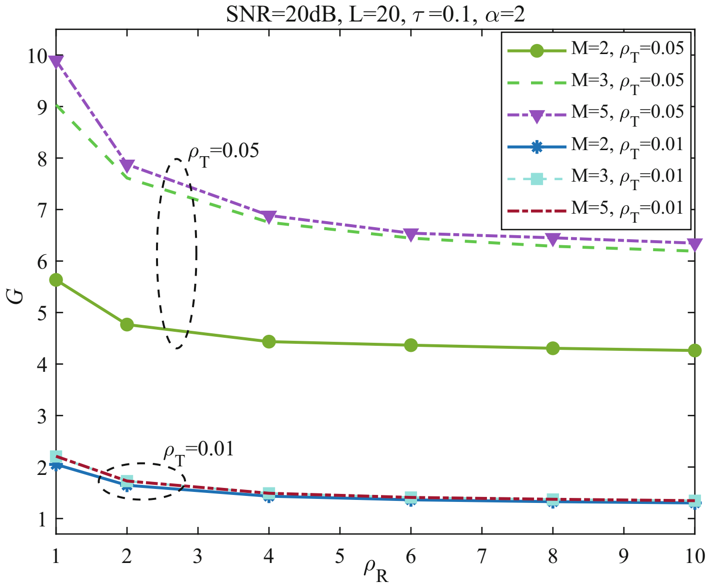


Fig. 2.5 Gain G versus ρ_R of multistage detectors for path loss plus LoS channels and $\rho_T \in \{0.01, 0.05\}$

$M = 2, 3, 5$. The analytical results in solid lines are obtained under the asymptotic assumption $L \to \infty$. The empirical results shown by markers are obtained for $L = 20$. Simulations show an excellent match between the asymptotic performance and the empirical results. In the case of Rayleigh fading, as favorable propagation conditions are satisfied, the performance gap between matched filter and multistage detectors tends to vanish, and gain G becomes negligible as ρ_R increases while ρ_T is kept constant. Then, for ρ_R sufficiently large, the matched filter achieves almost-optimal performance. On the contrary, in the LoS case, the performance gap between the matched filter and the multistage detectors is dramatic with an increase in SINR of about 140% even for systems with 1000 receive antennas per transmitter per unit area. It is interesting to note that for the considered channel models, this dramatic performance enhancement can be attained already with a very simple 2-stage detector, where higher complexity multistage detectors offer only incremental improvements at least at very low system loads.

In Fig. 2.5, we analyze the effect of ρ_T/ρ_R, the system load per unit area, in the case of transmit and receive antennas in LoS. Figure 2.5 shows gain G for $\rho_T = 0.05$ and $\rho_T = 0.01$ as the intensity of receivers varies. Increasing the system load, the SINR increase offered by a 2-stage detector increases enormously, and for higher load, also the increments offered by higher order multistage detectors over a 2-stage detector become significant.

2.2.7 Conclusion

In Sect. 2.2, we considered a CF massive MIMO system in uplink, comprising a *massive* number of *distributed* transmit and receive antennas. In our DAS, transmit and receive antennas are distributed according to homogeneous point processes (PP), and the received signals are processed jointly at a CPU. We showed analytically that the favorable propagation conditions are not satisfied in CF massive MIMO system with transmit and receive antennas in LoS motivating the use and analysis of multistage receivers. On the contrary, they hold in the case of Rayleigh fading. Then, we analyzed the performance of multistage detectors in these two scenarios and showed the relevance of their use especially in systems with antennas in LoS. Simulation results of the favorable propagation conditions and the performance of multistage detectors for finite systems validated the asymptotic analytical results.

2.3 Distributed Antenna Systems: Cell-Free MIMO and Precoding

Massive MIMO has emerged as a key technology in implementing 5G because of its advantages such as high spectral efficiency, energy efficiency, and extended coverage enabling many users to be served simultaneously. Massive MIMO is deployed using a large number of antennas at the network end that can serve multiple users at the same time over the same frequency resource. Implementation can be either centralized or distributed [8, 22]. The sub-6-GHz band used in 4G LTE will not be able to meet the increased data rates in enhanced mobile broadband (eMBB), and hence, mmWave transmission will be required. The huge amount of mmWave spectra with small wavelength (1–10 mm) uses the technology of low-power CMOS radio frequency (RF) miniaturization allowing deployment of a large number of antenna elements with small form factors. With the appropriate beamforming, mmWave massive MIMO schemes can provide sufficient antenna gain to compensate for the free-space path loss incurred at these higher frequencies [23]. Distributed massive MIMO offers fair coverage by exploiting base station diversity or macro-diversity and path loss along with other advantages such as large throughput, coverage probability, and energy efficiency [8]. A large number of works have considered cell-free or distributed massive MIMO, but these mostly consider the band below 6 GHz. The comparison between collocated and fully distributed antenna systems has been made in [24–27]. In [28–30], a distributed approach is studied assuming multi-antenna APs, but not explicitly arranged in antenna arrays such as will be required at mmWave. In this chapter, a practical line-of-sight (LoS) model is used, considering the 3GPP standard for high frequency to analyze the key trade-offs in cell-free or distributed massive MIMO compared to collocated massive MIMO, between concentrating the antenna elements in one location versus distributing the same number over multiple APs in arrays with smaller numbers of elements. Therefore, the processing power within the infrastructure is identical in both cases. We exploit these arrays to perform explicit beamforming, directing a beam toward each user, and hence, full-pilot-based channel estimation is not necessary. The main contributions in sect. (2.3) are as follows:

- Focus on the advantages of using distributed antenna arrays rather than single antennas at each AP; the reduced wavelength at mmWave allows this within a physically small AP. We describe this as ***partially distributed massive MIMO*** rather fully distributed MaMIMO.
- Antenna arrays are configured to perform explicit 3D beamforming, therefore directing the beams toward the desired user.
- Novel reduced complexity ***partially centralized zero-forcing (PC-ZF)*** scheme to eliminate the inter-user interference.

2.3.1 3D Beamforming

3D beamforming is one potential enabling technology for 5G, alongside mmWave, massive MIMO, etc. Many current networks employ horizontal uniform linear beamforming arrays to generate a beam pattern in the azimuth plane: this is known as 2D beamforming (2DBF). Here, we assume uniform planar arrays are used, allowing beam steering in both azimuth and elevation: this is 3D beamforming (3DBF) [31]. 3DBF allows larger arrays giving greater gain and narrower beamwidths, leading to higher user capacity, less inter-user interference, higher energy efficiency, improved coverage, and increased spectral efficiency. Furthermore, with 3DBF, it is possible to serve users at different heights such as high-rise buildings.

2.3.2 System and Channel Model

We consider a pure LoS model based on the standard 3GPP channel model for frequencies from 0.5 to 100 GHz [32] to observe the performance of distributed 3D massive MIMO in the mmWave frequency band for an outdoor environment. In this chapter, we assume LoS propagation only, with no multipath. We use this model to evaluate performance of a distributed antenna array in a $100 \times 100\,\text{m}^2$ open square area where M APs, with $N_h \times N_v$ antenna elements each, serve single or multiple UEs each equipped with a single-antenna element at 26 GHz (https://www.ofcom.org.uk/..data/assets/pdf.file/0014/104702/5Gspectrum-access-at-26-GHz.pdf; https://5gobservatory.eu/europe-to-harmonise-radio-spectrum-in-the-26-ghz-band/). We arranged the AP positions in two ways: one in which all the APs are randomly distributed along the edges of the square area with a minimum 5 m distance between adjacent APs, while in the other, the APs are at fixed positions with an equal distance from each other around the edges. In each arrangement, the advantage of distributed antenna arrays over a single or centralized is observed by distributing a constant number of antenna elements between these smaller antenna arrays compared to centralizing them in one array. In both cases, the users are randomly distributed throughout the service area. For the random distribution of APs, the constant total number is $E_{total} = 100$. For the distributed antenna system, we divide the total elements among different numbers of APs as listed in Table 2.1.

For 2 and 11 APs, E_{total} are 98 and 99, respectively, due to the use of square planar antenna arrays. The overall random distribution of APs and UEs for an instance is illustrated in Fig. 2.6. For the fixed positioned distribution, the total number of antenna elements E_{total} is set to 96, where the antenna arrays are not assumed to be square arrays. The array distribution and the diagram for the

Table 2.1 Array distribution

Number of APs	Array size
1	10 × 10
2	7 × 7
4	5 × 5
11	3 × 3
25	2 × 2

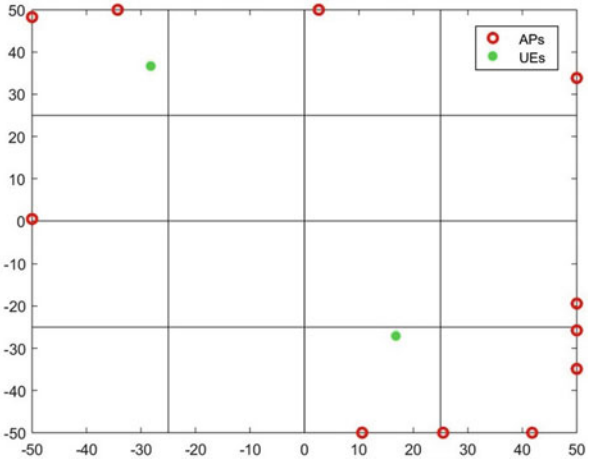


Fig. 2.6 Distribution for the APs and UEs for an instance

Table 2.2 Array distribution for fixed positions of APs

Number of APs	Array size
4 collocated APs at the centre	4 × 6
4 distributed APs	4 × 6
8 distributed APs	3 × 4
16 distributed APs	2 × 3

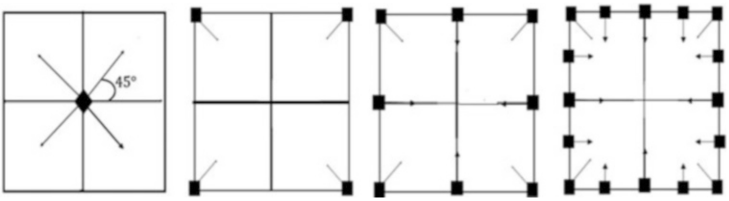


Fig. 2.7 Distribution of APs along the edges of the coverage area with fixed coordinates

distribution are shown in Table 2.2 and Fig. 2.7. The heights of APs and UEs are assumed to be 6 and 1.5 m, respectively. The height of the AP is chosen based on the height of a streetlight in a residential area in the UK, while for the UE it is based on the average height of a person. The total transmitted power is maintained at a constant level of 33 dBm in every scenario over the bandwidth of 1.5 GHz, and the receiver has a noise figure of 10 dB.

Beamforming weights are applied at the APs to direct the signal toward the LoS path, which is assumed to be unobstructed between the AP and the UE. The signals on the direct paths from each AP are assumed to be phase controlled so as to combine coherently at the user. The propagation model follows the inverse square law given by the Friis path loss equation. The received power, $P_{UE,m}$ from the mth AP, is calculated as

$$P_{UE,m} = P_{AP} G_{AP} G_{UE} \left(\frac{\lambda}{4\pi d_m} \right)^2, \tag{2.23}$$

where P_{AP} is the transmitting power. $L = \left(\frac{4\pi d_m}{\lambda}\right)^2$: that is, the path loss depends on the distance d_m and $\lambda = \frac{f_c}{c}$ is the wavelength, where f_c is the carrier frequency and c is the speed of light. G_{AP} is the transmitting antenna gain and is a function of the azimuth and elevation angles between the AP and the UE. Here, we used a square planar antenna arrays with $N_h \times N_v$ elements. Considering the gain of each of the $N_h \times N_v$ elements, $G_{n_h,n_v}(\theta, \phi)$, the array factor $AF(\theta, \phi)$ can be written as

$$\begin{aligned} AF(\theta, \phi) = &\sum_{n_h=1}^{N_h} \sum_{n_v=1}^{N_v} w_{n_h,n_v} \sqrt{G_{n_h,n_v}(\theta, \phi)} \\ &\times e^{j[(n_h-1)kd_x \sin\theta\cos\phi + (n_v-1)kd_y \sin\theta \sin\phi]}, \end{aligned} \tag{2.24}$$

where $k = \frac{2\pi}{\lambda}$ is the propagation constant in free space, w_{n_h,n_v} is the weight of each element and the weights are defined so as to direct a beam from the array in the direction of the LoS to the UE, the antenna spacing is $\frac{\lambda}{2}$, and θ and ϕ are the azimuth and elevation angles, respectively. Assuming 100% efficiency, the gain G_{AP} is given by

$$G_{AP} = \frac{|AF(\theta, \phi)|^2}{\int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} |AF(\theta', \phi')|^2 d\theta' d\phi'} \tag{2.25}$$

G_{UE} is the receiving antenna gain. We assumed the broadsides of the APs are perpendicular to the edges of the square area for the random distribution, and for the fixed positioned distribution, the broadsides are shown in Fig. 2.7. Due to the short link length of AP-UE, we have neglected the atmospheric and other losses.

2.3.3 Zero-Forcing Beamforming

Zero-forcing beamforming (ZFBF) or null steering is a linear precoder that nulls interference between users. ZF could be applied at the APs, or at a central processing unit (CPU) that has access to the signals at all antenna elements on all APs: the former would require channel state information (CSI) to be shared between APs; the latter would require fronthaul connections between CPU and APs to carry signals for all elements. We therefore propose a novel concept, ***Partially Centralized Zero-forcing Beamforming*** (PC-ZFBF), in which APs need only local CSI and only signals for users served by an AP need to be conveyed over the fronthaul. The required direction for the steered beam is obtained from the uplink signal received from that UE, which can be calculated locally [33]. Each beamformer thus has K_{UE} inputs and $N_{tot} = N_h N_v$ outputs: we will represent the beamformer at the m^{th} AP by the matrix U_m, and the channel between this AP and the UTs by the $(K_{UE} \times N_{tot})$ matrix H_m. The overall channel between the inputs of the beamformers and the UTs can then be represented by the $(K_{UE} \times M_{AP} K_{UE})$ composite matrix:

$$H = [H_1 U_1 \quad \dots \quad H_{M_{AP}} U_{M_{AP}}]. \tag{2.26}$$

Note that each element $H_m U_m$ is dominated by its diagonal in that the beamformer should steer a strong signal to each user from its corresponding input: however, it will also in principle have small but non-zero off diagonals since side lobes of the beam pattern may also interfere with the other UEs. We can nevertheless perform zero-forcing precoding using a global precoding matrix U applied at the central processor, which is the pseudo-inverse of H:

$$U = H^H (H H^H)^{-1}. \tag{2.27}$$

The diagonal dominance of the component matrices means that $H H^H$ is also diagonally dominant, and hence, the inversion is numerically stable and will lead to negligible noise enhancement.

2.3.4 Simulation Results and Discussion

In this section, we present initial numerical results from the simulation, demonstrating the distributed antenna arrays for mmWave communication in the 3D environment. Here, we compared the system performance of distributed antenna arrays with that of a single-antenna array keeping the total number of elements constant in a $100 \times 100\text{m}^2$ square area serving multiple users each equipped with a single antenna. The total transmitted power P_{AP} is 33 dBm, and this is also maintained for the distributed cases and is independent of the number of users. We applied the aforementioned zero-forcing precoding to eliminate the inter-user interference. We show the received signal power for random distributions of the users through a cumulative distribution function (CDF) graph. It can be seen that the signals combine coherently in the direct path as the signal amplitudes are summed at the UE from each AP and are then squared to obtain the LoS power. Then, the total LoS powers give the total received power at the user, and the ratio of the total received power to the thermal noise provides the SNR values to plot the overall performance.

Figure 2.8 shows the CDF of the SNR experienced by the two UEs, where the APs are distributed randomly at the edges of the coverage area and the UEs are randomly distributed throughout the coverage area. (The cumulative probability on the vertical axis can also be regarded as the outage probability for the given SNR.) These results are for a pure line-of-sight environment, where we assumed no

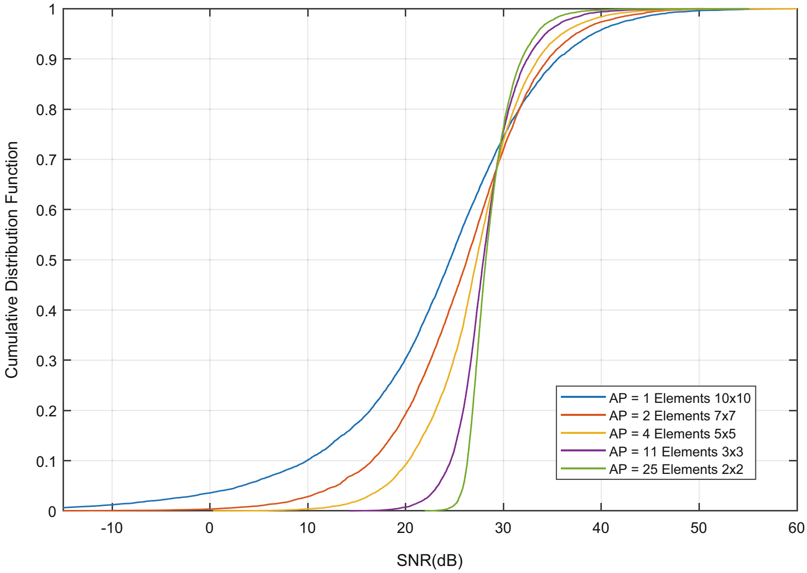


Fig. 2.8 Cumulative distribution function of user SNR for pure LOS channel

multipath propagation. It can also be seen that at 10% outage, the SNR improves by 16 dB from the single-antenna array to the 25 smaller antenna arrays. At mmWave frequencies, distributing the APs over the area provides diversity gain to overcome the path loss, hence increasing the performance. Clearly, it can be stated that splitting the total number of elements between different sites significantly enhances the SNR performance and provides better coverage.

Figure 2.9 shows the CDF of the SNR at the two randomly distributed UEs. Here, APs are specified to certain positions on the sides of the coverage area. For this plot, the total number of antenna elements is 96. Figure 2.7 shows the fixed locations of the APs for four different scenarios, and Table 2.2 shows the array distribution of the 96 elements. In the first scenario, the 4 APs are collocated at the centre and the broadside is at 45° angle. In the second one, 4 APs are distributed at the four corner of the coverage area where the broadside is pointed at 45°. The other two scenarios are shown in Fig. 2.7 with their broadside directions. It is observed from Fig. 2.9 that the distributed arrays perform better than centralized arrays; from collocated APs to 16 distributed arrays, at 10% the SNR improves almost by 10 dB. However, if we compare the performance with 4 APs between random locations, fixed locations, and with the centralized array, at 10% outage, the performance of the randomly distributed APs improves by 3 dB compared to the collocated antenna array, and further, the SNR is increased by 5 dB for the fixed APs compared to the randomly distributed APs. It can be stated that it is better to set up the APs explicitly rather than for them to be positioned randomly. It can also be seen that the improvement between 8 and 16 arrays is not significant as the APs are approaching

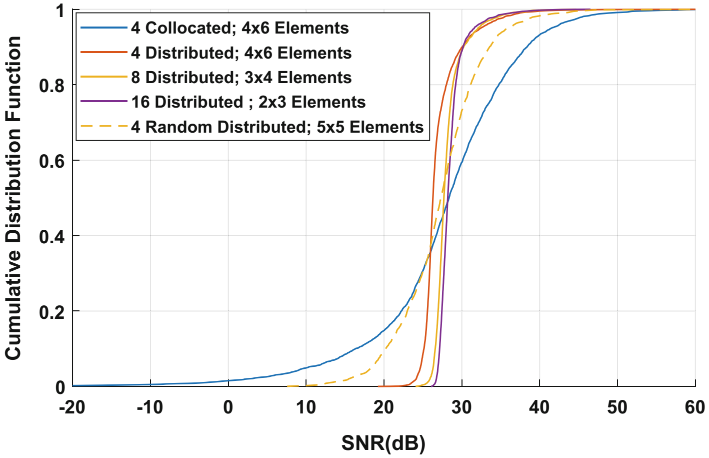


Fig. 2.9 Cumulative distribution function of user SNR for pure LOS channel where the APs are in fixed position

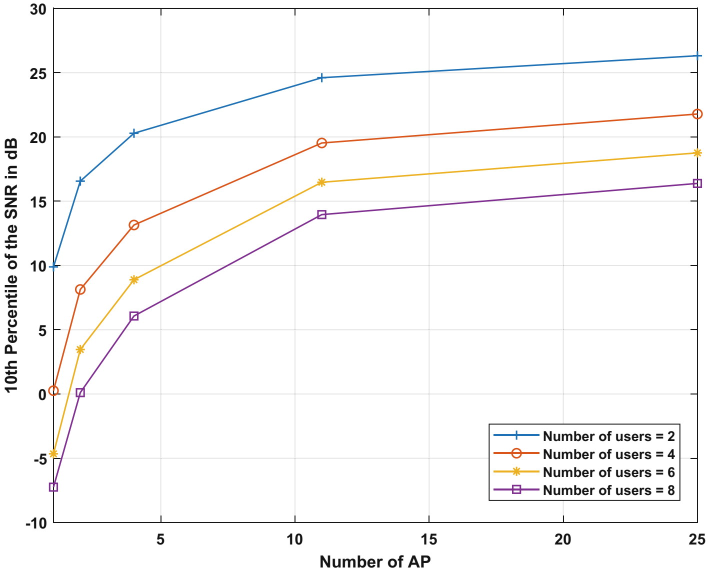


Fig. 2.10 10th percentile of the overall SNR for a different number of UEs

the saturation state, suggesting that it may not be worthwhile to increase the number of APs beyond 8 in this case.

Figure 2.10 shows the 10th percentile of the overall SNR against the number of APs for different numbers of UEs. It is observed from the plot that for any number of users, the distributed approach is performing better than a single array: the higher the number of split arrays the better the performance. On the other hand, as the number of users increases the SNR decreases. This is largely due to the constant total transmitting power being shared among more users. This situation could be improved by an adaptive transmit power scheme at the processor.

2.4 Interference Cancellation for FBMC Waveforms

CP-OFDM is used as the physical layer waveform for 5G and mmWave applications. However, as we head toward 6G, wireless networks will become highly dense and unplanned, especially for MTC or IoT applications. The strict synchronization requirements of CP-OFDM will render it ineffective in a network that considers such massive deployment of sensor devices. The synchronization overhead (CP and guard band) associated with CP-OFDM may be unmanageable [34]. Filter bank based multicarrier (FBMC) systems have attracted increasing research attention in recent years as a potentially feasible option to CP-OFDM since it has better spectral efficiency and robustness to synchronization errors, both features of fundamental importance in future networks [35].

FBMC utilizes an advanced prototype filter on each subcarrier, resulting in improved frequency confinement. This results in a considerable reduction in the out-of-band (OOB) leakage to adjoining frequency bands, compared to CP-OFDM that employs a rectangular pulse shaping filter. This makes FBMC systems suitable for asynchronous transmissions and significantly reduces the overhead due to synchronous communication in CP-OFDM [36]. One of the most well-known FBMC systems is the offset quadrature amplitude modulation (OQAM)-based FBMC (FBMC-OQAM). FBMC-OQAM modulates subcarriers by transmitting the in-phase and quadrature samples with a shift of half the symbol period between them. As a result, FBMC-OQAM system satisfies the orthogonality condition only in the real field. Therefore, in the presence of complex fading channels, intrinsic interference occurs. As a result, the implementation of MIMO techniques such as maximum-likelihood detection (MLD) and space–time block codes (STBC) is not straightforward.

To resolve the problems in FBMC-OQAM, the quadrature amplitude modulation (QAM)-based FBMC has been investigated [37, 38]. By transmitting QAM signals, the FBMC systems cannot guarantee complex domain orthogonality, resulting in high intrinsic interference, i.e., inter-symbol interference (ISI) and inter-carrier interference (ICI). To overcome the intrinsic interference problem, we study an iterative interference cancellation (IIC)-based bit-interleaved coded modulation with an iterative decoding (BICM-ID) receiver [39]. The proposed receiver combines two component decoders.

The received signal is fed to the *inner decoder*, which performs FBMC-QAM demodulation and demapping and passes the soft information at the output of the demapper to the *outer decoder*. The outer decoder (LDPC decoder) performs iterative decoding by passing the received information between its variable node decoder (VND) and check node decoder (CND) for a predefined number of iterations. The signal output by the outer decoder is again fed back to the inner decoder for interference estimation and cancellation. Thus, the intrinsic interference is removed following a scheduled number of iterations of the proposed receiver. Thus, by combining FBMC with the IIC-based BICM-ID receiver, the performance can be comparable with that of CP-OFDM.

2.4.1 Coding-Based FBMC-QAM Modulation

At the transmitter of the coding-based FBMC-QAM system, a stream of information bits is encoded by a channel encoder. A low-density parity-check code (LDPC) encoder is considered in this chapter due to its widespread application in modern communication systems, such as IEEE 802.11n and 5G new radio (NR). The encoded bits are randomly interleaved, QAM modulated, and passed through the FBMC-QAM modulator. The FBMC-QAM modulator consists of a serial-to-parallel converter, inverse fast Fourier transform (IFFT), synthesis filter bank (SFB) block, and parallel-to-serial converter as shown in Fig. 2.11. The signal at the output of the FBMC-QAM modulator is given as

$$\mathbf{x}_n = \mathbf{G}\boldsymbol{\Phi}^H\mathbf{a}_n, \tag{2.28}$$

where $\mathbf{x}_n = [x_{0,n}, x_{1,n}, \ldots, x_{L_p-1,n}]$ is the vector of the transmitted data, $L_p = K \times M$, and K is the overlapping factor and M is the total number of subcarriers. Also, $\mathbf{a}_n = [a_{0,n}, a_{1,n}, \ldots, a_{M-1,n}]$ is the $M \times 1$ symbol vector in the frequency domain before IFFT, $a_{m,n}$ is the data at the m-th subcarrier of the n-th symbol, and $\boldsymbol{\Phi}$ is the $M \times M$ unitary discrete Fourier transform (DFT) matrix whose entry on the l-th row and t-th column is $(1/\sqrt{M})e^{-j\frac{2\pi lt}{M}}$. Note that $\mathbf{G}$ is the poly-phase network (PPN) matrix whose l-th row and t-th column are given as

$$[\mathbf{G}]_{lt} = \begin{cases} g[(l-t)M], & \text{for } 0 \le l-t < K \\ 0, & \text{otherwise} \end{cases} \tag{2.29}$$

with $g[i]$ as the prototype filter coefficient. The PHYDYAS prototype filter [40] is employed in this chapter.

In the presence of a frequency-selective channel, the discrete-time domain FBMC-QAM signal vector at the receiver $\mathbf{y}_n = [y_{0,n}, y_{1,n}, \ldots, y_{L_p-1,n}]$ is given by

$$\mathbf{y}_n = \mathbf{H}\mathbf{x}_n + \mathbf{z}_n, \tag{2.30}$$

where $\mathbf{H}$ denotes the $L_p \times L_p$ multipath channel matrix and $\mathbf{z}_n$ is the $L_p \times 1$ AWGN vector. The received signal is passed through the FBMC-QAM demodulator, i.e.,

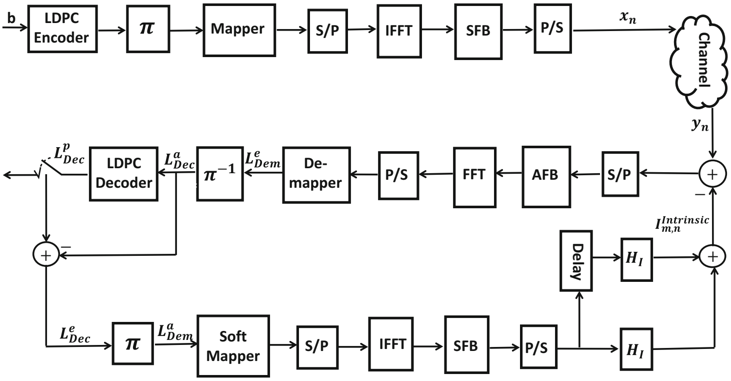


Fig. 2.11 Block diagram of transmitter and BICM-ID receiver

receive filtering, down sampling by a factor K, and then FFT. The output of the FFT is given by

$$\begin{aligned}\mathbf{y}_n^f &= \mathbf{\Phi}\mathbf{G}^H\mathbf{y}_n \\ &= \mathbf{\Phi}\mathbf{G}^H\mathbf{H}\mathbf{G}\mathbf{\Phi}^H\mathbf{a}_n + \mathbf{\Phi}\mathbf{G}^H\mathbf{z}_n \\ &= \tilde{\mathbf{H}}\mathbf{a}_n + \mathbf{z}_n^f,\end{aligned} \tag{2.31}$$

where $\tilde{\mathbf{H}} = \mathbf{\Phi}\mathbf{G}^H\mathbf{H}\mathbf{G}\mathbf{\Phi}^H$ represents the $M \times M$ effective channel matrix after FFT and $\tilde{\mathbf{z}} = \mathbf{\Phi}\mathbf{G}^H\mathbf{z}_n$ is the colored noise. Thus, the received FBMC-QAM signal associated with m-th subcarrier and the n-th symbol, $r_{m,n}$ is expressed as

$$r_{m,n} = \tilde{\mathbf{H}}_{m,n}a_{m,n} + I_{m,n}^{intrinsic} + \tilde{z}_{m,n}, \tag{2.32}$$

where $I_{m,n}^{intrinsic}$ represents the intrinsic interference that is given by

$$I_{m,n}^{intrinsic} = \underbrace{\sum_{i \neq m} \tilde{\mathbf{H}}_{i,n}^{ICI} a_{i,n}}_{ICI} + \underbrace{\sum_{j \neq n} \sum_{m=0}^{M-1} \tilde{\mathbf{H}}_{m,j}^{ISI} a_{m,j}}_{ISI}, \tag{2.33}$$

where $\tilde{\mathbf{H}}_{i,n}^{ICI}$ and $\tilde{\mathbf{H}}_{m,j}^{ISI}$ are the residual channels that lead to ICI and ISI, respectively. For desirable performance, the ICI and ISI terms must be estimated and cancelled from the received signal. To do this, the interference channels $\tilde{\mathbf{H}}_{i,n}^{ICI}$ and $\tilde{\mathbf{H}}_{m,j}^{ISI}$ must be estimated at the receiver. The structure of the interference channel matrix is shown in [41]. Before detection and decoding, a simple one-tap zero-forcing (ZF) equalizer is applied. The resulting signal is represented as

$$\tilde{r}_{m,n} = \frac{r_{m,n}}{\tilde{\mathbf{H}}_{m,n}} = a_{m,n} + \frac{I^{intrinsic}_{m,n}}{\tilde{\mathbf{H}}_{m,n}} + \frac{\tilde{z}_{m,n}}{\tilde{\mathbf{H}}_{m,n}}. \tag{2.34}$$

The equalized signal is deinterleaved and passed to the LDPC decoder for decoding and detection. The receiver processing is repeated for several iterations to get rid of the intrinsic interference through decoding and IIC.

2.4.2 IIC-Based BICM-ID Receiver for FBMC-QAM Systems

In order to recover the transmitted bits, an iterative detection and decoding receiver is proposed as shown in Fig. 2.12. It is made up of two component decoders:

- *Inner Decoder*—which consists of the soft mapper, soft demapper, FBMC-QAM modulator, and demodulator (See Fig. 2.13), and an iterative interference cancellation (IIC) operation

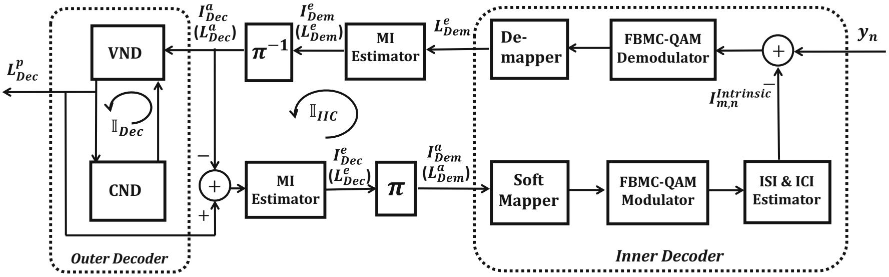


Fig. 2.12 IIC-based BICM-ID system model for EXIT chart analysis

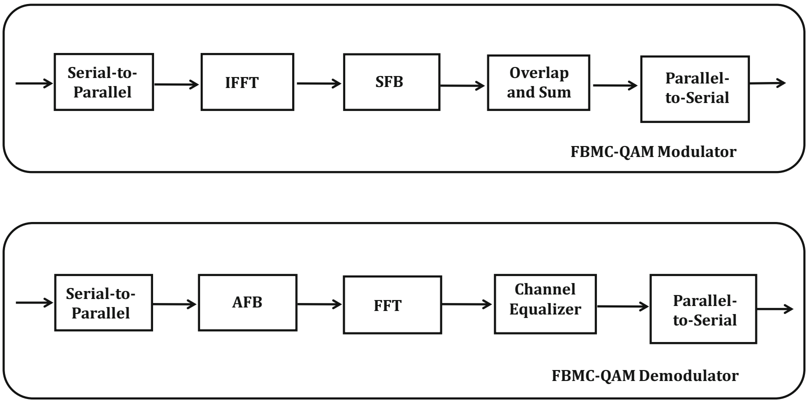


Fig. 2.13 FBMC-QAM modulator and demodulator

- *Outer Decoder*—representing the LDPC decoder, which consists of two types of nodes: VND and CND

To remove the intrinsic interference in FBMC-QAM, the proposed BICM-ID receiver performs two iterative processes: (i) the exchange of mutual information (MI) between the VND and CND of the outer decoder and (ii) the exchange of MI between the inner decoder and the outer decoder. The inner decoder takes the received signal $\mathbf{y}_n$ and the *a priori* information of the coded bits, $L^{q,n}_{Dem,a}$, from the outer decoder in the previous iteration ($L^{q,n}_{Dem,a} = 0$ in the first receiver iteration) and computes the extrinsic log-likelihood ratios (LLRs). The LLR values can be calculated using the maximum *a posteriori* demapping algorithm as shown in [42]. The extrinsic LLR values, $L^{q,n}_{Dem,e}$, are deinterleaved and passed to the outer decoder as *a priori* LLR, $L^{q,n}_{Dec,a}$, for channel decoding. After a number of iterations between the CND and VND of the outer decoder, it computes *a posteriori* LLR, $L^{q,n}_{Dec,p}$. The outer decoder extrinsic information is reinterleaved and passed to the soft mapper as *a priori* information, $L^{q,n}_{Dem,a}$. To estimate and cancel the intrinsic interference term in (2.32), the output of the soft mapper is FBMC-QAM modulated, and the estimated symbols are subtracted from the received signal for the next iteration. We denote the number of iterations between the inner decoder and the outer decoder by $\mathbb{I}_{IIC}$, and the number of iterations within the outer decoder by $\mathbb{I}_{Dec}$. After the final iteration in both component decoders, $L^{q,n}_{Dec,p}$ is used to generate the hard-decision estimates of the transmitted bits.

2.4.3 Simulation Results

The performance of the proposed receiver is presented in the subsection. For the LDPC decoder, an irregular parity-check matrix has been used, whereas the soft-decision message-passing algorithm, sum–product decoding, is employed as demapper [43]. The optimized anti-gray mapping scheme, $M16^r$, is adopted in this simulation. For comparison, synchronous CP-OFDM is considered as a benchmark. For the benchmark implementation, we assume that the CP-OFDM system has sufficient CP and guard band in order to maintain orthogonality and synchronization between subbands. To study the effect of fading on the proposed system, we consider the 3GPP standardized channel models: Extended Pedestrian A (EPA), Extended Vehicular A (EVA), and Extended Typical Urban (ETU) [39]. Perfect channel state information is assumed at both the transmitter and the receiver.

Figure 2.14 shows the bit-error rate (BER) performance of the proposed IIC-based BICM-ID receiver for FBMC-QAM and synchronous CP-OFDM over the LTE EPA channel. As can be seen, the BER performance of FBMC-QAM can be significantly improved after a few inner and outer decoder iterations. For example, measured at a BER of 10^{-6} for $\mathbb{I}_{Dec} = 1$, the proposed receiver shows about 3, 4, and 5 dB SNR gain for $\mathbb{I}_{IIC}$ values of 1, 2, and 3, respectively, when compared with

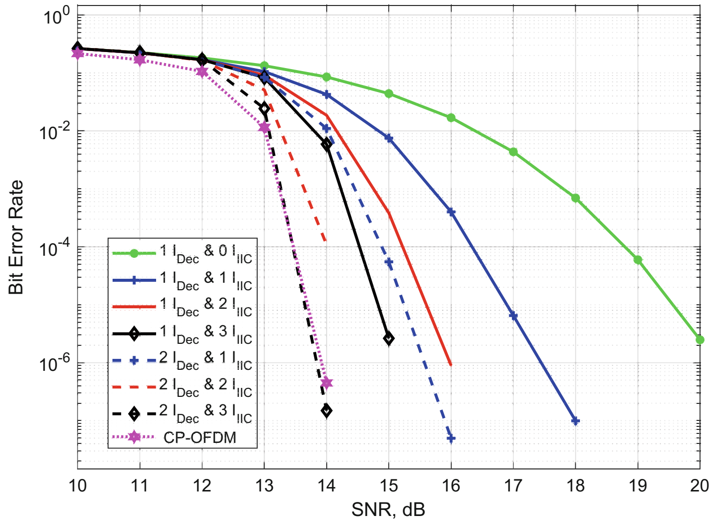


Fig. 2.14 BER performance of the proposed IIC-based BICM-ID system with M16^{r} mapped 16-QAM over EPA channel

a receiver with no inner decoder iterations. Moreover, the decoding performance can be further improved by increasing $\mathbb{I}_{Dec}$ from 1 to 2 . Compared to the benchmark CP-OFDM system, FBMC-QAM shows about 6 dB loss at 10^{-6} BER when no inner decoder iterations are applied. Setting $\mathbb{I}_{Dec} = 2$ and $\mathbb{I}_{IIC} = 3$, FBMC-QAM achieves the same BER performance as the CP-OFDM benchmark.

Similarly, the BER performances of the proposed IIC-based BICM-ID receiver over the EVA and ETU channels are shown in Figs. 2.15 and 2.16, respectively. From Fig. 2.14, it can be seen that with $\mathbb{I}_{Dec} = 1$ and $\mathbb{I}_{IIC} = 2$, a BER of 10^{-6} is obtained for the EPA channel at an SNR of 16 dB. However, an SNR of about 19.8 dB is required to achieve a similar BER performance for the EVA channel as shown in Fig. 2.15. This is mainly due to the high-frequency selectivity of the EVA channel compared to EPA, which degrades the BER performance. Notice from Figs. 2.15 and 2.16 that for $\mathbb{I}_{Dec} = 2$ and $\mathbb{I}_{IIC} = 3$, FBMC-QAM has 0.5 and 1 dB SNR losses compared to the benchmark CP-OFDM for EVA and ETU, respectively.

The results show that the proposed IIC-based LDPC BICM-ID receiver can remove the intrinsic interference in FBMC-QAM systems under different time-varying channels. With the capability to address the intrinsic interference problem in FBMC-QAM, the proposed receiver allows the use of FBMC as a viable physical layer waveform for future wireless networks due to its ultra-low OOB emission, which causes very low leakage interference between asynchronous users as well as its better spectral efficiency.

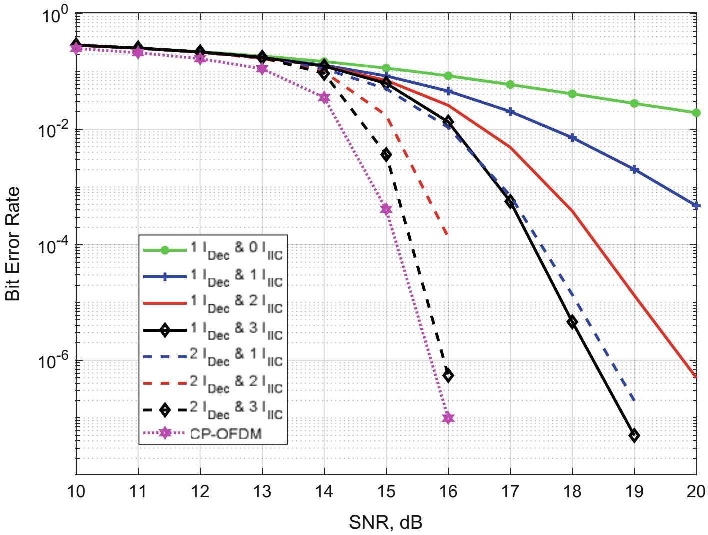


Fig. 2.15 BER performance of the proposed IIC-based BICM-ID system with M16^r mapped 16-QAM over EVA channel

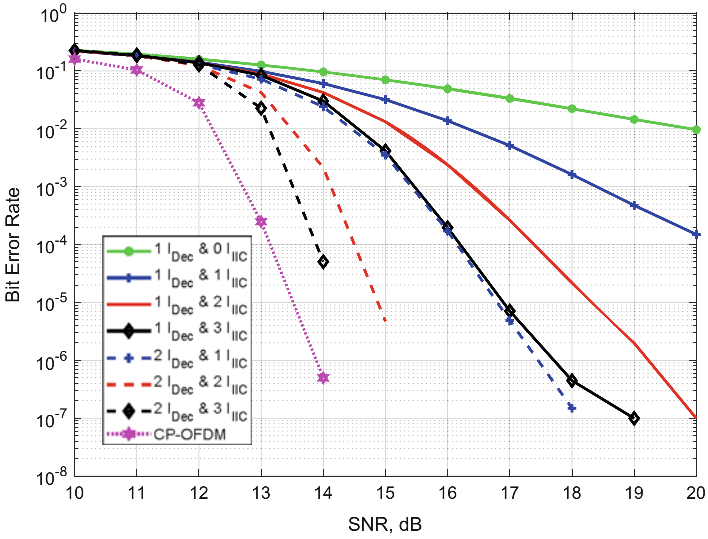


Fig. 2.16 BER performance of the proposed IIC-based BICM-ID system with M16^r mapped 16-QAM over ETU channel

2.5 Conclusion

In Sect. 2.2, we considered a cell-free MIMO system in uplink, comprising a massive number of distributed transmit and receive antennas. In our DAS, transmit and receive antennas are distributed according to homogeneous point processes (PP), and the received signals are processed jointly at a CPU. We showed analytically that the favorable propagation conditions are not satisfied in CF massive MIMO system with transmit and receive antennas in LoS motivating the use and analysis of multistage receivers. On the contrary, they hold in the case of Rayleigh fading. Then, we analyzed the performance of multistage detectors in these two scenarios and showed the relevance of their use especially in systems with antennas in LoS. Simulation results of the favorable propagation conditions and the performance of multistage detectors for finite systems validated the asymptotic analytical results.

Sect. 2.3 presents a LoS mmWave propagation model serving multiple users. In this model, it is assumed that the same total number of antenna elements is distributed among different sites positioned randomly at the sides of a square-shaped coverage area and the user equipment (UE) are also randomly distributed throughout that area. The results show significant performance enhancement with this distributed approach compared to collocated antenna array due to base station (BS) diversity. A zero-forcing (ZF) precoding scheme is applied to eliminate the interference from multiple users.

In Sect. 2.4, FBMC was investigated as an alternative to CP-OFDM for beyond 5G applications, due to its high spectral efficiency and ultra-low OOB emission, which make it suitable for asynchronous transmissions. The key drawback of FBMC systems is the high intrinsic interference caused by the loss of complex orthogonality between subcarriers. To address the interference problem, we propose and analyze the performance of an IIC-based BICM-ID receiver for FBMC systems. The results show that the proposed IIC-based LDPC BICM-ID receiver can remove the intrinsic interference in FBMC systems under time-varying channels. In summary, FBMC is a promising multicarrier waveform for asynchronous applications due to its ultra-low OOB emission, which causes very low leakage interference between asynchronous users, and the proposed iterative receiver is capable of effectively addressing the intrinsic interference problem adequately.

Acknowledgments This work was supported by H2020-MSCA-SPOTLIGHT project under Grant 722788.

References

1. Lin, Y., & Yu, W. (2014). Downlink spectral efficiency of distributed antenna systems under a stochastic model. *IEEE Transactions on Wireless Communications, 13*(12), 6891–6902.
2. Li, J., Wang, D., Zhu, P., Wang, J., & You, X. (2017). Downlink spectral efficiency of distributed massive MIMO systems with linear beamforming under pilot contamination. *IEEE Transactions on Vehicular Technology, 67*(2), 1130–1145.

3. Dai, L. (2011). A comparative study on uplink sum capacity with co-located and distributed antennas. *IEEE Journal on Selected Areas in Communications, 29*(6), 1200–1213.
4. Cottatellucci, L. (2014). Spectral efficiency of extended networks with randomly distributed transmitters and receivers, in *2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)* (pp. 673–677). Piscataway: IEEE.
5. Cottatellucci, L. (2014). Capacity per unit area of distributed antenna systems with centralized processing. In *Global Communications Conference (GLOBECOM), 2014 IEEE* (pp. 1746–1752). Piscataway: IEEE.
6. Marzetta, T. L. (2010). Noncooperative cellular wireless with unlimited numbers of base station antennas. *IEEE Transactions on Wireless Communications, 9*(11), 3590–3600 (2010)
7. Ngo, H. Q., Larsson, E. G., & Marzetta, T. L. (2014). Aspects of favorable propagation in massive MIMO. In *2014 22nd European Signal Processing Conference (EUSIPCO)* (pp. 76–80). Piscataway: IEEE.
8. Ngo, H. Q., Ashikhmin, A., Yang, H., Larsson, E. G., & Marzetta, T. L. (2015). Cell-free massive MIMO: uniformly great service for everyone. In *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)* (pp. 201–205). Piscataway: IEEE.
9. Ngo, H. Q., Ashikhmin, A., Yang, H., Larsson, E. G., & Marzetta, T. L. (2017). Cell-free massive MIMO versus small cells. *IEEE Transactions on Wireless Communications, 16*(3), 1834–1850.
10. Chen, Z., Björnson, E. (2018). Channel hardening and favorable propagation in cell-free massive MIMO with stochastic geometry. *IEEE Transactions on Communications, 66*(11), 5205–5219.
11. Gholami, R., Cottatellucci, L., & Slock, D. (2020). Channel models, favorable propagation and multistage linear detection in cell-free massive MIMO. In *2020 IEEE International Symposium on Information Theory (ISIT)* (pp. 2942–2947). Piscataway: IEEE.
12. Cottatellucci, L., & Muller, R. R. (2005). A systematic approach to multistage detectors in multipath fading channels. *IEEE Transactions on Information Theory, 51*(9), 3146–3158.
13. Moshavi, S. (1996). Multi-user detection for DS–CDMA communications. *IEEE Communications Magazine, 34*(10), 124–136.
14. Goldstein, J. S., Reed, I. S., & Scharf, L. L. (1998). A multistage representation of the Wiener filter based on orthogonal projections. *IEEE Transactions on Information Theory, 44*(7), 2943–2959.
15. Tulino, A., & Verdú, S. (2004). *Random matrix theory and wireless communications*. *Foundations and trends in communications and information theory* (vol. 1). Norwell: Now Publishers.
16. Skipetrov, S., & Goetschy, A. (2010). Eigenvalue distributions of large Euclidean random matrices for waves in random media. arXiv:1007.1379.
17. Nyberg, A. (2014). *The Laplacian spectra of random geometric graphs*, Ph.D. Dissertation.
18. Gray, R. M. (2006). Toeplitz and circulant matrices: A review. *Foundations and Trends®in Communications and Information Theory, 2*(3), 155–239.
19. Gholami, R., Cottatellucci, L., & Slock, D. (2020). Favorable propagation and linear multiuser detection for distributed antenna systems. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5190–5194). Piscataway: IEEE.
20. Cottatellucci, L., Muller, R. R., & Debbah, M. (2010). Asynchronous CDMA systems with random spreading—Part II: Design criteria. *IEEE Transactions on Information Theory, 56*(4), 1498–1520.
21. Cottatellucci, L., Müller, R. R. (2007). CDMA systems with correlated spatial diversity: A generalized resource pooling result. *IEEE Transactions on Information Theory, 53*(3), 1116–1136.
22. Bjornson, E., Larsson, E. G., & Marzetta, T. L. (2016). Massive MIMO: Ten myths and one critical question. *IEEE Communications Magazine, 54*(2), 114–123.
23. Femenias, G., & Riera-Palou, F. (2019). Cell-free millimeter-wave massive MIMO systems with limited fronthaul capacity. *IEEE Access, 7*, 44596–44612.

24. Taygur, M. M., & Eibert, T. F. (2017). Investigation of distributed and collocated base stations in a large urban massive MIMO scenario. In *European Conference on Antennas and Propagation (EUCAP)* (pp. 1577–1581)
25. Liu, Z., & Dai, L. (2014). A comparative study of downlink MIMO cellular networks with co-located and distributed base-station antennas. *IEEE Transactions on Wireless Communications, 13*(11), 6259–6274.
26. Liu, Z., & Dai, L. (2013). Asymptotic capacity analysis of downlink MIMO systems with co-located and distributed antennas. In *IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)* (pp. 1286–1290).
27. Kamga, G. N., Xia, M., & Aissa, S. (2016). Spectral-efficiency analysis of massive MIMO systems in centralized and distributed schemes. *IEEE Transactions on Communications, 64*(5), 1930–1941.
28. MacCartney, G. R., & Rappaport, T. S. (2019). Millimeter-wave base station diversity for 5G coordinated multipoint (CoMP) applications. *IEEE Transactions on Wireless Communications, 18*(7), 3395–3410.
29. Alonzo, M., Buzzi, S., Zappone, A., & D'Elia, C. (2019). Energy-efficient power control in cell-free and user-centric massive MIMO at millimeter wave. *IEEE Transactions on Green Communications and Networking, 3*(3), 651–663.
30. Alonzo, M., Buzzi, S., & Zappone, A. (2018). Energy-efficient downlink power control in mmWave cell-free and user-centric massive MIMO. In *IEEE 5G World Forum (5GWF)* (pp. 493–496).
31. Razavizadeh, S. M., Ahn, M., & Lee, I. (2014). A new enabling technology for 5G wireless networks three-dimensional beamforming. *IEEE Signal Processing Magazine, 31*(6), 94–101.
32. *5G; Study on channel model for frequencies from 0.5 to 100 GHz*, 3GPP TR 38.901 V.14.1. ed. (2017)
33. Yang, H., & Choi, S. (2013). Implementation of a zero-forcing precoding algorithm combined with adaptive beamforming based on WiMAX system. *International Journal of Antennas and Propagation, 2013*, 1–7.
34. Chen, D., Tian, Y., Qu, D., & Jiang, T. (2018). OQAM-OFDM for wireless communications in future Internet of Things: A survey on key technologies and challenges. *IEEE Internet of Things Journal, 5*, 3788–3809 (2018)
35. Bellanger, M. (2010). FBMC physical layer: A primer. PHYDYAS, Tech. Rep.
36. Sexton, C., Bodinier, Q., Farhang, A., Marchetti, N., Bader, F., & DaSilva, L. A. (2018). Enabling asynchronous machine-type D2D communication using multiple waveforms in 5G. *IEEE Internet of Things Journal, 5*(2), 1307–1322.
37. Nam, H., Choi, M., Han, S., Kim, C., Choi, S., & Hong, D. (2016). A new filter-bank multicarrier system with two prototype filters for QAM symbols transmission and reception. *IEEE Transactions on Wireless Communications, 15*(9), 5998–6009.
38. Mahama, S., Harbi, Y. J., Burr, A. G., & Grace, D. (2019). Iterative interference cancellation in FBMC-QAM systems. In *Wireless Communications and Networking Conference (WCNC)* (pp. 1–5)
39. Mahama, S., Harbi, Y. J., Burr, A. G., & Grace, D. (2020). Design and convergence analysis of an IIC-based BICM-ID receiver for FBMC-QAM systems. *IEEE Open Journal of the Communications Society, 1*, 563–577.
40. Kim, J., Park, Y., Weon, S., Jeong, J., Choi, S., & Hong, D. (2018). A new filter-bank multicarrier system: The linearly processed FBMC system. *IEEE Transactions on Wireless Communications, 17*(7), 4888–4898.
41. Harbi, Y. J., & Burr, A. G. (2016). On ISI and ICI cancellation for FBMC/OQAM system using iterative decoding and ML detection. In *Wireless Communications and Networking Conference (WCNC)* (pp. 1–6)
42. Li, Q., Zhang, J., & Epple, U. (2016). Design and EXIT chart analysis of a doubly iterative receiver for mitigating impulsive interference in OFDM systems. *IEEE Transactions on Communication, 64*(4), 1726–1736.
43. Johnson, J. S. (2009). *Iterative error correction: Turbo, low-density parity-check and repeat-accumulate codes*, C. U. Press, edn. Cambridge: Cambridge University Press.

Chapter 3
Radio Resource Management and Access Polices for B5G

Michalis Eliodorou, Tafseer Akhtar, Muhammad Tayyab, and Subin Narayanan

Abstract This chapter will address radio resource management for B5G in a bid to maximize cell capacity and energy efficiency. This needs to be done in a holistic manner that takes into account the network state parameters. In this context, this chapter will consider radio resource allocation for 5G UDNs using cooperative game theory to optimize the network's performance. In addition, the mobility which is already a major challenge will become even more demanding in future mobile networks. Important aspects of mobility are presented along with effective HO solutions being widely studied for the cell edge. Finally, the chapter provides an analysis on the random access policy that can be particularly suited for NB-IoT by optimally allocating "repetition" and "retransmission" to reduce the collision rate.

3.1 Introduction

In future fifth generation (5G) and beyond wireless technologies, ultradense networks (UDNs) will be employed to serve a massive number of devices with mobile access. Although UDNs can provide a basis for high-throughput systems, their capability is still largely potential rather than practical due to the challenging conditions of the networking environment and stringent design 5G design requirements. In particular, 5G is planned to handle a very high volume of data traffic, high peak data

M. Eliodorou
epic ltd, Nicosia, Cyprus
e-mail: michalis.eliodorou@epic.com.cy

T. Akhtar
University of Patras, Patra, Greece

M. Tayyab
Huawei Technologies, Helsinki Area, Finland
e-mail: muhammad.tayyab5@huawei.com

S. Narayanan (✉)
National and Kapodistrian University of Athens, Athens, Greece
e-mail: snarayanan@di.uoa.gr

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_3

rate, and very low latency, among others, that has spurred various new approaches and innovations into enhancing and optimizing the deployed 5G infrastructure [1], where undoubtedly any future uptake will become part a future 5G release.

One of the major challenges in ultradense networks (UDNs) is intra- and intercell interference due to the close deployment of base stations. The 5G environment is also heterogeneous where various small cells (micro, pico, and femto cell) are deployed densely to create the hyper dense network environment [2]. Game theory can be used for modelling various RRM problems; games can be of cooperative or non-cooperative nature where players can consider overall collective payoffs or individual payoff [3]. Cooperative games can effectively deal with user association, which is essential for dealing with optimization problems including intra- and intercell interference, energy minimization for IoT devices using mobile edge computing (MEC), etc. In this chapter, we aim to precisely address the RMM Challenge. Section 3.2 introduces coalitional games for characterizing the user association problem that relies on cooperation among players, followed by examples of game-theoretic algorithms that target to provide solutions for various optimization problems. In Sect. 3.3, the cooperative game approach is further extended to demonstrate how they can be utilized to design RRM for 5G and beyond networks use cases.

Another key challenge in 5G and beyond is mobility in small cell networking environments. In the design of future mobile networks, the mobility requirements are becoming more challenging, both in terms of robustness against handover (HO) failure and reducing energy consumption. Current mobility solutions in Long-Term Evolution (LTE) and New Radio (NR) come at the expense of increased signaling overheads due to measurement reports over the air interface, especially at the cell edges. In this context, Sect. 3.4 will investigate the mobility problem in mobile networks employing small technology such as 5G and how UL (Uplink) measurement-based RS (received signal) HO schemes can reduce the HO signaling overheads and power consumption in contrast to legacy downlink measurement-based approaches.

The legacy cellular networks that are typically designed for conventional human-type communications (HTC) are evolving to support machine-type communications (MTC) or the Internet of Things (IoT). According to analyst firm Gartner, the number of IoT devices will reach up to 20.4 billion in 2020, which will lead to a scenario called massive IoT system (MIoT). The MIoT refers to the billions of mobile or stationary devices that communicate with each other or to a centralized system through some wireless technologies. The technological solutions to support MIoT can be broadly classified into two: (i) unlicensed spectrum technologies and (ii) licensed spectrum technologies or cellular IoT (CIoT) technologies. The licensed spectrum technologies can support a wide range of IoT/MTC use cases with better device management and enhancement in service provisioning, compared to unlicensed technologies. Resource optimization in MTC is gaining increased attention, being responsible for attaching the MTC devices to the network, that is, creating huge challenges due to the massive amount of device that will create excessive collisions at the RAN. In Sect. 3.5, we analyze how the NB-IoT (CIoT technology) RAN parameters “repetition” and “retransmission” can be optimally allocated to reduce the collision rate.

3.2 Coalition Games for Resource Allocation in 5G and Beyond Networks

In future fifth-generation (5G) and beyond wireless technologies, ultradense networks (UDNs) will be employed to serve a massive number of devices with mobile access. One of the major challenges in UDNs is user association, which is essential for dealing with intra- and inter-cell interference. In this chapter, two user association problems are solved via a game theoretical approach. Specifically, coalition game algorithms are formulated which exploit the cooperation among the players of the game to maximize the overall sum rate in the first system model and minimize the energy consumption on the second. The proposed algorithms, using various techniques including non-orthogonal multiple access (NOMA), show that the overall performance of the system can be improved significantly, providing near-optimal solutions, while keeping the complexity low.

3.2.1 Introduction

Game theory could be described as the study of mathematical models related to rational decision-makers in situations involving conflict of interest or cooperation. Game theory has been studied for a variety of applications in multiple disciplines such as economics, political sciences, philosophy, and, more recently, engineering [4]. Non-cooperative game theory studies the decision-making of a single player. In contrast, cooperative game theory such as coalitional game theory and matching games, being the topic of this section, focuses on what groups of players can achieve together rather than on what individual players can do alone [5]. More analytically, coalitional game theory seeks for an optimal coalition structure of the players to maximize the value of each coalition. Also, coalitional games are divided into two types: with transferable utility (TU) and with non-transferable utility (NTU). In TU games, a group of players is associated with a fixed number which can be distributed in any way among the group members defining each player’s payoff value. In NTU games, which is the type of game applied in the following network models, the sum of the payoff value of the members of each coalition is not fixed; hence, the value of a coalition depends on the selected structure [6]. In what follows, two coalition games are studied for two different network models where the coalition structure set to be optimized is applied on a set of users.

3.2.2 Game Theory for User Association

Coalition games can solve optimization problems by reaching a final state that benefits all of the players. In general, coalition games have been proven to be very efficient in many multiplayer scenarios [7–9]. A user association problem (i.e., associating users to BSs) can be defined as a coalition game $(\mathcal{K}, \mathcal{X}, U)$ with a non-transferable utility U [10], where $\mathcal{K}$is the player set consisting of the users, set $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$ is the set consisting of the vectors indicating the user-BS association, and U is the achievable data rate of all the players for a given association $\mathcal{X}$. A partition of the players, among the available BSs, is denoted by $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$, where S_m is the coalition consisting of the users associated with the m-th SBS. For each coalition, $S_m \in \mathcal{S}, m \in \mathcal{M}$, the conditions $S_m \cap S_l = \varnothing$, $\forall\ m \neq n$ and $\bigcup_{m=1}^{M} S_m = \mathcal{K}$ are both satisfied. To formulate the algorithms presented in this chapter, the following three definitions are introduced:

Definition 1 (Preference condition) For any user $k \in \mathcal{K}$, we use the symbol $\succ_k$ to denote its preference between two different partition sets $\mathcal{S}$ and $\mathcal{S}'$.

The binary decision of a user k depends on whether the utility value of the game with the new partition will increase, i.e.,

$$\mathcal{S}' \succ_k \mathcal{S} \iff U\left(\mathcal{S}'\right) > U\left(\mathcal{S}\right), \tag{3.1}$$

where the utility value $U\left(\mathcal{S}\right)$ is the overall sum rate given a partition set $\mathcal{S}$.

Definition 2 (Split and merge operation) Given two different partition sets $\mathcal{S}$ and $\mathcal{S}'$, a user $k \in \mathcal{K}$ decides to leave its current coalition $S_m \in \mathcal{S}$, to join another one $S_{m'} \in \mathcal{S}$, where $m, m' \in \mathcal{M}, m \neq m'$,if and only if its preference condition (Definition 1) is satisfied. The split and merge operation can be written as

$$\{S_m, S_{m'}\} \rightarrow \{S_m \setminus \{k\}, S_{m'} \cup \{k\}\}. \tag{3.2}$$

Note that for the above operation, the user k joins the other partition if there is room for an additional player. Otherwise, according to predefined limit, a user k in coalition $S_{m'}$ is selected at random and swapped with k based on the following definition.

Definition 3 (Swap operation) Two users are said to be swapped, if and only if the preference condition (Definition 1) is satisfied for both. Then, the partitions are updated accordingly as

$$\{S_m, S_{m'}\} \rightarrow \left\{S_m \setminus \{k\} \cup \left\{k'\right\}, S_{m'} \setminus \left\{k'\right\} \cup \{k\}\right\}. \tag{3.3}$$

Initially all players (i.e., users) can be allocated randomly to the available resources or based on a conventional method. At each iteration, a user associated with a coalition, say m, is randomly selected. By selecting a different coalition $m^{'}$,

$m \neq m^{'}$, we check if the preference condition is satisfied. In the case where this is true, operations Split and merge or Swap are applied accordingly.

In what follows, a proof is provided, showing that any game theoretic algorithm is able to converge at a final state which is D_p stable.

Convergence: Starting at any initial combination, the user association game of Algorithm 1 is guaranteed to converge at a final state.

Proof In order to increase the game utility U, the users perform either one of the operations described above, which results in a constantly modifying partition set. Consider two successive iterations i and $i + 1$, and assume that partition $S_{i + 1}$ was formed from S_i, after an operation is applied. Both operations take place if and only if the game utility U is strictly increased. This can be written as

$$S_i \rightarrow S_{i+1} \iff U\left(S_i\right) < U\left(S_{i+1}\right) \tag{3.4}$$

Therefore, the game utility value is always increasing, that is,

$$S_{ini} \rightarrow S_1 \rightarrow S_2 \rightarrow \cdots \rightarrow S_{fin} \tag{3.5}$$

where S_{ini} and S_{fin} is the initial and final partition set of the game, respectively. Hence, the sum rate is guaranteed to improve at each new partition set. Sine the number of players is finite and the number of actions of each player is finite as well means that the number of partition sets is also finite and is based on the Bell number [11]. Therefore, the sequence of the partition sets formulated by the algorithm is guaranteed to converge to the final state S_{fin}.

D_p stability: The final partition set S_{fin} is D_p stable.

Proof A partition S is D_p stable, if for any other partition $\mathcal{S}^{'} \neq \mathcal{S}$, $U\left(\mathcal{S}\right) > U\left(\mathcal{S}'\right)$. Suppose the final partition S_{fin} of Algorithm 1 is not D_p stable. Then, there must exist a user $k \in \mathcal{K}$ that prefers to leave its current coalition and join another. This will form a new partition S_{tmp}, where $S_{tmp} \succ_k S_{fin}$ which contradicts the fact that S_{fin} is the final partition. Therefore, the final partition of Algorithm 1 is D_p stable.

3.2.3 User Association Coalition Games for Sum-Rate Maximization

In [12], a user association problem in a cellular downlink network is presented, and a solution using coalition game theory is formulated. Specifically, an algorithm is presented where both Zero-Forcing (ZF) and regularized ZF (RZF) are applied at the small base stations (SBSs). Simulation results show that the proposed algorithms can significantly outperform the conventional minimum-distance association scheme in terms of the network's sum rate. The benefits of NOMA are also considered to increase the number of users being served, and it is shown that the joint consideration of ZF with NOMA provides substantial gains to the sum-rate performance. The proposed algorithms via coalition games are of great importance for future networks, as they are of low complexity and can achieve near-optimal solutions.

3.2.3.1 System Model and Throughput Enhancement Techniques

A downlink cellular network is considered focusing in a circular area in which M SBSs and K users are randomly located, with $K \geq M$. We denote by $\mathcal{M} = \{1, 2, \ldots, M\}$ and $\mathcal{K} = \{1, 2, \ldots, K\}$ the sets of the SBSs and the users, respectively. Each SBS transmits with power P_t and is equipped with N antennas. The set of users associated with the j-th SBS is indicated with $\mathcal{K}_j$, and the cardinality of the set is denoted by K_j, where $K_j \leq N_{RF}$ and $\sum_{j=1}^{M} K_j = K$. All channel coefficients are modeled as block Rayleigh fading with unit variance, i.e., $h_{k,j} \sim$ (0,1), and the SBS are assumed to have full channel state information (CSI). The path-loss model is considered to be $d_{k,j}^{-\alpha}$, where $d_{k,j}$ is the distance between user k and SBS j with α being the path-loss exponent. All links contain additive white Gaussian noise with variance σ^2.

Precoding Schemes (ZF and RZF)

Precoding techniques used in MU-MIMO antennas can enhance throughput. In this work, conventional methods, i.e., ZF or RZF [13], are applied where up to N_{RF} users can be served by each BS simultaneously, where N_{RF} is the number of available radio-frequency (RF) chains.

Non-orthogonal Multiple Access

NOMA is a multiple access scheme which allows additional users to be served using pairs [8]. Each pair requires a strong and a weak user using the same resources apart from the power, which is separated between the two users [14], i.e., the weak user requires more power since its channel conditions are poorer. The strong user can perform a successive interference cancellation (SIC) technique [8], while the weak user treats the strong user's signal as interference. We denote by p_w and p_s the power allocation coefficients of the weak and strong user, respectively, with $p_w > p_s$ and $p_w + p_s = 1$. In the case where ZF and NOMA is jointly applied, we consider $K^{'}$ additional users which are no longer served with the ZF scheme.

Sum-Rate Maximization Problem

The user association problem is formulated aiming to maximize the overall DL rate of all small cells. The data rate of user k served by the m-th SBS is $R_{k,\,m} = B\log_2(1 + SINR_{k,\,m})$,where B is the available bandwidth. The association of each user is critical as it affects the inter-cell interference caused to the rest of the network's users. Therefore, the overall data rate is highly depended on the user selection. The user association problem based on the utility is formulated as follows:

$$\max_{\left\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\right\}} \sum_{m=1}^{M} x_{k,m} R_{k,m}, \tag{3.6}$$

s.t.

$$x_{k,m} \in \{0, 1\}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \tag{3.6a}$$

$$\sum_{m=1}^{M} x_{k,m} = 1, \forall m \in \mathcal{M}, \tag{3.6b}$$

$$\sum_{i=1}^{K} x_{i,m} \leq N_{RF}, \forall m \in \mathcal{M}, \tag{3.6c}$$

where $x_{k,\,m}$ is a binary value denoting whether or not the k-th user is associated with the m-th SBS and $\boldsymbol{x}_i = \left\{x_{k,i}\right\}, k \in \mathcal{K}$, is the set of cardinality K, defining each user's association with the i-th SBS. The constraint in (3.6b) ensures that each user is associated with only one SBS. The last constraint in (3.6c) guarantees that the number of users associated with a SBS does not exceed the number of available RF chains, N_{RF}. The formulated problem is non-convex and difficult to transform into a convex problem [8]. However, by treating it as a coalition game, the problem can be solved.

3.2.3.2 Game Theoretic Algorithms for UE-to-BS Association

The first game-theoretic algorithm uses only the ZF/RZF schemes. Initially, all users are allocated randomly to the available SBSs. At each iteration, a user associated with SBS, say m, is randomly selected. By selecting a different SBS $m^{'}$, $m \neq m^{'}$, thus selecting another coalition, we check if the preference condition is satisfied. In the case where this is true, operations Split and Merge or Swap are applied accordingly. The pseudocode of the proposed algorithm is provided in Fig. 3.1.

Fig. 3.1 Pseudocode for the coalition game algorithm with ZF/RZF

```
Algorithm 1 Coalition game with ZF/RZF
Initializing users with a random parition $\mathcal{S}_{ini}$
Denote current partition $\mathcal{S}_c \leftarrow \mathcal{S}_{ini}$
repeat
    Randomly select a user $k$ of coalition $S_m \in \mathcal{S}_c$
    Randomly select a user $k'$ of coalition $S_{m'} \in \mathcal{S}_c$
    if $|S_{m'}| = N_{\mathrm{RF}}$ then
        Assume $\mathcal{S}_{tmp} \leftarrow$ swap user $k$ with user $k'$
        if $\mathcal{S}_{tmp} \succ_k \mathcal{S}_c$ then
            $\mathcal{S}_c \leftarrow \{\mathcal{S}_c \setminus \{S_m, S_{m'}\}\} \cup \{S_m \setminus \{k\} \cup \{k'\},$
                                        $S_{m'} \setminus \{k'\} \cup \{k\}\}$
    else
        Assume $\mathcal{S}_{tmp} \leftarrow$ user $k$ joins $S_{m'}$
        if $\mathcal{S}_{tmp} \succ_k \mathcal{S}_c$ then
            $\mathcal{S}_c \leftarrow \{\mathcal{S}_c \setminus \{S_m, S_{m'}\}\} \cup \{S_m \setminus \{k\}, S_{m'} \cup \{k\}\}$
until
```

Complexity: Each iteration executes K number of computational operations, to calculate the data rate of each user. The complexity of the algorithm is (CK), where C is the number of iterations. The complexity of an exhaustive search is (C^K) which is significantly higher [7].

The NOMA+ZF case is considered as well, where a dominant and a weak user must be paired, thus increasing the number of users served by the network. Algorithm 1 is executed first, and $K^{'}$ additional users participate in a second game to be paired. Again, starting from a random pair allocation, $\mathcal{S}_{ini}^{NOMA}$, in a similar manner, a user $k \in K^{'}$ is selected, and a split and merge or a swap operation is accepted only when the sum rate of the pair is increased. Note that the SIC condition must be satisfied to ensure that the pair can apply NOMA. Like the previous algorithm: convergence and stability are satisfied. In Fig. 3.2, the pseudocode for Algorithm 2 is presented.

The Simulated Annealing algorithm (SAA) is developed as a performance metric allowing us to approximate the global optimum solution [14]. This is achieved by allowing the algorithm to accept a new partition set $\mathcal{S}_{i+1}$, even when the utility value of the new partition, i.e., $U\left(\mathcal{S}_{i+1}\right)$, is lower than the current one. To do so, we use a probabilistic approach, the Metropolis-Hastings algorithm [15]. The probability of a partition $\mathcal{S}_{i+1}$ being accepted, despite $U\left(\mathcal{S}_i\right) > U\left(\mathcal{S}_{i+1}\right)$, is decided by the following probability:

$$P_{SAA} = \tau \exp\left(\frac{U\left(\mathcal{S}_{i+1}\right) - U\left(\mathcal{S}_{max}\right)}{U\left(\mathcal{S}_{max}\right)}\right), \tag{3.7}$$

where τ is the temperature of the SAA and $\mathcal{S}_{max}$ is the maximum value up to that point. By using a large number of iterations, we can ensure that $\mathcal{S}_{max}$ approximates the global optimum.

Algorithm 2 Coalition game algorithm with ZF and NOMA

```
**Algorithm 1** is executed
Randomly pair $K'$ additional users with the $K$ users of the first game, i.e. $S_{ini}^{\text{NOMA}}$
Denote current partition $\mathcal{S}_c^{\text{NOMA}} \leftarrow \mathcal{S}_{ini}^{\text{NOMA}}$
**repeat**
    Select a NOMA user $k$ of coaltion $S_m \in \mathcal{S}_c^{\text{NOMA}}$
    **if** $|S_{m'}| = N_{\text{RF}}$ **then**
        Select a user $k'$ of coalition $S_{m'} \in \mathcal{S}_c^{\text{NOMA}}$
        $\mathcal{S}_{tmp}^{\text{NOMA}} \leftarrow$ swap case of NOMA users $k$ and $k'$
        **if** $\mathcal{S}_{tmp}^{\text{NOMA}} \succ_k \mathcal{S}_c^{\text{NOMA}}$ **then**
            $\mathcal{S}_c^{\text{NOMA}} \leftarrow \{\mathcal{S}_c^{\text{NOMA}} \setminus S_m, S_{m'}\} \cup \{S_m \setminus \{k\} \cup$
                        $\{k'\},\ S_{m'} \setminus \{k'\} \cup \{k\}\}$
    **else**
        $\mathcal{S}_{tmp}^{\text{NOMA}} \leftarrow$NOMA user k pairs with a user of $S_{m'}$
        **if** $\mathcal{S}_{tmp}^{\text{NOMA}} \succ_k \mathcal{S}_c$ **then**
            $\mathcal{S}_c^{\text{NOMA}} \leftarrow \{\mathcal{S}_c^{\text{NOMA}} \setminus S_m, S_{m'}\} \cup \{S_m \setminus \{k\}, S_{m'} \cup k\}$
**until**
```

Fig. 3.2 Pseudocode for the coalition game algorithm with ZF + NOMA

3.2.3.3 Numerical Results

Numerical results demonstrate the performance of the proposed coalition game algorithms. The following parameters were used: $M = 5$, $N = 6$, $N_{RF} = 6$, $K = 60$, $\sigma^2 = -90$ dBm, $R_D = 50$ m, $B = 20$ MHz, and $\alpha = 2$ for the LOS case and $\alpha = 4$ for the NLOS and $\tau = 0.2$. The power coefficients of each NOMA pair are $p_w = 0.7$ and $p_s = 0.3$ for the weak and the strong user, respectively.

Figure 3.3a shows the system sum rate achieved by the proposed schemes over the number of iterations. The minimum-distance-based user association (MDUA) is also included. The sum rate of both algorithms at iteration 0 is lower than the MDUA scheme, but as the iteration number increases, we reach a final state where the sum rate is significantly higher. It is shown that 1500 iterations are sufficient for the game to converge. Algorithm 2 serves K' additional users; hence, the data rate of all the users has a higher value. To ensure that the SAA algorithm approximates the global optimum, 10^5 iterations were used. It is observed that the final values $U(\mathcal{S}_{fin})$ achieved by Algorithm 1 and Algorithm 2 successfully provide a near-optimal solution. Figure 3.3b presents the converged sum-rate value achieved by the proposed schemes along with the MDUA scheme for three different number of users (20, 40, and 60). As we can see, the algorithms outperform the MDUA scheme, regardless of the number of users. RZF achieves a slightly higher sum rate compared to ZF, but ZF + NOMA can outperform both precoding schemes when employed ($K > 30$). In the case of 60 users, a performance downgrade is observed for Algorithm 2, indicating the effect of inter-cell interference caused by the additional K′ users.

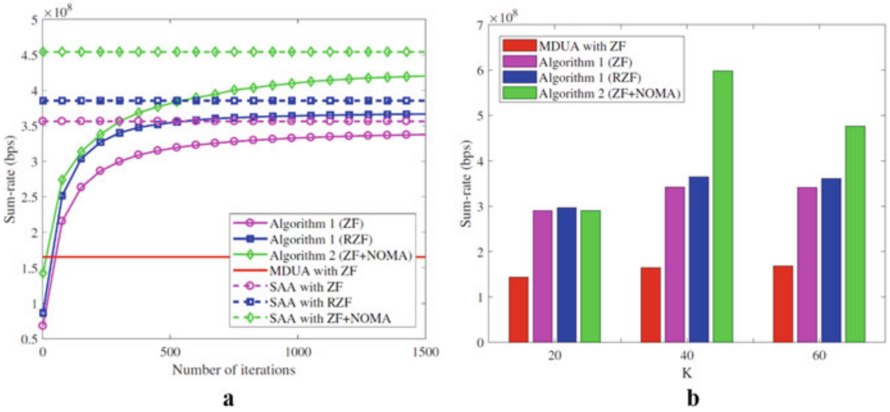


Fig. 3.3 (**a**) Sum rate versus the number of iterations, (**b**) sum-rate at iteration 1500 versus the number of users

3.2.4 User Association Coalition Games for Energy Minimization in MEC Networks

In [16], the user association problem is investigated for a mobile edge computing (MEC) non-orthogonal multiple access network. Multiple users can access the MEC server to offload data simultaneously. The fact that resources are shared among the users can potentially impact the required transmit power for offloading, hence increasing the energy consumption of the devices. Aiming to minimize the total energy consumption of the network's devices, we formulate an optimization problem where user association, optimal power allocation, data rate, and offloaded data are jointly considered. Two coalition game algorithms are proposed to efficiently reduce the total energy consumption.

3.2.4.1 System Model and MEC

In this scenario, a SBS with N number of subcarriers is located at the center of a circular area with radius R_D and K users randomly located, with $K \geq N$. We denote by $N = \{1, 2, \ldots, N\}$ the set of subcarriers and $K = \{1, 2, \ldots, K\}$ the set of users. Each user must execute a task with L input bits and is able to offload all or part of the data by utilizing the MEC scheme. Each subcarrier can receive data from up to N_{RF} users at the same time, which is the number of available radio frequency (RF) chains.

Mobile Edge Computing (MEC) with Partial Offloading

The SBS is equipped with a MEC server to help users execute their computational tasks within a time slot of duration T. We denote by L the overall bits of the task. The process of remote execution is partitioned in three phases: T_{UL}, users upload the data to the SBS; T_{EX}, the processing of the data takes place at the MEC server; and T_{DL} for the downlink transmission of the final result back to the user. Considering the advanced resources of the MEC server and considerably low data of the result, we assume that the time for the execution stage and the final stage is negligible, i.e., $T_{EX} \approx 0$ and $T_{DL} \approx 0$ [17]. The time block can be written as

$$\mathrm{T} = T_{UL} + T_{EX} + T_{DL} \approx T_{\mathrm{UL}} \tag{3.8}$$

The rate (bits/sec/Hz) for the offloaded data of each user can be expressed as $r_{k,\,n} = \log_2(1 + \mathit{SINR}_{k,\,n})$. The capacity region $C(\boldsymbol{p})$ of the uplink channel is characterized by the set of all rates $\left(r_1, \ldots, r_{K_n}\right)$ [17], satisfying the conditions of the polymatroid, i.e.,

$$C\left(\boldsymbol{p}\right) = \left\{ \boldsymbol{r} \in \mathbb{R}^{K \times 1} : \sum_{k \in \mathcal{J}} r_k \leq \log_2 \left(1 + \sum_{k \in \mathcal{J}} p_k |h_k|^2 \right), \forall \mathcal{J} \subseteq \mathcal{K} \right\} \tag{3.9}$$

where $\boldsymbol{p}$ is the power allocation vector $\left[p_1, \ldots, p_{K_n}\right]$. The selection of $\boldsymbol{p}$ must consider the constraints set by the above capacity region.

The total energy consumption for task execution while using partial offloading consists of the necessary energy to transmit the offloaded data the local energy consumption. More details about it can be found in [16].

Energy Minimization Problem

The association between user and subcarrier is critical since it affects the achievable data rate as shown by the capacity region. The user association problem is formulated jointly with the transmit power allocation $\boldsymbol{p}$, the achievable data rate $\boldsymbol{r}$, and the offloaded data as

$$\min_{\boldsymbol{r}, \boldsymbol{p}, \boldsymbol{l}, \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}} \sum_{n=1}^{N} x_{k,n} \left(E_k^{\mathrm{tx}} + E_k^{\mathrm{loc}} \right), \tag{3.10}$$

s.t.

$$x_{k,n} \in \{0, 1\}, \forall k \in \mathcal{K}, \forall n \in \mathcal{N} \tag{3.10a}$$

$$\sum_{n=1}^{N} x_{k,n} = 1, \quad \forall n \in \mathcal{N}, \tag{3.10b}$$

$$\sum_{i=1}^{K} x_{i,n} \leq N_{RF}, \quad \forall n \in \mathcal{N}, \tag{3.10c}$$

$$r \in C\left(\boldsymbol{p}\right) \; with \; r_k \geq \frac{L - l_k}{B\,T}, \tag{3.10d}$$

$$0 \leq p_k \leq P_{\max}, \forall k \in \mathcal{K}, \tag{3.10e}$$

$$0 \leq l_k \leq L, \tag{3.10f}$$

where $\boldsymbol{x}_n = \{x_{k,\,n}\}$ defines the set of users associated with the n-th subcarrier and $x_{k,\,n}$ is a binary value denoting whether or not the k-th user is associated with the n-th subcarrier. The constraint of (3.10b) ensures that each user is associated with only one subcarrier. The third constraint, (3.10c), guarantees that the number of users associated with a subcarrier does not exceed the number of available RF chains N_{RF}. Equation 3.10d ensures that the uplink rates lie within the capacity region $C(\boldsymbol{p})$. Finally, the last two constraints ensure the values for p_k and l_k are non-negative and within a permitted maximum value which are denoted as $P_{\max}$ and L, respectively. In case where the fourth constraint cannot be satisfied, we assume that the entire task is executed locally.

The formulated problem is non-convex and difficult to transform into a convex problem. However, as described in [17], the sub-problem of the last three constraints is convex and can be solved with numerical tools such as CVX [18]. Treating the general formulated problem as a game, we can reach to a final solution.

3.2.4.2 Coalition Game-Based Algorithm Formulation for Energy Minimization

In the first coalition game algorithm, all users are initially allocated randomly to the available subcarriers. At each iteration, a user associated with subcarrier, say n, is again randomly selected. By selecting a different subcarrier $n^{'}$, $n \neq n^{'}$, thus selecting another coalition, we check if the above definitions are satisfied. In the case where this is true, split and merge or swap operations are applied accordingly. The pseudocode of the proposed algorithm is provided in Fig. 3.4.

In the second algorithm, for each subcarrier n, $n \in N$, one user is selected based on the best channel conditions and distance. The selected user is called coalition head (CH), and all remaining users are called coalition members (CMs). The CHs

Fig. 3.4 Pseudocode for Algorithm 1: Random Coalition Game Algorithm

```
Algorithm 1 Random Coalition Game Algorithm
Initializing users with a random parition $\mathcal{S}_{ini}$
Denote current partition $\mathcal{S}_c \leftarrow \mathcal{S}_{ini}$
repeat
    Randomly select a user $k$ of coalition $S_n \in \mathcal{S}_c$
    Randomly select a user $k'$ of coalition $S_{n'} \in \mathcal{S}_c$
    if $|S_{n'}| = N_{\mathrm{RF}}$ then
        Assume $\mathcal{S}_{tmp} \leftarrow$ swap user $k$ with user $k'$
        if $\mathcal{S}_{tmp} \succ_k \mathcal{S}_c$ then
            User $k$ leaves $S_n$ and joins $S_{n'}$
            User $k'$ leaves $S_{n'}$ and joins $S_n$
            Update current partition:
            $\mathcal{S}_c \leftarrow \{\mathcal{S}_c \setminus \{S_n, S_{n'}\}\} \cup \{S_n \setminus \{k\} \cup \{k'\},$
                $S_{n'} \setminus \{k'\} \cup \{k\}\}$
    else
        Assume $\mathcal{S}_{tmp} \leftarrow$ user $k$ joins $S_{n'}$
        if $\mathcal{S}_{tmp} \succ_k \mathcal{S}_c$ then
            User $k$ leaves $S_n$ and joins $S_{n'}$
            Update current partition:
            $\mathcal{S}_c \leftarrow \{\mathcal{S}_c \setminus \{S_n, S_{n'}\}\} \cup \{S_n \setminus \{k\}, S_{n'} \cup \{k\}\}$
until
```

make proposals sequentially and invite the CMs to join their coalitions. In each iteration, the answer of any CM is to either accept or reject the CH's proposal based on the sum of the two coalition utility values. The pseudocode is shown in Fig. 3.5. Convergence and stability are satisfied; therefore, a final partition will be reached within a limited number of iterations converging to a final solution.

Numerical Results

Numerical results are presented in Fig. 3.6. The following parameters were used: $N = 4$, $N_{RF} = 4$, $K = 12$, $\sigma^2 = -90$ dBm, $\alpha = 3$, $R_D = 60$ m, $B = 200$ KHz, $T = 1$ sec, and the overall number of bits of the task is $L = 200$ bits.

Figure 3.6a shows the total energy consumption of the algorithms for partial and full offloading. For the case of full offloading, the appropriate adjustments were made [16]. Local computation is also included for comparison. As it can be observed, partial offloading provides significant improvement over full offloading and local computation. As the number of iterations increases, the user association changes reaching a final state where both algorithms converge. Algorithm 1 can provide a better outcome but requires more iterations than Algorithm 2 which makes the sequential algorithm better for time-critical communication scenarios with mobility. In Fig. 3.6b, we present the total energy consumption achieved by Algorithm 1 at iteration 100 versus the distance for $K = 8$ users. We point out that the performance of full offloading degrades exponentially with distance. On the

Fig. 3.5 Pseudocode for Algorithm 2: Sequential Coalition Game Algorithm

Algorithm 2 Sequential Game Algorithm

```
Initialization of $S_{ini}$:
    All CHs are associated with different subcarriers
    All CMs are randomly associated with the subcarriers
Denote current partition $\mathcal{S}_c \leftarrow \mathcal{S}_{ini}$
repeat
    For any CH $i$ of cluster $S_i$, $i \in \{1, 2, \ldots, N\}$, user
    $i$ makes a new proposal $\sigma_i$ to all CMs sequentially
    For any CM $k$, $k \in S_j, j \neq i$:
    if $|S_i| = N_{\mathrm{RF}}$ then
        Select a CM $k'$ of $S_i \in \mathcal{S}_c$ to investigate swap
        if utilities of $S_i$ and $S_j$ are increased then
            The proposal is accepted by the CMs $k$ and $k'$
        else
            The proposal is declined (no operation)
    else
        if utilities of $S_i$ and $S_j$ are increased then
            The proposal is accepted by the CM $k$
        else
            The proposal is declined (no operation)
until
```

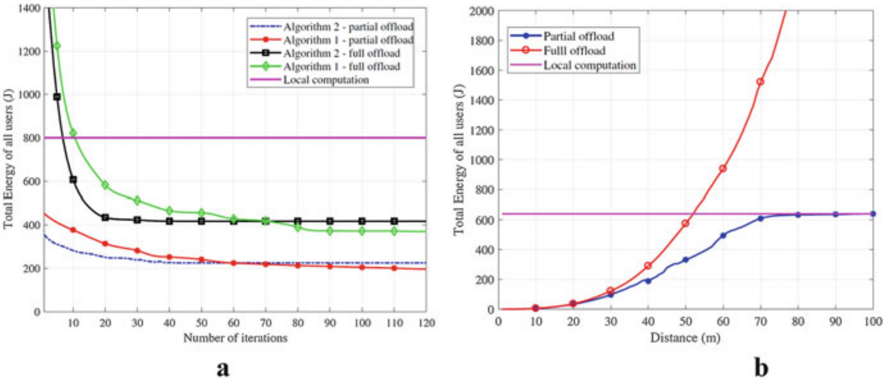


Fig. 3.6 (**a**) Sum rate versus the number of iterations, (**b**) sum rate at iteration 1500 versus the number of users

other hand, the partial offloading scheme used in our work not only will choose the best option between the two conventional schemes (local computation and full offload) but can combine them achieving a significantly higher performance gain (30–70 m).

3.2.5 Conclusion

In conclusion, we have shown that although Cloud-RANs with dense deployment can be very effective for 5G and beyond networks, interference can significantly limit their potential gain. However, user association schemes can deal with challenges emerging from UDN deployment effectively. In this section, a game-theoretic approach is applied ensuring the benefits of digital precoding techniques and NOMA, which are both susceptible to user association, and the results verify that a near-optimal solution can be achieved where the overall sum rate has significantly improved compared to conventional schemes. Another formulation is presented aiming to optimize the energy consumption of the devices for a MEC partial offloading scheme. The impact of user association and its significance is again demonstrated, and simulation results show that the two proposed algorithms can successfully converge at a final state where the total energy consumption has been reduced compared to conventional methods. Sequential swapping can reach its final state faster than random swapping. However, random selection provides higher performance gain, reducing the energy consumption even further. The results verify that coalition games are generally useful for optimization problems, and the four game theoretic algorithms can provide a solution with low complexity.

3.3 Game-Based RRM for Coordinated Multipoint Transmission

Fifth-generation (5G) mobile communication architecture and requirements demand drastic change in the existing communication. Every existing communication technological advancements and breakthroughs will be needed for efficient implementation and delivery of 5G. Tremendously high volume of data traffic, high data rate, and very low latency are just a few of the key requirements of 5G. Various new approaches and innovations in existing techniques are required to satisfy the demands of 5G system [1]. The nature of 5G, however, is not exclusive but inclusive, and previous generation will coexist with the new. However, previous generation devices, approaches, and architecture will require customization and innovation to effectively work along with new generation [19].

Radio resource management (RRM) in 5G will be more challenging than the previous generations due to its architecture and standard requirements. Intelligent and robust RRM technique is essential to handle the issues of resource management. RRM is a complex and challenging problem because it requires solution of multiple issues such as interference management, spectrum utilization, fairness, quality of service (QoS) requirement, etc. An efficient RRM approach must consider various issues simultaneously without compromising the other key aspect of network requirement [20].

In a hyper-dense network interference and spectrum utilization are two major issues, as the distribution of transmitter and receivers (base stations and users) is random and highly dense. Co-tier interference due to close deployment of base stations is very common in hyper-dense environment where user's quality of experience suffers because of nearby interfering base stations. 5G environment is also heterogeneous where various small cells (micro, pico, and femto cell) are deployed densely to create the hyper-dense network environment [2].

Heterogeneous network environment suffers from cross-tier interference as well, which make 5G environment even more challenging for effective RRM technique design. Coordination and cooperation can be a key technique to solve the complex issues of RRM in 5G environment. A cooperative network is preferred over non-cooperative network where key information exchanges can help the overall network to handle various RRM issues. Cooperation can happen at various levels (i.e., between base stations, users, users and base station) although cooperation between base stations can play an important role for designing an efficient RRM technique, but cooperation will always come with the cost due to additional signaling and information exchange.

Game theory is an effective mathematical approach for solving the complex modern problems which involve payoff and cost according to players actions. Various issues and problems can be modeled into the game with different utility functions that provide the set of payoffs while considering the respective cost involved with these actions. Game theory can be used for modeling various RRM problems, and games can be of cooperative or non-cooperative nature where players can consider overall collective payoffs or individual payoff [3]. A cooperative game approach for designing the RRM technique can be used for solving various issues in 5G and beyond.

The capacity requirements can be met by the network densification which will create a dense heterogeneous and multi-tier network consisting of coexisting macro cells and small cells which will also improve the spectral efficiency and the coverage of the network [21]. This densification can create potential problems in terms of inter-cell interference. One possible solution is the implementation of the coordinated multipoint (CoMP) technique [22]. CoMP consists of different modes of operation ranging from interference avoidance mode (coordinated scheduling and beamforming) to the more complex diversity gain mode, where the same data is transmitted from multiple cell sites, called joint transmission (JT). CoMP C-RAN is also considered a key enabler [23] for the deployment of small cells by providing efficient base band processing over the cloud and has attracted intense research interests from both industry and academia [24, 25]. Figure 3.7 shows the 5G CoMP architecture for hyper-dense and heterogeneous environment with various interfaces and connections (X2 and next-generation layer 2/3 (NG2/3) links). The sections are organized as follows: Sect. 3.3.1 briefly describes CoMP and the various applications, while Sect. 3.3.2 presents RRM assuming underlying CoMP as a technological enabler and existing works. Section 3.3.3 discusses game theory in synergy with RRM and how game theory schemes can play a key role in designing efficient RRM. Section 3.3.4 presents cooperative coalitional games with CoMP

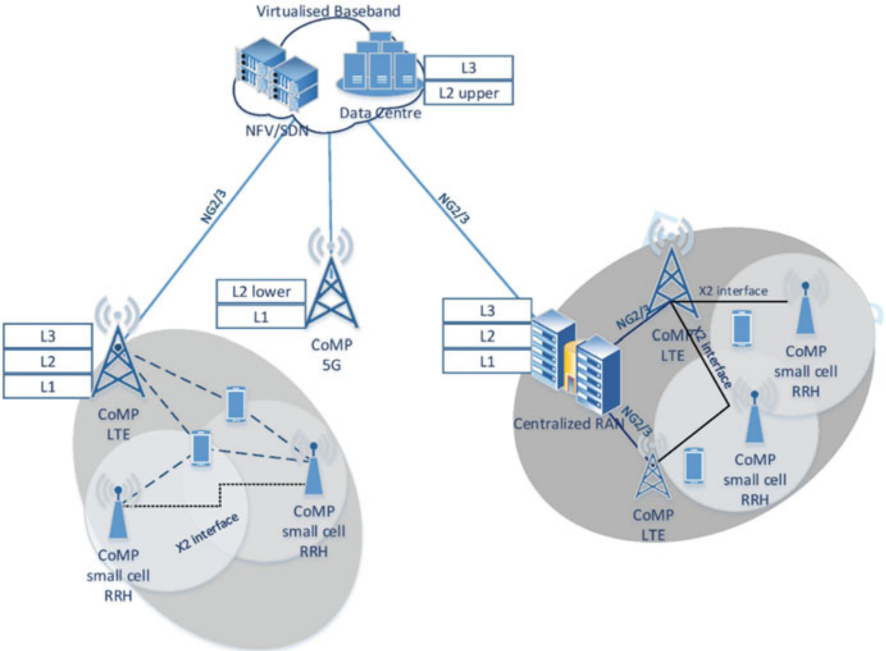


Fig. 3.7 Heterogeneous 5G architecture with its various components [26]

based on existing works, while Sect. 3.3.5 describes the proposed work along with results shown in Sect. 3.3.6, and finally, Sect. 3.3.7 concludes this part of chapter.

3.3.1 Coordinated Multipoint (CoMP) Transmission

CoMP enables inter-cell interference to be utilized as a useful signal; hence, users at the cell edge receive higher throughput and increased overall network performance gain. This can be achieved by employing efficient coordination between multiple transmission points (i.e., antennas, sectors), which form a CoMP area. A key element for the coordination in the CoMP area as well as the basis for transmission decisions and adaptations is the channel state information (CSI), reported by user equipment (UE). Each report transmitted by a user forms the basis for different transmission decisions of the cooperating BSs in the CoMP area. Such reports include a variety of measurements (i.e., channel quality indicator, CQI; rank indicator, RI; precoder matrix indicator, PMI), which are utilized by the cooperating network elements.

According to 3GPP [27], CoMP can be classified as inter-site CoMP and intra-site CoMP, as illustrated in Fig. 3.8. In the former, the coordination is performed

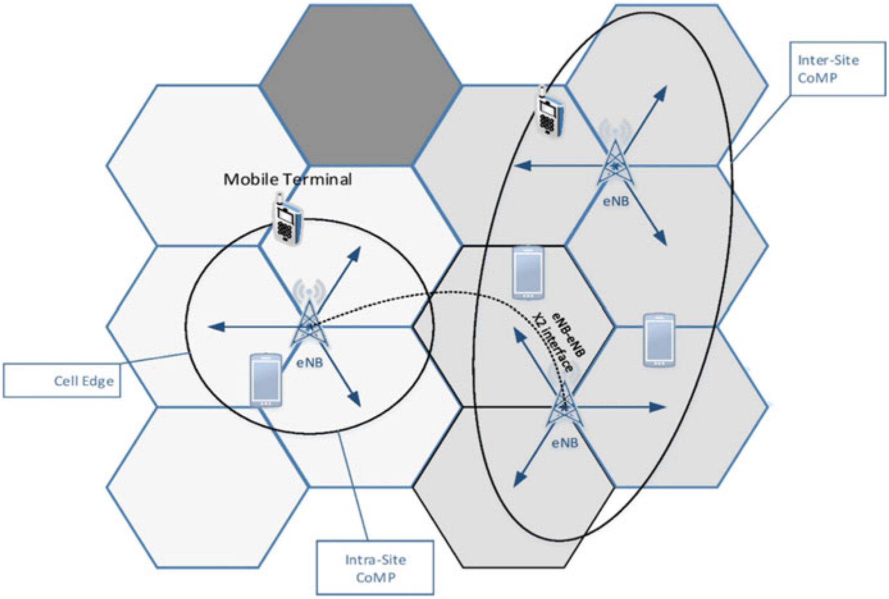


Fig. 3.8 Intra and inter-site CoMP

between BSs located at separate geographical areas, whereas intra-site CoMP allows the coordination between sectors of the same BS using multiple antenna units.

There are three different scenarios where CoMP can be applied:

- **Homogeneous network with intra-site CoMP:** In this scenario, the coordination area is restricted to the sectors (cells) of a single site controlled by an eNB. This scenario has the benefit that there is no need for external connections between different sites and no need for fiber backhaul connection.
- **Homogeneous network with inter-site CoMP:** Compared to the first scenario, the coordination area is expanded to include the cells of different sites. This scenario may be realized by having a single eNB coordinate with multiple high-transmission-power RRHs at different cells or multiple eNBs at different sites coordinate with each other. The level of performance in this scenario is dependent on the number of cells involved and the latency of the site connections.
- **HetNets with low-power pico cells within the macro cell coverage area:** In this scenario, coordination occurs between a macro cell and multiple low-transmission-power RHHs. The transmission/reception points created by the RHHs have different cell IDs from the macro cell ID. Although CoMP is applied in both uplink and downlink communication paths of mobile networks, the focus of this study is primarily on the downlink case and particularly in the joint transmission (JT) scenario. JT-CoMP is an advanced scenario of CoMP implementation according to which each UE active in the CoMP area is capable of receiving the same data from multiple transmission points by using

the same radio resource index (i.e., time, frequency) in order to coherently or non-coherently improve the received signal quality and throughput. The data requested by the user are simultaneously available at all transmission points in the CoMP set, and each transmission point transmits the same resource block, thus improving the reception quality of the user [27] studies coherent JT CoMP, which assumes precoders able to exploit the phase and amplitude correlations between channels of different transmission points. In coherent JT CoMP, the transmission signal from multiple TPs is jointly pre-coded to achieve coherent combining in the wireless channel.

3.3.2 RRM and CoMP Approaches

Ultradense network (UDN) deployment in 5G RAN can cause serious co-channel interference if single frequency is used for BSs which ultimately leads to bad QoS experience for cell-edge users [28, 29]. UDNs can be clustered to embrace multifrequency system as 5G RAN architecture always has various RATs and tiers that work in non-overlapping spectrum of frequency using coordinated control of transmission power. The motivations for RRM in 5G RAN are the following: (1) hyper-dense heterogeneous deployment of devices, (2) traffic load imbalance and coverage issue because of different transmission powers of various BSs, (3) different kinds of access restrictions results in various interference levels, and (4) different frequency priorities for accessing channel and strategies of resource allocation [30].CoMP is suggested by many schemes to counter co-channel interference, where many BSs jointly serve receivers with simultaneous data transfer which will result in better data rate and cell-edge performance [31]. This scheme highly depends upon CSI and data sharing between transmission points (TPs) which indicates toward high CSI feedback overhead, backhaul latency and capacity, and synchronization issues between TPs [32]. Game theory-based clustering and cooperation in CoMP is suggested where lower signaling overhead and better performance are observed [33–35]. Cluster-wise beam forming is done jointly with coalition game formation in the small cell for BS cluster. The recursive core is achieved using merge algorithm which has low complexity with better QoS performance in UDNs; however, cross-tier interference management in multi-RAT environment costs high CSI overheads. Another approach includes the optimization of mutual information in a distributed manner for MIMO Gaussian channel where the Nash equilibrium (NE) is achieved by using metrics of transmit covariance [36]. For the activation probability of the relay nodes (RNs), it is suggested that both cooperative and non-cooperative transmissions are considered jointly to improve the overall sum rate. A control model for the channel is created with cooperative clustering, and three architectures of CoMP are used, first is centralized, next is semi-distributed, and finally the fully distributed [37] approach. This model improves the reliability of 5G RAN with performance gain, but signaling overhead can be high during heavy traffic condition [38]. Figure 3.9 shows the typical CoMP setup to help users at the cell edge.

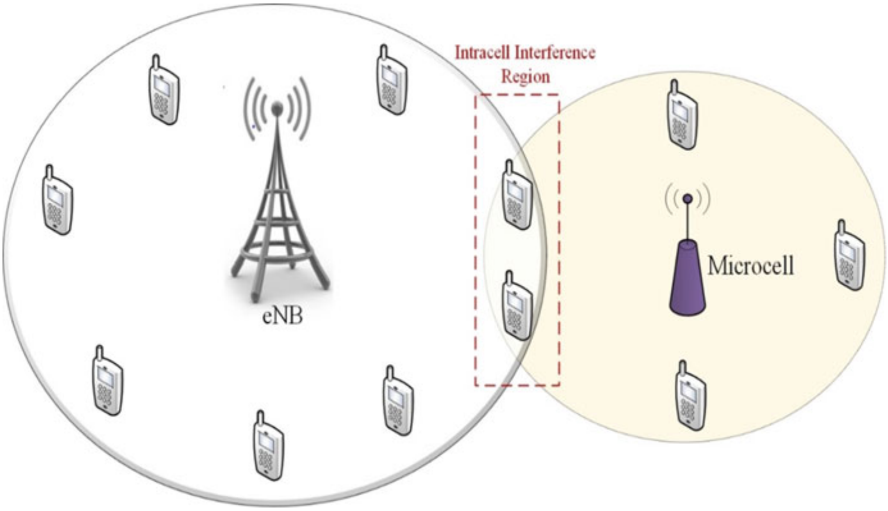


Fig. 3.9 Typical CoMP Scenario

3.3.3 Game Theory for RRM

Game theory is a mathematical decision-making technique that can be used in the context of resource allocation, where decisions are taken by BSs for efficient resource management. The existing cognitive radio approach [39], a strategic game can be formed where every eNB has a strategic profile which is linked with a set of probabilities; these probabilities indicate the use of fix Physical Resource Blocks (PRBs) out of the available PRBs, whenever available PRBs are sensed [40]. The expected payoff is avoidance of co-tier interference for these fix PRBs. The Nash equilibrium [41] is reached at a certain iteration of the game, if the payoff of every eNB is the same for all fixed PRBs corresponding to some probability, and no other profile linked with eNB can provide better payoff than the current one. A utility function is created with three components, the first component targets the data rate demand linked with the eNB, second is related with the fairness of PRBs consumption by eNBs, and the last one corresponds to transmission power control for co-tier interference management; the maximization of this utility function locally will be the goal of the strategic game. The concept of correlated equilibrium is introduced for global and local fairness assurance [42, 43]. A cooperative game approach is also designed with cognitive femtocell networks, where coalitions of UEs are created and they are assigned to one BS; these coalitions maximize the throughput of network with guaranteed fairness. A utility function is formed that shows the throughput achieved by the UEs which belong to the coalition – the payoff of the game will be the increase in terms of throughput using utility function after joining the coalition. More constraints are added to this utility function, which include ensuring the payoff sum should remain equal to the total coalition revenue, subscription of UE will belong to the closed access femtocell, and ensuring a predefined payoff for every UE joining any coalition, otherwise it will remain as a singleton without any BS, and the last constraint relates to fair allocation of resources which ensure no other coalition for UEs exist with better payoff than the current one. These utility function constraints make the optimization problem more accurate for cooperative coalition games, and its solution will satisfy most of the requirements needed for efficient resource allocation [44]. Optimal payoff allocation or solution is referred to as the core, and reaching the core is the overall objective of game. Distributed coalition is also suggested for attaining the core earlier. The implementation complexity of this approach is less; however, reaching the core is a computationally extensive problem.

3.3.4 Cooperative Game Approach for RRM

Inter-cell interference is one of the popular problems RRM which has attracted ample research attention due to evolution of LTE/LTE-A toward 5G and hyper-dense HetNets. A game-based approach is proposed to solve inter-layer interference in [45], which ensures the service to every UE remains unaltered with the increase in the presence of small BSs, though a comprehensive investigation of intra-layer interference is missing. In [46], a resource allocation scheme and joint association algorithm are proposed while considering the maximization of sum utility for rates.

A coalition game scheme is considered for improving spectral efficiency through cooperation by choosing the best coalition based on the total payoff of distributed players [10]. An interference mitigation solution is presented in [47], where cooperation to form coalition is considered in C-RAN environment, although aggregate interference is not considered in the utility formation. In another approach, Markov chain model-based coalition analysis approach is also presented for finding the best cooperative coalitions to improve user's SINR [3]. CoMP has been proposed in [27]; the JT-CoMP technique has been shown to be capable of providing the highest gains in terms of cell capacity in dense homogeneous and heterogeneous cell deployments among the various CoMP schemes. However, several challenges for its effective implementation have been identified by these studies, including the cell clustering and backhaul capacity and latency constraints. Regarding clustering in CoMP, various research papers consider static clustering methods [48]–[50]. In [51], a coalition formation game is modelled to cluster the small cell base stations so that they can perform cooperative beamforming to mitigate the effects of inter-cell interference and shadow fading. In [3, 52], a coalition formation game is formulated to form cooperation clusters to mitigate inter-cell interference and improve user performance via Time-Division Multiple Access (TDMA)-based transmissions.

3.3.5 Coalition Games for CoMP Formation

The proposed coalition-game-based JT CoMP involves three main stages: the formation of the list of interfering RRHs, the formation of coalitions of interfering RRHs, and the coalition game.

Initially, the small cell network includes all non-cooperative RRHs (K), which are referred to as singleton coalitions.

A search of potential coalition patterns among RRHs begins as soon as all users (U) are scheduled and receive interference from the neighboring singleton RRHs. Each UE at the edge of every RRH's cell (edge UE) calculates a matrix of carrier-to-interference ratio values between the serving and interfering RRHs. The interfering RRHs are assigned a unique ID by each user, and this ID is forwarded to the C-RAN. If a RRH is serving multiple edge UEs, the proposed functionality at the C-RAN averages their values, resulting in (K) total interference matrices, which are then sorted in ascending order based on the carrier-to-interference values. The corresponding priority list consisting of the IDs of the interfering RRHs is formed, based on the entries in the top row of the matrices. This priority list indicates the order in which each coalition between the most interfering RRHs will be tested.

In the third stage, all possible coalitions are tested for a duration of few transmission time intervals (TTIs), and their members are coordinated based on the JT CoMP technique. A coalition among two or more interfering RRHs is formed if a set of constraints are satisfied. First, the tested coalition increases (or at least does not decrease) the throughput of every edge UE (i.e., the game's payoff). Second, the throughput of the non-edge UEs is not reduced below a certain threshold value set beforehand. Third, the backhaul capacity constraint is satisfied. The coalition that is formed remains unaltered until the completion of the game. If a coalition has already been tested, the next one (based on the priority list) is examined. Also, if a RRH is already a member of a coalition and its turn, based on the priority list, comes, the total coalition between the new candidate and all the existing members of the coalition is tested. A formed coalition will split only when this split results in the increase of the throughput of at least one edge UE of a member RRH, while the throughput of the other members' edge UEs do not decrease. Following this process, the next rows of the interference matrices are considered, and a new priority list is made. This stage will be repeated until all the considered values of the interference matrices exceed the predefined threshold, in which case, the game ends.

Figure 3.10 depicts the flow of the described coalition formation game algorithm that enhances the JT-CoMP functionality by providing self-organization attributes to the participant RRHs.

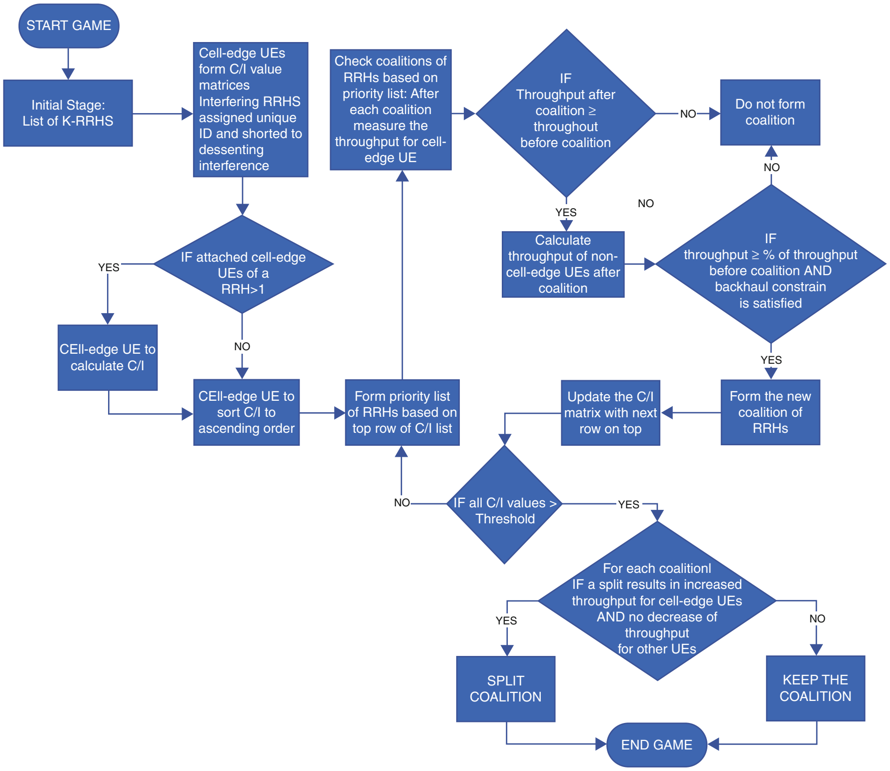

START GAME
Initial Stage: List of K-RRHS
Cell-edge UEs form C/I value matrices Interfering RRHS assigned unique ID and shorted to dessenting interference
Check coalitions of RRHs based on priority list: After each coalition measure the throughput for cell-edge UE
IF Throughput after coalition ≥ throughout before coalition
NO
Do not form coalition
NO
NO
YES
Calculate throughput of non-cell-edge UEs after coalition
IF throughput ≥ % of throughput before coalition AND backhaul constrain is satisfied
YES
IF attached cell-edge UEs of a RRH>1
YES
CEll-edge UE to calculate C/I
NO
CEll-edge UE to sort C/I to ascending order
Form priority list of RRHs based on top row of C/I list
Update the C/I matrix with next row on top
Form the new coalition of RRHs
NO
IF all C/I values > Threshold
YES
For each coalitionl IF a split results in increased throughput for cell-edge UEs AND no decrease of throughput for other UEs
YES
NO
SPLIT COALITION
KEEP THE COALITION
END GAME

Fig. 3.10 Flow chart of proposed cooperation game with CoMP

3.3.6 Results

Figure 3.11 and 3.12 show the impact of cooperative game JT-CoMP on both the edge and normal users, where throughput improvement is shown with and without our cooperative coalition game approach.

More transmission power is allocated to the edge users as they are weak, which results in improved throughput reflecting their SINR values. However, every RRH can allocate a fixed number of RBs, and when extra RBs are allocated to the edge user after coalition formation, it will be detrimental to the normal users attached to this RRH. Therefore, the limitation of the normal user's throughput drop must be defined. The acceptable drop for a normal user throughput from its previous value must not be more than 30% for our scenario. Also, after restricting the drop in normal user throughput and observing the improvement in the value of edge user's throughput, it can be observed that there is a significant increase in overall cell throughput. This indicates the optimal management point of the radio resources of the cell.

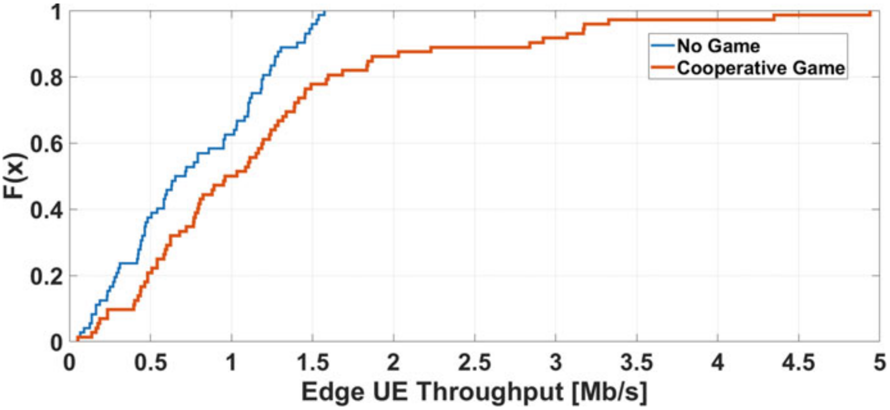


Fig. 3.11 CDF of edge user throughput

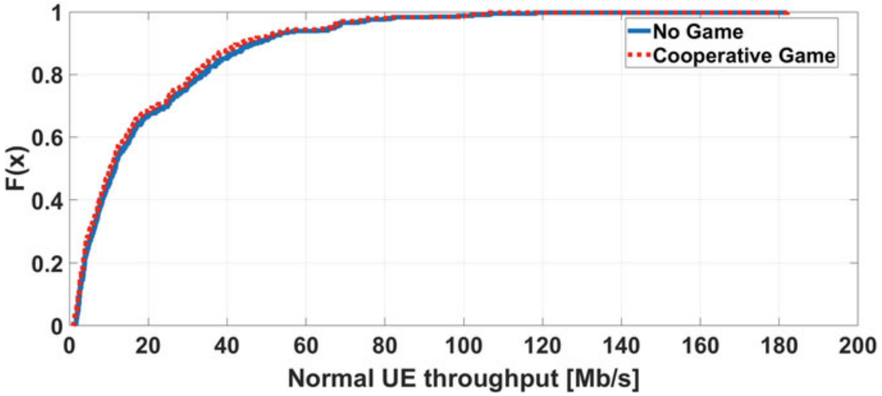


Fig. 3.12 CDF of normal user throughput

3.3.7 Conclusion

The 5G era imposes a set of strict requirements for achieving ultralow latency, high reliability, and high throughput across wireless mobile devices. Such requirements are more difficult to achieve in densely connected networks. Several technology enablers such as CoMP and C-RAN are being considered as candidates for ensuring the migration from the legacy networks to the future wireless radio access architectures. In this section, the benefits of JT CoMP in increasing spectrum efficiency and interference management are explained with the adaptation of a coalition formation game among entities. A system-level simulation was performed that indicates an obvious advantage of the proposed game-theory-based CoMP in terms of throughput and SINR achieved for users located at the cell edge.

3.4 Energy-Efficient Handover for Small Cell Technology

3.4.1 Introduction

3.4.1.1 Importance of Reliable and Signaling-Efficient Handover

To satisfy the increasing data traffic demands in future cellular networks, the ultra-cell densification approach is introduced by the 3rd Generation Partnership Project (3GPP). This approach shrinks the footprints of base stations (BSs) and thus reduces the number of users connected to each BS, consequently improving both the spectral efficiency and frequency reuse. One of the limiting factors in densified deployments is the increased HO rate, i.e., the successive change of handling BS for a moving user. Using the cell densification approach, the gain in capacity is counterbalanced by increased HO rates. The signaling overheads during the HO procedure interrupt the data flow and thus reduce the user throughput [53]. Frequent HOs, unnecessary HOs with ping-pong effects, and HO failures (HOFs) result in high power consumption for both the network and the user device. Specifically, this causes a more detrimental effect on the user battery lifetime.

Future cellular networks need to support data-hungry applications with enhanced data rates, among others, via cell densification (a.k.a. small cells). In addition to providing high data rates, it is equally important to provide a reliable HO mechanism that directly impacts on the perceived quality of experience (QoE) for the end user. A large portion of existing works in the area of small cells overlook the HO procedure and concentrate solely on capacity and throughput studies. Nonetheless, the true challenge of maintaining mobility while providing high data rates for moderate-to-high speed users in urban environments remains.

The mobility challenges described above only become more exacerbated when considering advanced mobility deployments such as on-demand mobile small cells (MSCs). In such scenarios, MSCs can be enabled by mobile relay nodes (MRNs) or by individual user equipments (UEs) via Device-to-Device (D2D) communication in order to provide cellular coverage on the move. In the former case, MRNs which are roof-mounted on vehicles provide a MSC servicing users on-board. As addressed by SECRET’s (H2020 ETN project) future MSC-based vision [54], such mobility scenarios require robust, signaling, and energy-efficient HO schemes that should also allow the users to enter and leave MSC seamlessly without interruptions.

3.4.1.2 Legacy Handover and Its Drawbacks

Mobility is a key feature of cellular networks. In the design of future mobile networks, the mobility requirements are getting more demanding, both in terms of robustness against HOFs and reducing energy consumption. HO schemes in both legacy Long-Term Evolution (LTE) and New Radio (NR) systems rely on the measurement and subsequent measurement reporting done by the UEs over downlink (DL) RS (Reference Signal) transmitted by nearby BSs. The current mobility solution in LTE and NR (New Radio) comes at the cost of the increased signaling overheads of measurement reports over the air interface, especially at the cell edges [55, 56]. Therefore, this procedure (coined hereon DL-HO) is problematic in future ultradense networks in terms of HO performance, higher HO signaling, and power consumption. In these scenarios, where the serving BSs may rapidly change, it is important to minimize the measurement of DL RSs to reduce the energy consumption of the UE. In addition, reducing the HO related air-interface signaling messages has also a great potential for improving the UE's battery lifetime.

3.4.1.3 Proposed Uplink Reference Signal-Based Handover Scheme

Motivated by the importance of a reliable HO scheme in fulfilling the future cellular network objectives, we introduce a UL RS-based HO solution (henceforth UL-HO) aimed at reducing the HO signaling overheads and power consumption. Using this method, the UEs transmit UL RSs that are received possibly at several BSs, allowing the network to make intelligent proactive decisions on which BS shall serve a given user, instead of relying on the UE measuring DL RS and reporting back measurements. This method eliminates the measurement report signaling during the HO procedure; consequently, it contributes to reducing the HO delays, HOFs, and energy consumption [55]. The UL-HO is also suitable for MSCs as it is signaling and power efficient and reduces the chances of a MSCs HOF between donor BSs (i.e., the BS serving the MSCs via a wireless backhaul). In [55], we found that the measurement report signaling contributed the highest over the air interface, which can be reduced by employing the proposed UL-HO solution. The UL-HO scheme can be easily implemented in both LTE and NR, since both systems share the same HO principles, with only some terminology differences [56]. In this section, the power consumption of DL-HO and UL-HO is compared to quantify the potential benefits of the latter.

The rest of this section is organized as follows: in Sect. 3.4.2, the concepts of the DL-HO and UL-HO procedure, power consumption model, system model, and simulation analysis are presented. Finally, Sect. 3.4.3 summarizes the outcomes of this work.

3.4.2 From DL Handover to UL Handover Scheme

Essential to any HO procedure is its effect to the transmitted and received power in both user devices and the network. A sound model of the power consumption for HO signaling is required when modelling the system and evaluating HO performance in terms of energy efficiency. In this subsection, a comparative performance analysis is given based on a definition of both DL and UL HO procedures and a formulation

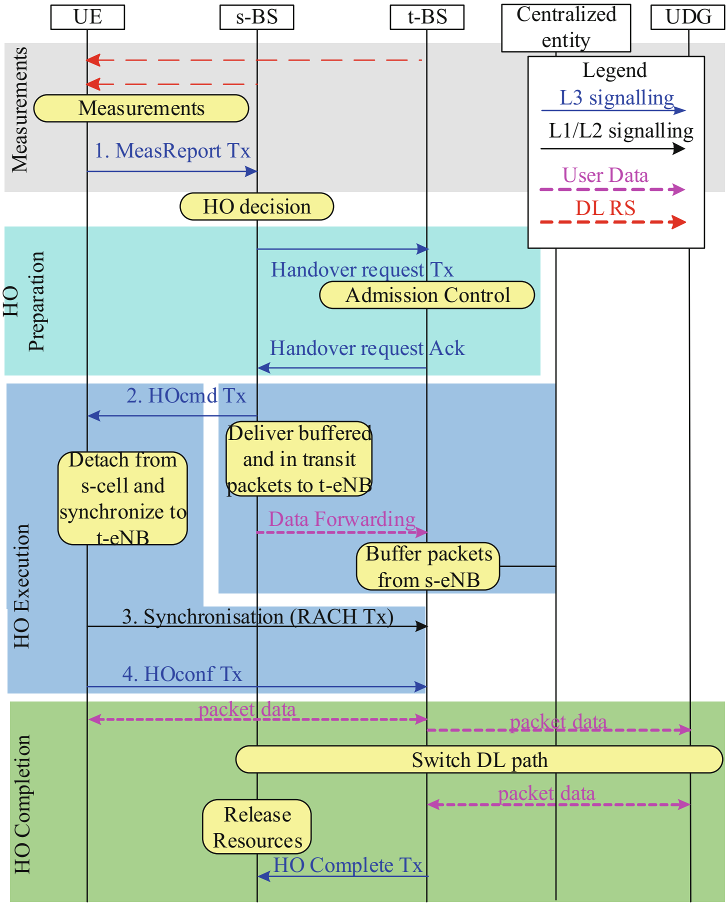


Fig. 3.13 Current cellular networks HO procedure (Adapted from [57, 58])

of the transmitted and received power consumption models for their corresponding HO signaling procedures.

3.4.2.1 A Brief Overview of Legacy Handover Scheme and Pitfalls

A brief overview of the current (legacy) HO procedure in 3GPP cellular networks is shown in Fig. 3.13 [57, 58], which relies on the DL measurement-based HO (DL-HO) procedure for both LTE and NR. As it will become apparent below, in comparison to the UL-HO scheme, the DL-HO scheme requires increased power consumption, especially at the UE side as the UE has to measure the DL RSs transmitted by the nearby BSs and report back to the serving BS (s-BS) with a list of target BSs (t-BSs) and associated signal strength measurements.

The DL-HO scheme is divided into three stages: HO preparation, HO execution, and HO completion stage. The UE performs DL signal strength measurements over specific RS resources to evaluate the reference signal received power (RSRP) from the s-BS as well as the nearby cells. After processing the RSRP measurements, if an entry condition is fulfilled, a measurement report (MeasReport) is transmitted to the s-BS. The A3 event is used as an entry condition expressed as the RSRP of the t-BS being higher than that of the s-BS plus a hysteresis margin (called A3 offset). This entry condition has to be maintained during a time defined by the Time-to-Trigger (TTT) timer [57]. Once the MeasReport is correctly received at the s-BS, the HO preparation stage starts with a HO request transmitted from the s-BS to the t-BS. Upon successful admission, the t-BS accepts the HO request sent by the s-BS and prepares for HO. Subsequently, a HO command (HOcmd) is transmitted from the s-BS to the UE. If successful, the HO execution stage begins in which the UE accesses the t-BS by means of synchronization and a random access (RA) procedure, followed by the transmission of a HO confirmation (HOconf) message. Finally, the t-BS transmits a HO complete message to the s-BS to inform the success of the HO when the DL data path is switched from the user data gateway (UDG) toward the t-BS. Subsequently, the s-BS releases the allocated resources.

3.4.2.2 Energy-Efficient UL Handover (UL-HO) Scheme and Requirements

In contrast to the DL-HO procedure presented in the previous subsection, Fig. 3.14 shows the UL-HO measurement procedure. Hereby, the UE transmits UL RSs which are possibly received at several nearby BSs, allowing the network to perform UL signal strength measurements. These measurements are processed in a central network controller to make intelligent proactive decisions on which BS shall serve a given user. Analogously to the DL-HO case, if the UL-RSRP of the s-BS is less than a t-BS by an “A3 UL-offset,” and this condition is maintained during an “UL Time-To-Trigger (UL-TTT)” time, the controller may decide that the t-BS shall serve this UE. If so, the s-BS sends a HO request to the t-BS. The rest of the HO procedure remains the same as in LTE and NR starting from the HO preparation phase in Fig. 3.13 and described in the previous subsection.

The aforesaid UL RS can be effectively implemented in both LTE and NR by using the sounding reference signal (SRS) (see [59, 60]), noting that reusing the SRS for HO procedures requires no extra signaling overhead. Another benefit of using UL measurements for UE mobility is that it is possible to improve the network performance by changing the HO optimization parameters (i.e., TTT and A3 offset) as per the scenario requirements by network-side upgrades without UE impact. The benefits of using the UL-HO scheme come at the cost of some new requirements, for example, time synchronization between BSs, as several BSs need to receive the UL RSs simultaneously, and coordination of UL RS resources between different cells to avoid pilot contamination between UL RS (i.e., SRS).

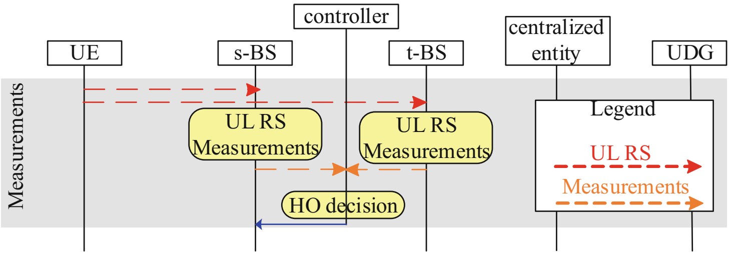


Fig. 3.14 Uplink reference signal-based HO (UL-HO) measurement procedure (Adapted from [55, 57, 58])

Handover Power Consumption Model

This subsection presents transceiver power consumption model for both the BS and the UE air interface signaling that occurs during HO, the details of which are covered in [55].

Transceiver Power Consumption Model at the BS

The supplied power to the BS, necessary to either transmit or receive signaling s, is denoted by $P_{BS,\text{sup}}^{s,Tx/Rx}$ and can be calculated as follows [55]:

$$P_{eBS,\text{sup}}^{s,Tx/Rx} = P_{eBS,Tx/Rx}^{s}/\eta + N_{TB}^{s}/N_{TB}^{DL} \cdot \left(P_{RF,BS} + P_{BB}'\right), \tag{3.11}$$

where $P_{BS,Tx/Rx}^{s}$ is the allocated BS transmitted or received power (in W) per signaling message s and η is the power amplifier efficiency. $P_{RF,\,BS}$ denotes the supply power contribution of the RF equipment, which is conveniently scaled by the portion of utilized resources by signaling message s. Similarly, P_{BB}' is the basic baseband unit (BBU) consumption in W (see Table 3.1).

Transceiver Power Consumption Model at the UE

The supply power required for the UE to transmit or receive signaling message, s, is denoted by $P_{UE,\text{sup}}^{s,Tx/Rx}$ and is given by [55]

$$P_{UE,\text{sup}}^{s,Tx/Rx} = P_{UE,Tx/Rx}^{s} + N_{TB}^{s}/N_{TB}^{UL} \cdot \left(P_{RF,UE} + P_{Tx/RxBB}\right), \tag{3.12}$$

where $P_{UE,Tx/Rx}^{s}$ is the allocated UE transmitted or received power (in W) per signaling message s and where the supply power contribution to the RF and BB part is also scaled by the portion of utilized resources by signaling s. $P_{Tx/RxBB}$ is the transmitted or received UE BBU power (see Table 3.1) where R_{Rx} is the received data rate that is a multiplication of signaling rate and the carried bits in a transport block (TB). The time-averaged supply power to capture the time-domain system dynamics is given by

$$\overline{P}^{s,Tx/Rx}_{x,\text{sup}} = P^{s,Tx/Rx}_{x,\text{sup}} \cdot T^{s}_{x} \cdot R^{s}_{x}, \tag{3.13}$$

where T^{s}_{x} is the signaling duration in seconds and $\mathrm{R}^{\mathrm{s}}_{\mathrm{x}}$ is the signaling rate which will be obtained from system-level simulations.

System Model

A hexagonal grid of 16 tri-sectored BSs is considered in a MATLAB-based system-level simulator. In order to ensure fair interference conditions across the simulation scenario, a cell wraparound feature is included. A set of 100 UEs (with fixed speed and random directions [0°, 360°]) are placed randomly over the simulation scenario. The considered simulation time is 60 seconds which is large enough for statistical confidence. Further details regarding the simulator modeling are covered in [55]. The simulation implementation is largely based on the LTE standard. Table 3.1 summarizes the main simulation assumptions.

Table 3.1 Simulation parameters and assumptions

Feature	Implementation
Network topology	A hexagonal grid of 16x3=48 cells (wraparound included)
Inter-site distance	500 m
System bandwidth	B_{sys}=5 MHz (paired FDD), with $N_{RB}^{DL} = N_{RB}^{UL} = 25$ RBs at carrier frequency f_c=2.1GHz, 1TB=6 RBs, $N_{TB}^{DL} = N_{TB}^{UL} = \lfloor 25/6 \rfloor$
BS DL power	P_{eNB}= 43 dBm
UE power	P_{UE}= 23 dBm
Antenna patterns	3D model (specified in [64] – Table A.2.1.1.2-2)
Channel model	6 tap model, Typical Urban
Shadowing	Log-normal shadowing mean 0 dB, standard deviation 8dB
Propagation model	$L = 130.5 + 37.6\log_{10}(R)$, R in km
UE speed	30 km/h
RLF detection parameters	T310=1s, N310=1, N311=1 as specified in [65] $\mathrm{Q_{in}}$=-4.8 dB; $\mathrm{Q_{out}}$=-7.2 dB as specified in [66]
HO parameters	TTT= 32 ms, A3 offset = 1 dB, L3 filter coefficient K=4
Number of TBs per each signaling message	$N_{TB}^{MR} = 1$ TB; $N_{TB}^{HOcnf} = 1$ TB; $N_{TB}^{HOcmd} = 2$ TBs; N_{TB}^{RA} =1 TB [55]
Power consumption calculation parameters	$\eta = 0.311$ (31.1%), $P_{RF,\,BS} = 12.9$ W, $P_{RF,\,UE} = 2.35$ W, $P'_{BB} = 29.4$ W, $P_{TxBB} = 0.62$ mW, $P_{RxBB} = 0.97.$ $R_{Rx} + 8.16$ (mW) [55]
Signaling timing	$T_{eNB}^{HOcmd} = 1$ms; $T_{UE}^{MR} = 1$ms; $T_{UE}^{HOcnf} = 1$ms; $T_{UE}^{RACHtx} = 1$ ms [55]

Uplink Reference Signal Model

In this work, we will use the sounding reference signal (SRS) to implement the UL RS for HO purposes [59, 60]. In current LTE/NR standards, SRSs are used to estimate the UL channel to maintain uplink synchronization, perform accurate link adaptation, support frequency selective scheduling, and determine the channel quality information in the UL direction. The BS configures the sounding bandwidth, periodicity, subframe offset, and frequency via higher-layer signaling on a cell-wide basis. In addition, each UE is individually configured with different sounding bandwidths, periodicities, sequence, and hopping patterns to achieve resource orthogonality.

Table 3.2 presents the considered numerical values for SRS parameters. The total bandwidth to be sounded (within the system bandwidth) is defined by the SRS bandwidth configuration parameter ($\boldsymbol{C_{SRS}} \in \{\mathbf{0}, \mathbf{1}, \ldots, \mathbf{7}\}$) and the SRS bandwidth parameter ($\boldsymbol{B_{SRS}} \in \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}\}$) along with the partial sounded bandwidth ($\boldsymbol{m_{SRS,b}}$, with $\boldsymbol{b} = \boldsymbol{B_{SRS}}$) at each SRS transmission. The frequency hopping pattern followed by different SRS transmissions is defined by a hopping parameter ($\boldsymbol{b_{hop}} \in \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}\}$) to sound a portion or the entire sounding bandwidth, i.e., at $\boldsymbol{b_{hop}} = \boldsymbol{B_{SRS}}$, frequency hopping is done over the full sounding bandwidth. The subcarriers occupied by an SRS transmission bandwidth are defined by a comb parameter $\boldsymbol{K_{TC}}$, i.e., $\boldsymbol{K_{TC}} = \mathbf{2}$ assign a comb index $\boldsymbol{k_{TC}} = \{\mathbf{0}, \mathbf{1}\}$ for odd and even subcarriers to multiplex SRS transmissions over the same bandwidth. Zadoff-Chu sequences are allocated to different UEs to provide code-domain multiplexing. To guarantee orthogonality in the code domain, UE-specific sequence cyclic shift index ($\mathrm{n}_{\mathrm{SRS}}^{\mathrm{cs}} \in \{0, 1, \ldots 7\}\Big)$ and a cell-specific sequence identifier ($\mathrm{n}_{\mathrm{ID}}^{\mathrm{SRS}}$)) are defined in the standards. SRS transmissions are configured with SRS periodicities ($\mathrm{T}_{\mathrm{SRS}}$) ranging from 2 ms to 320 ms [61]. The SRS Configuration Index ($\mathrm{I}_{\mathrm{SRS}} \in \{0, 1, ..636\}$) calculates the different SRS periodicities. In this work, we will simulate different SRS periodicities to find an optimum SRS periodicity value in terms of the lowest power consumption. Figure 3.15 shows the abovementioned SRS parameters for the case of two SRS transmissions.

Simulation Evaluation

In this section, a simulation analysis is provided for the UL-HO procedure in terms of both signaling and power consumption costs with a detailed comparison of power consumption between DL-HO and UL-HO methods. In [55, 62, 63], it is found that the measurement report transmission has the highest contribution over the air interface. Also, the largest contributor to UE power consumption is the MeasReport transmission by the UE [55]. The UL-HO scheme is utilized to cope with the problem of high power consumption due to MeasReport signaling.

In [55, 62], it is found that the optimum ISD out of many simulated cases is the ISD 500 m case. Based on this finding, the ISD is fixed to 500 m and speed to 30 km/h to have a fair comparison of DL-HO and UL-HO schemes. The values

Table 3.2 SRS parameters and values [55, 59, 60]

Feature	Values
SRS periodicity (T_{SRS})	{5 ms, 10 ms, 20 ms, 40 ms, 80 ms, 160 ms, 320 ms} [61]
Number of transmission combs (K_{TC})	2 [59]
Number of OFDM symbols per SRS resource	1
SRS number of symbol per slot	1
Cyclic shift (n^{cs}_{SRS})	{0, 1, 2,. . . . . . ,7} for $K_{TC} = 2$ [59]
Bandwidth configuration (C_{SRS})	7, choices {0, 1, 2, ... ,7} [59]
Sounding bandwidth in PRBs	24
SRS frequency hopping parameter (b_{hop})	0, choices {0, 1, 2, 3} [59]
SRS bandwidth parameter (B_{SRS})	0, choices {0, 1, 2, 3} [59]
SRS configuration index (I_{SRS})	Range = 0 to 636, [59]
SRS duration	0.5 ms (1 slot)

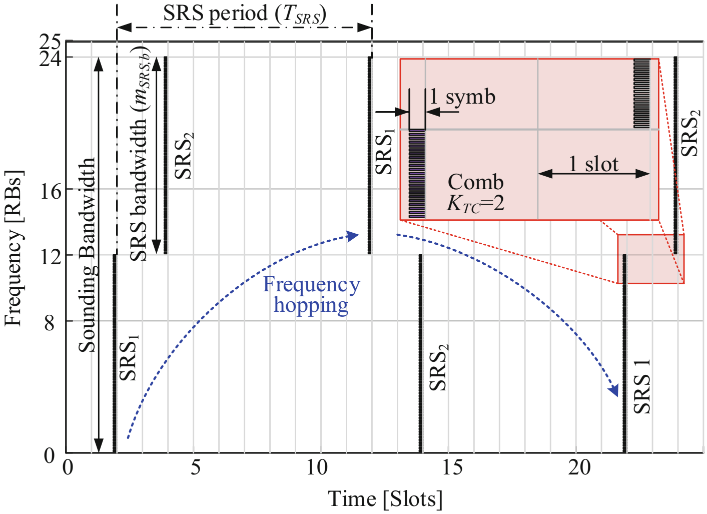


Fig. 3.15 Example of two orthogonal SRS transmissions (SRS1 and SRS2 in the graph) along with main SRS design parameters

of A3 UL and DL offset, UL, and DL TTT are also fixed to 1 dB and 32 ms, respectively. The SRS periodicity values are varied from 5 ms to 320 ms to find an optimum periodicity value that has the lowest power consumption during the HO procedure.

Figure 3.16 shows the UE transmitted (due to transmission of MeasReport, RACH, and HOconf messages) and received (due to the reception of HOcmd message) average supply power consumption comparison for DL-HO and UL-HO schemes. Using UL-HO, no measurement report transmission from the UE to BS is

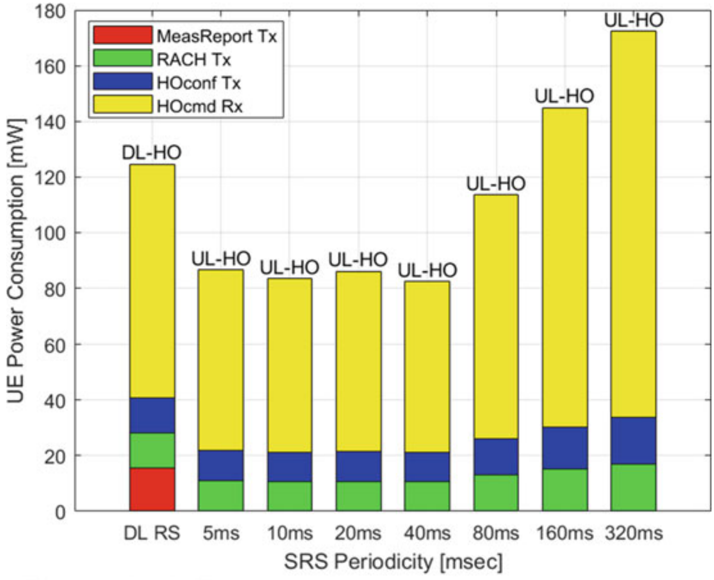


Fig. 3.16 UE transmitted and received average supply power consumption comparison for DL-HO and UL-HO schemes

required; thus, there is no power consumption related to MeasReport signaling. The graph shows that the UL-HO method outperforms the DL-HO for SRS periodicity values up to 80 ms. The lowest UE power consumption is found for an SRS periodicity of 40 ms, almost 42 mW lower than the DL-HO case. Increasing the SRS periodicity beyond 40 ms, the UE power consumption starts increasing because of the high HO rate and PP rate as observed in [55].

Figure 3.17 presents a comparison of average supply BS power consumption resulting from the transmission of HO command along with the reception of MeasReport, RACH, and HOconf signaling. The BS power consumption decreases for low SRS periodicity values, and then it starts increasing at high periodicity values because of a high number of HO rates due to the high PP rate observed in [55]. The minimum BS transmitted power consumption is obtained for the SRS periodicity of 40 ms, almost 500 mW lower than the DL-HO method.

Figure 3.18 shows the total average supply power consumption for each air interface HO signaling message (the addition of transmitted and received power consumption) where the percentage of power increase or decrease in comparison to the DL-HO scheme is also shown. The graph shows that the total power consumption exhibits a decreasing trend in comparison to DL-HO until a sweet spot of 40 ms SRS periodicity case arrives, and then it again starts increasing. Increasing the SRS period beyond 40 ms, the high power consumption is due to frequent HOs we noted in [55]. The lowest total average supply power consumption is found for the SRS periodicity of 40 ms, 30% lower than the DL-HO scheme.

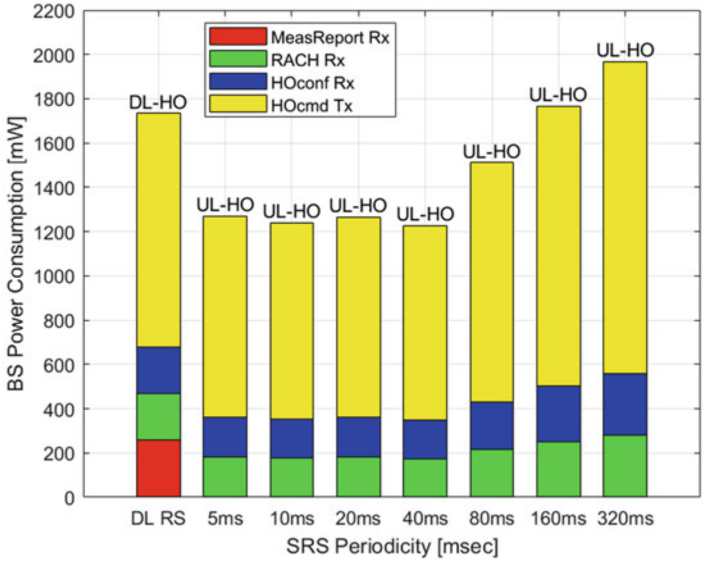


Fig. 3.17 BS transmitted and received average supply power consumption comparison for DL-HO and UL-HO schemes

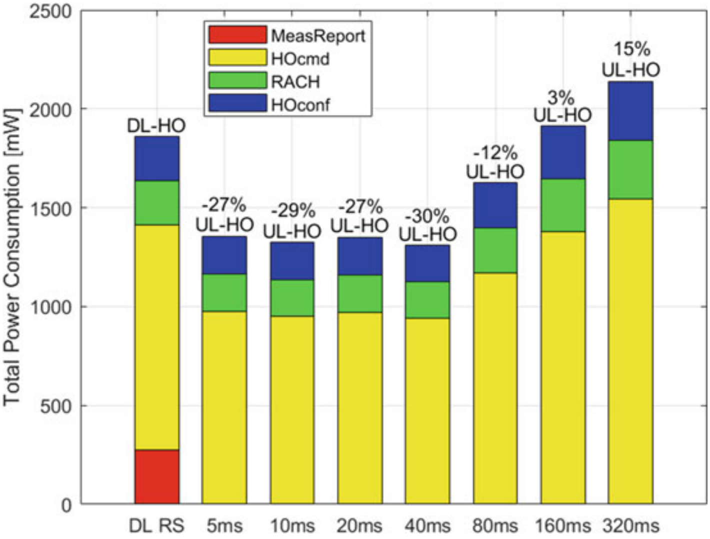


Fig. 3.18 Total average supply power consumption comparison for DL-HO and UL-HO schemes

3.4.3 *Concluding Remarks*

A simulation analysis is performed to determine the power consumption during the handover (HO) procedure for both UL and DL reference signal (RS)-based HO schemes, UL-HO and DL-HO, respectively. We utilize the UL RS-based method to make the measurement procedure more power-efficient because no measurement report transmission is required using this method. This method supports mobility of the network by reducing both the signaling overhead and energy consumption. The simulation analysis shows that the proposed UL-HO method reduces the average supply power consumption of both the UE and BS by almost one third if the UL RS periodicity is carefully chosen (40 ms in our experiments) in comparison to the DL-HO method. The proposed method is power-efficient at both UE and BS side thus reducing the operational expenditures (OPEX) and the environmental effects. It is also an excellent candidate for an energy-efficient HO procedure in future releases of the 3GPP standards heading toward denser network deployments and higher frequencies. These features make the UL-HO scheme appropriate for mobile small cells (MSCs) in 5G and B5G networks. Extending the UL-HO concept to MSCs or the MRNs is expected to bring similar gains.

3.5 Repetition vs Retransmission Scheme for NB-IoT Random Access Channel

A key use case for 5G era is massive-scale IoT connectivity, where the 3GPP Release 13 has been defined as Narrow Band Internet of Things (NB-IoT) technology, among other variants. The performance of NB-IoT technology at the access level utilizes repetition and retransmission mechanisms as an effort to increase the detection probability of preamble transmission and to provide extra time diversity gain for preamble transmission, respectively. We address the challenge of deciding the number of repetitions and retransmissions for a given amount of random access (RA) resources under a massive access scenario. We ended up with a trade-off approach, in terms of access success probability, access latency, power consumption, and time vacancy ratio.

3.5.1 *Introduction*

Legacy cellular networks that are typically designed for conventional human-type communications (HTC) are evolving to support machine-type communications (MTC) or the Internet of Things (IoT) [67–70]. According to analyst firm Gartner, the number of IoT devices will reach up to 20.4 billion in 2020, which will lead to a scenario called massive IoT system (MIoT) [71, 72]. The MIoT refers to the billions of mobile or stationary devices that communicate with each other or to a centralized system through some wireless technologies.

3.5.1.1 MTC Technologies in the 5G Era

The technological solutions to support MIoT can be broadly classified into two: (1) unlicensed spectrum technologies and (2) licensed spectrum technologies or cellular IoT (CIoT) technologies.

- Unlicensed Spectrum Technologies
  Unlicensed spectrum MTC technologies operate at license-free frequency band. Popular unlicensed spectrum technologies are Zigbee, LoRa, SigFox, Ingenu, and Weightless. Use of unlicensed spectrum makes these technologies cheaper (cost-wise) for the user. The infrastructure for these technologies is not widely deployed, and is not 3GPP standardized. Hence, these technologies need to work with non-3GPP air-interface.
- Cellular IoT (CIoT) Technologies
  The licensed spectrum technologies are deployed over the existing cellular network. The licensed spectrum technologies can support a wide range of IoT/MTC use cases with better device management and enhancement in service provisioning, compared to unlicensed technologies. Considering this aspect, the 5G and beyond networks focus on licensed spectrum technologies as major wireless solutions to support MIoT.

3.5.1.2 The NB-IoT Technology

The 3GPP has already introduced three licensed spectrum IoT technologies (also known as cellular IoT: CIoT) in Release 13, namely, (1) Narrowband Internet of Things (NB-IoT), (2) Enhanced MTC (eMTC), and (3) EC-GSM (Extended Coverage-GSM) [73, 74]. The EC-GSM is deployed on the GSM spectrum, while NB-IoT and eMTC are deployed on the existing 4G networks [75]. These CIoT technologies are further enhanced and optimized in the subsequent release of 3GPP from Rel.14 till down to Rel.16 [76–78]. Till today, out of the three CIoT technologies, the NB-IoT is leading the race with 56 NB-IoT networks deployed worldwide, as of May 2019 [79].

The NB-IoT is a cellular technology which is configured to work with one physical resource block (PRB) of LTE spectrum (180 KHz). The existing LTE infrastructure can be software updated to incorporate NB-IoT services together with other LTE UEs. The NB-IoT provides three flexible deployment options, i.e., in-band, guard band, and stand-alone operation. Three coverage enhancements (CE) levels are defined in NB-IoT in terms of maximum coupling loss (MCL): (1) CE-0 (MCL of 144 dB), (2) CE-1 (MCL of 155.7 dB), and (3) CE-2 (MCL of 164 dB). The NB-IoT is suitable for low-rate, ultralow power-consuming MTC applications.

3.5.2 Random Access Procedure in the NB-IoT

3.5.2.1 Four-Step Random Access Procedure

The NB-IoT devices access the network by following a random access procedure. Same as with LTE, in NB-IoT, the devices perform a four-step random access (RA) procedure to establish a connection with the network [80, 81]:

1. Preamble transmission: The main purpose of the preamble transmission is to indicate to the base station the presence of the random-access attempts and to allow the base station to estimate the round-trip delay. This estimated delay will be used in the second step for uplink synchronization. A device starts the RA by transmitting the NB preamble on a randomly chosen NB-physical random-access channel (NPRACH) subcarrier.
2. Random access response: In this step, the base station will send a response to the detected random-access attempt to terminals. The Re-Auth-Request (RAR) message contains:
    - The timing correction calculated by the random-access preamble receiver
    - A scheduling grant, indicating resources the terminals will use for the transmission of the message in the third step
    - A temporary cell identity (TC-RNTI) used for further communication between the terminal and network.
3. Terminal identification: The device sends the Radio Resource Control (RRC) connection request using scheduled uplink resources obtained from the RAR message. A terminal identity is included in the uplink message. This identity is used for the contention resolution in the fourth step.
4. Contention resolution: Multiple devices performing the random access simultaneously with the same subcarrier in the first step have the same terminal identifier. In this case, multiple terminals will receive the same TC-RNTI and resource allocation in step 2. Hence, a contention resolution message is addressed through TC-RNTI. The device checks the match between the identities that it sends in the third step with the identity that it received in the contention resolution message. If the match is found, then the RA procedure is declared as successful, and TC-RNTI is promoted to C-RNTI.

Preamble Format

The preamble consists of four symbol groups and a symbol group comprised of five symbols and one cyclic prefix. The four symbols groups are transmitted over four different subcarrier frequencies (frequency hopping) to facilitate the accurate timing estimation. Subcarrier location of the first symbol group is chosen randomly. The position of the next three preamble group is determined by an algorithm which depends on the subcarrier index of the first preamble group. The subcarrier selection

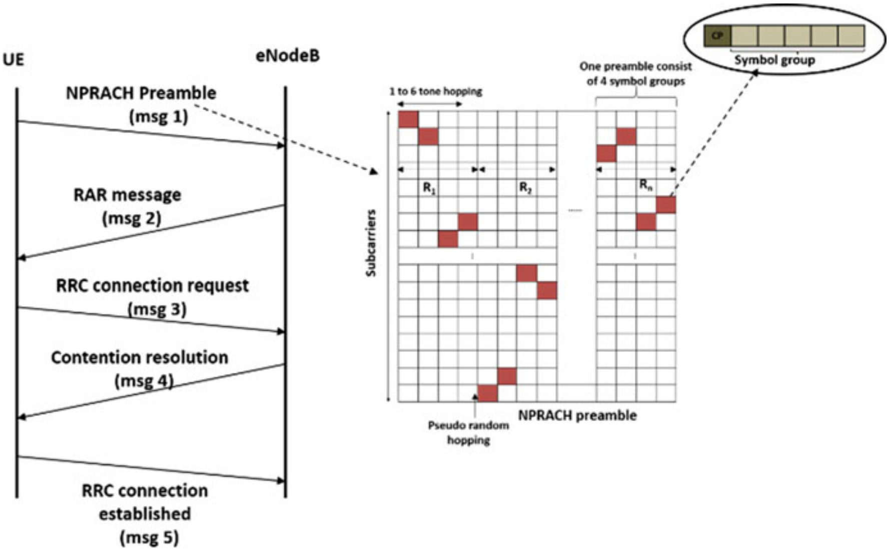


Fig. 3.19 NB-IoT RA preamble format

of the first preamble group after each repetition is determined by pseudo-random hopping. The format of NB-IoT preamble is shown in Fig. 3.19.

Repetition and Retransmission Scheme

Our work concentrates on the first step of the RA procedure wherein the device requests the RA resource by sending a preamble on a randomly chosen RA resources. When a massive number of devices are trying to access the network with the limited available RA resources, *collision* will happen at the base station [82]. The collision happens when two or more users simultaneously transmit their preamble using the same RA resources. Another reason for access failure is *signal outage.* The *signal outage* happens when the transmitted preamble does not possess enough signal strength to be detected at the base station. To reduce the access failure due to collision and signal outage, *retransmission* and *repetition* schemes are used in the NB-IoT.

In the retransmission scheme, the collided devices perform a uniform back-off and retransmit the preamble in the next available RA resource. The basic idea of the retransmission scheme is to provide time diversity by offloading the network traffic over time, under uncorrelated channel conditions. The maximum number of retransmissions that an NB-IoT device can perform is limited by the latency requirements of the NB-IoT system. To overcome the signal outage, a repetition scheme is used in the NB-IoT, wherein in the repetition mechanism, the subframes containing the preambles are repeated for a repetition number defined by the base station. The repeated preambles are combined at the base station to achieve a higher SNR.

3.5.2.2 Problem Under Study

A proper value of repetition helps improve the channel efficiency; at the same time, this may lead to lower detection probability at the base station. On the other hand, a redundant repetition guarantees higher detection probability but leads to resource deficiency for Narrowband Physical UplinkShared CHannel (NPUSCH). One solution in such cases is to compromise on the repetition number by providing extra time diversity gain through additional retransmissions. Since both the repetition and retransmission scheme consume extra RA resources, it is important to figure out under what scenario it is feasible to provide additional retransmissions to the devices. At the same time, extensive use of retransmission may result in extra power consumption and an increase in the access delay. In this context, for a given amount of RA resources, defined in terms combination of repetition and retransmission, we analyzed which combination yields better access success probability and what the effects of the combination on the other access parameters are like access latency, time vacancy ratio, and device power consumption.

3.5.3 System Model

In our simulation, we consider a multi-cell, multi-user scenario over an area of A where the NB-IoT devices and base stations are distributed following a Homogeneous Poisson Point Process (HPPP), with a density λ_B and λ_D, respectively. We assume an equal distribution of devices among three CE levels and an independent and identically distributed (i.i.d) Rayleigh fading channel, where the channel power gain is assumed to be an exponentially distributed random variable with unit mean.

3.5.3.1 Multi-band Slotted Aloha System

The NB-IoT supports three coverage enhancement (CE) levels to support devices working in different (poor to good) coverage areas. The CE levels are defined in terms of maximum coupling loss (MCL) [83].

The RA procedure starts by sending a preamble over a randomly chosen subcarriers. The preamble repetition value is assigned at the beginning of the RA. The preamble value is set such that the preamble detection probability is at least 99% [84]. It is important to mention that while calculating the repetition value, the time diversity gain due to retransmission scheme is not considered. If the preambles are successfully decoded at the base station, the RA is declared as successful, else the RA procedure is considered as failed.

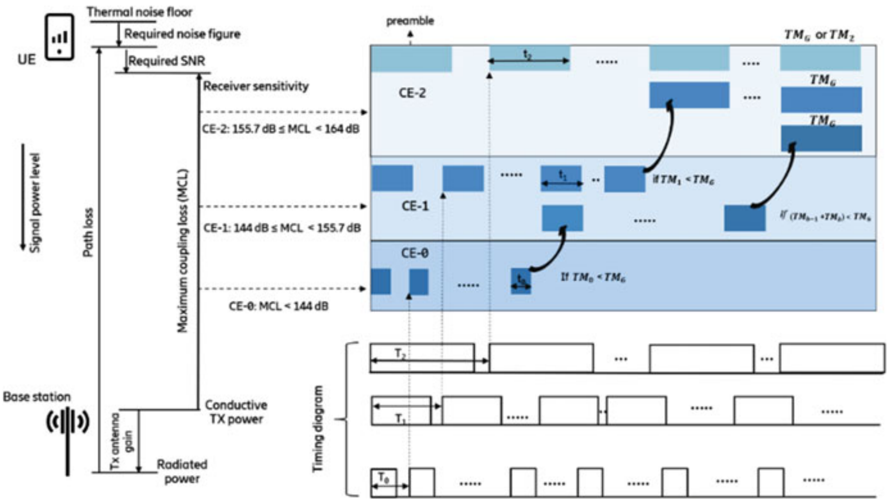


Fig. 3.20 The NB-IoT RA preamble transmission

There are two possibilities for RA failure, first one due to preamble collision and second due to signal outage as mentioned previously. Upon RA failure, the device performs a uniform back-off and retransmit the preamble in the next available RA slot. The NB-IoT system predefines a maximum number of preamble (re)transmission a device is allowed to perform in the entire NB-IoT system and in each CE level. We denote TM_b as the maximum number of retransmissions a device can perform at CE level b, and RT_G is the maximum number of retransmission a device can perform in the entire system, such that $TM_b \leq TM_G$. A device can perform retransmission in the CE level b until the device succeeds or till when it reaches the number of retransmission attempt TM_b. If either of these two conditions is reached, the device moves to the next higher CE level $b + 1$ and performs the preamble retransmission with a higher repetition number, provided that $TM_b < TM_G$ [85]. A schematic of the RA procedure in NB-IoT is shown in Fig. 3.20, where T_b is the RA periodicity in the CE level b.

3.5.3.2 Key Performance Indicators

We follow the simulation scenario mentioned in our previous work [86], to derive the expressions for key performance indicators mentioned in this section.

Access Success Probability

$$P_s = \frac{\sum_{b=0}^{2} \sum_{i=1}^{N_b} S_b\left[i\right]}{\sum_{b=0}^{2} \sum_{i=1}^{N_b} M_b\left[i\right]} \tag{3.14}$$

where $S_b[i]$ is the total number of successful devices in slot i and $M_b[i]$ is the total number of contending devices in slot i.

Access Latency

We define access latency as the average time required for an NB-IoT device to complete the RA procedure. The expressions for access latency is given by

$$D = \frac{\sum_{b=0}^{2} \sum_{i=1}^{N_b} S_b\left[i\right]\ T_b \times (i-1)}{P_s \times \sum_{b=0}^{2} \sum_{i=1}^{N_b} M_b\left[i\right]} \tag{3.15}$$

Power Consumption

The average power consumed by a device to complete the RA procedure is given by

$$\frac{\sum_{b=0}^{2} \sum_{k=1}^{N_b} \sum_{h=1}^{k} G_h\left[k\right] \times P_{trx} + (i-h) \times P_{sleep}}{P_s \times \sum_{b=0}^{2} \sum_{i=1}^{N_b} M_b\left[i\right]} \tag{3.16}$$

where P_{trx} and P_{sleep} are the time-averaged power consumption per device for preamble (re)-transmission and during the back-off, respectively. The $G_h[k]$ is the number of devices which attempt its G_h preamble retransmission in the slot k. The P_{trx} is calculated as

$$P_{trx} = P_{tx}\ \beta\ R_b + P_{sleep}\ \ (1-\beta) \tag{3.17}$$

where P_{tx} is the LTE normal mode device power consumption, $\beta = 5.6$ ms/TTI.

Time Vacancy Ratio

A parameter called *time vacancy ratio* (T_{NP}) is defined to reflect the time available for NPUSCH in a TTI (Transmission Time Interval), as given below:

$$gT_{NP} = \frac{\text{TTI} - 5.6 \times 10^{-3} \times Repetition\ Number}{TTI} \tag{3.18}$$

3.5.4 Numerical Results

In this section, numerical results are presented on the performance analysis of the RA for the various combination of repetition and retransmission mechanisms. The simulation settings are given in Table 3.3. For analysis, we present three schemes, as described below:

- Scheme A: We assigned the repetition value to the devices (mentioned as Rep, here onward) calculated based on [84]. The base station assigns a particular retransmission number for the devices (mentioned as RAO here onwards), such that the maximum RA resources available for each device are $\times RAO$.
- Scheme B: Here, we reduce the *Rep* value used in Scheme A by half and double the *RAO*, such that the maximum number of RA resources is the same for both Scheme A and Scheme B. The idea is to compensate for the reduction in detection probability with extra time diversity gain.
- Scheme C: In this scheme, we double the *Rep* value used in Scheme A and half the *RAO* value. By doing so, we keep the maximum RA resources per device the same for all the three schemes. In Scheme C, we compensate for a reduction in time diversity gain by increasing the detection probability.

Figure 3.21 shows the access success probability, average access delay, average power consumption/device, and time vacancy ratio for the three schemes mentioned above. The selection of *Rep* and *RAO* depends on the back-off window (BW) because the BW determines the distribution of devices over time. So, we have two sets of graphs in each plot with different values BW one for larger BW size and another for smaller BW size.

As we can observe from Fig. 3.21a, when the BW size is large (BW = 8192 ms in the numerical results), Scheme A gives better access success probability at lower values λ_D/λ_B, and Scheme C performance is better at high values λ_D/λ_B. This is because at higher values λ_D/λ_B, the network is congested, and by proving a lower number of *RAO* in Scheme C, we are reducing the number of contending devices in each RA slot. In this way, we lower the network congestion. Moreover, by doubling the *Rep* value, we are increasing the probability preamble detection of the contending devices. Even though Scheme C provides a better access success probability, from Fig. 3.21b, we can observe that the scheme increases the average access delay. This is because when the *Rep* value is high, the base station has to wait for all the repeated preamble to receive to start the decoding. Also, from Fig. 3.21b, we can observe that a high value of Rep decreases the time vacancy ratio, as repetition consumes more resources for the RA within a TTI, which results in lower channel resources for uplink data transmission. The access success probability of Scheme B is the lowest compared to Scheme A and C when BW = 8192 ms. This can be explained as in Scheme B, we compensate for the reduction in detection probability by providing extra time diversity. When BW size is large, when the collided devices perform the uniform back-off, the devices are spread over long duration time, so that a device may not be able to perform all the allocated

Table 3.3 Simulation settings [87, 88]

Parameter	Value
UE Tx power	23 dBm
Coverage level	CE-0 (144 dBm), CE-1(155.7 dBm), CE-2 (164 dBm)
SNR threshold at base station	−4.7 dB
Area	300 km^2
$\mathrm{T}_{\mathrm{max}}$	10 sec
TTI	640 ms
Base station noise figure	5 dB
Power control at UE	Full path-loss inversion power control, $\rho = -122$ dBm (CE-0), $\rho = -133$ dBm (CE-1), $\rho = -141$ dBm (CE-2)
Scheme A	Repetition: CE0 = 2, CE1 = 4, CE = 32Retransmission: CE0 = CE1 = CE2 = 4
Scheme B	Repetition: CE0 = 1, CE1 = 2, CE = 16Retransmission: CE0 = CE1 = CE2 = 8
Scheme C	Repetition: CE0 = 4, CE1 = 8, CE = 64Retransmission: CE0 = CE1 = CE2 = 2
Average power consumption	$P_{sleep} = 133.09$ mW, $P_{tx} = 2200$ mW

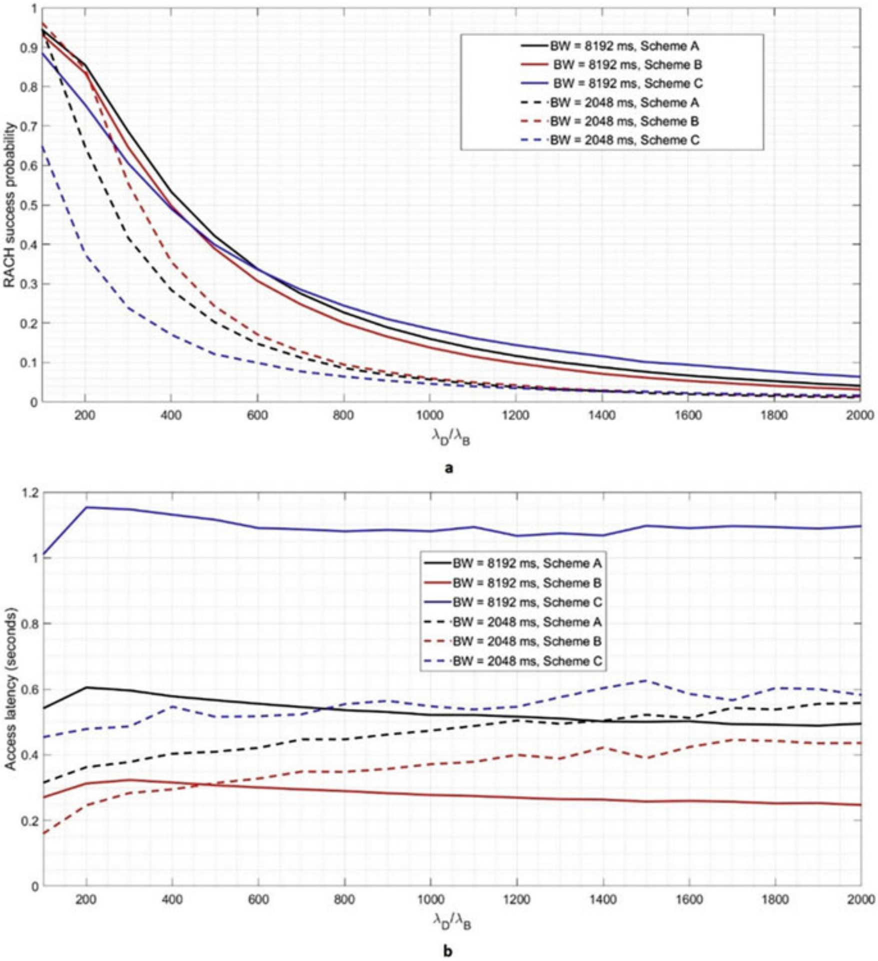


Fig. 3.21 Comparison of **a** random-access success probability and **b** access latency in a multi-cell, three-CE-level NB-IoT system within a predefined time-bound of 10 seconds for Scheme A, B, and C

RAO within the specified time duration, which results in a lower access success probability.

In cases when the NB-IoT applies a small BW size (BW = 2048 ms in the numerical results), Scheme B yields better access success probability for almost all the values of λ_D/λ_B. This is because, with a small BW, when the collided devices perform uniform back-off, the devices will spread over a reduced time frame, so that the devices have chances to perform all the allocated *RAO* within the given time duration. On the other hand, in Scheme C, with BW = 2048 ms, a higher value of *Rep* increases the detection probability of the preamble, which in turn increases the collision at the network, and subsequently results in insufficient time diversity gain to compensate for the increased collisions. Thus, the access success probability of Scheme C is the lowest with BW = 2048 ms. Scheme B however provides better access success probability, at the same time accounts for a high power consumption per device as shown by Fig. 3.22a, especially at high values of λ_D/λ_B. At high values of λ_D/λ_B, due to severe collision, the devices may try the maximum allowed retransmission attempts to successfully access the network. In this scenario, in addition to the power consumed for preamble (re-)transmission, the device consumes extra power while staying in the sleep mode (for the duration between two consecutive RA attempts), which results in increased power consumption.

3.5.5 Conclusion

The NB-IoT applies repetition and retransmission mechanisms to increase the random-access success probability. The repetition mechanism improves the effective signal-to-noise ratio at the base station, while the retransmission scheme provides extra time diversity gain for the preamble transmission under uncorrelated channel conditions.

From our analysis, it was concluded that with NB-IoT employing a large back-off window, providing more resources to the repetition mechanism with a fewer number of retransmissions, it yields better access success probability and low device power consumption in congested network traffic levels. At the same, providing more resources to the repetition mechanism increases the access delay and also results in a channel deficiency for uplink data transmission. In contrast, when the NB-IoT applies a small back-off window, providing more resources for retransmission with fewer resources for repetition, it provides higher access success probability and lower access delay. But this will increase the average power consumed by the devices to complete the random-access procedure.

3.6 Conclusion

The chapter has studied three significant aspects of RRM advancements for 5G and beyond networks based on radio resource allocation, Cloud-RAN design and mobility, and NB-IoT networks.

In Sect. 3.2, game theory is applied to the problem of resource allocation and user association in mobile networks. The main conclusions can be summarized as follows: while Cloud-RAN is very effective for ultradense networks and HetNets, interference can nonetheless limit performance requiring effective user association. The proposed association algorithms based on coalition games can achieve near-optimal performance for both NOMA and precoding schemes, which are both

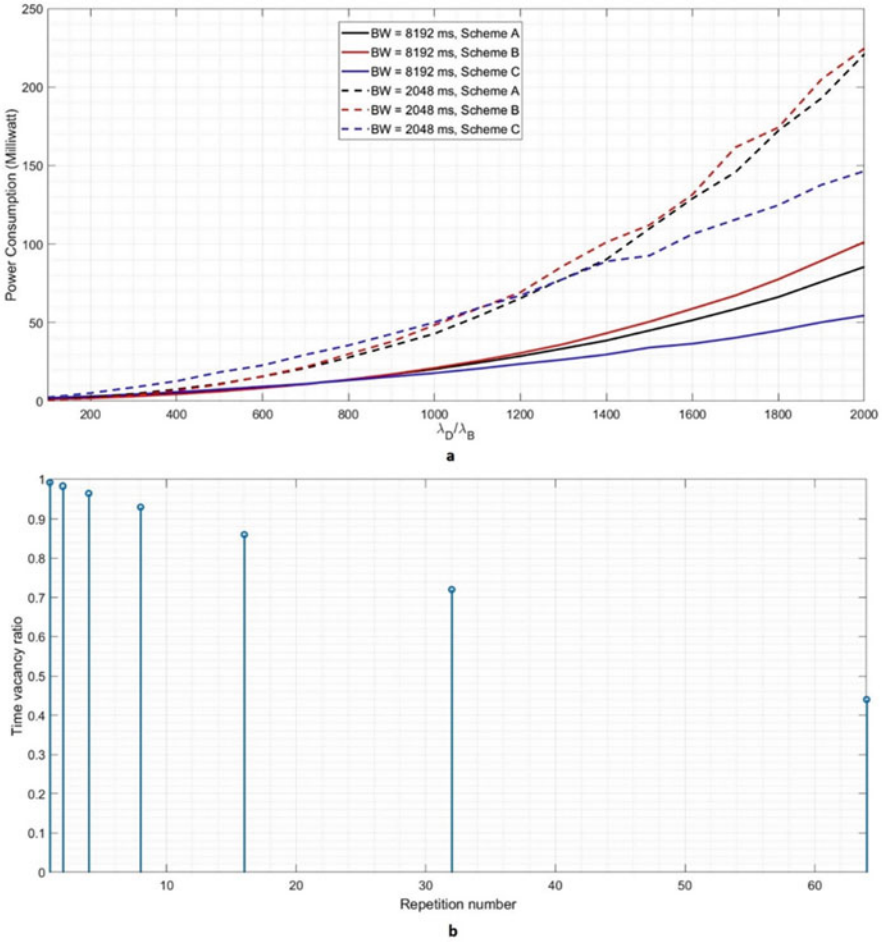


Fig. 3.22 (**a**) Power consumption in a multi-cell, three-CE-level NB-IoT system within a predefined time bound of 10 seconds for Scheme A, B, and C. (**b**) Dependency of time vacancy ratio on repetition number

sensitive to user association. Coalition games can also be similarly applied to the association problem of users with subcarriers to minimize the users' energy use in MEC systems. In this context, we have shown that the games converge to a state where energy consumption is significantly reduced. A comparison of the two algorithms based on this concept shows that sequential swapping converges more rapidly, but random swapping converges to a lower energy state.

Section 3.3 introduces several technology enablers such as CoMP and C-RAN which are part of the 4G/5G standard. The benefits of JT CoMP toward enhancing special efficiency and interference management were developed. We build on these concepts with the application of coalition formation game among entities as a type of optimization approach in order to enhance cell coverage/user throughput at the cell edge. Moreover, system-level simulations were performed that provided viable evidence on the advantages of the proposed game-based CoMP in terms of enhancing the throughput and SINR for users located at the cell edge in contrast to the baseline approaches that consider the absence of game theory.

In Sect. 3.4, the benefits of considering mobility are demonstrated through simulation analysis where the power consumption is determined during handover (HO) for both UL and DL based on the UL-HO and DL-HO, respectively. The UL RS-based method is exploited aiming to efficiently reduce the power needed for the measurement procedure, since no measurement report transmission is required for this method. The proposed method supports mobility of the network with modest signaling overhead and can reduce the average supply power consumption of both the UE and BS significantly. In addition, operational expenditures (OPEX) are reduced which demonstrates that the proposed method is an excellent candidate for an energy-efficient HO procedure in future releases of the 3GPP standards heading toward denser network deployments and higher frequencies.

Finally, Sect. 3.5 introduces the NB-IoT methodology and the repetition and retransmission mechanism which aim to increase random-access success probability. In the presented work, an analytical model of the NB-IoT random access channel is presented, and the trade-off in performance between repetition and retransmission mechanism for the NB-IoT random access channel is given for a given amount of radio resources. The main conclusions suggested that when a large back-off window is used for the NB-IoT, hence, more resources are allocated for the repetition, and a better access success probability is achieved, but at the expense of additional access delay. On the other hand, a small back-off window, and likewise fewer resources reserved for retransmission, can provide both higher access success probability and lower access delay. However, the average power consumption of the devices in order to complete the random-access procedure will significantly increase. Therefore, this work aimed to find an optimality point of operation.

Acknowledgments This research work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreements No 722788 (SPOTLIGHT) and No 722424 (SECRET).

References

1. Ali, M., Qaisar, S., Naeem, M., & Mumtaz, S. (2017). Joint user association and power allocation for licensed and unlicensed spectrum in 5G networks. Paper presented at the IEEE global communications conference, Singapore, 4–8 December 2017.
2. Khodashenas, P. S., Betzler, A., Lloreda, J., et al. (2017). *Ensuring Quality of Service in a multi-tenant cloud-enabled RAN environment*. Paper presented at the European Conference on Networks and Communications (EuCNC), Oulu, 12–15 June 2017.
3. Zhan, S., & Niyato, D. (2017). A coalition formation game for remote radio head cooperation in cloud radio access network. *IEEE Transactions on Vehicular Technology., 66*(2), 1723–1738.

4. Morgenstern O, Neumann J. V (1953) Theory of games and economic behaviour. Princeton University Press, .
5. Shoham, Y., & Leyton-Brown, K. (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.
6. Peleg, B., & Sudholter, P. (2007). *Introduction to the theory of cooperative games*. Springer.
7. Liu, Y., Fang, X., Zhou, P., & Cheng, K. (2017). Coalition game for user association and bandwidth allocation in ultra-dense mmWave networks. Paper presented at the IEEE/CIC international conference communications China, Qingdao, 22–24 October 2017.
8. Wang, K., Liu, Y., Ding, Z., & Nallanathan, A. (2018). *User association in non-orthogonal multiple access networks*. Paper presented at the IEEE Int. Conf. Commun., Kansas City, 20–24 May 2018.
9. Zhou, Z., Peng, J., Zhang, X., Liu, K., & Jiang, F. (2016). A game-theoretical approach for spectrum efficiency improvement in Cloud-RAN. Mobile Information Systems, Hindawi. Article ID 3068732. https://doi.org/10.1155/2016/3068732
10. Han, Z., Niyato, D., Saad, W., et al. (2012). *Game theory in wireless and communication networks – theory, models, and applications*. Cambridge University Press.
11. Rota, G.-C. (1964). The number of partitions of a set. *The American Mathematical Monthly., 71*, 498–504.
12. Eliodorou, M., Psomas, C., Krikidis, I., & Socratous, S. (2019). *User association coalition games with zero-forcing beamforming and NOMA*. Paper presented at the IEEE 20th international workshop on signal processing advances in wireless communications (SPAWC). Cannes, 2–5 July 2019.
13. Huang, S., Yin, H., Wu, J., et al. (2013). User selection for multiuser MIMO downlink with zero-forcing beamforming. *IEEE Transactions in Vehicular Technology, 62*, 3084–3097.
14. Ding, Z., Lei, X., Karagiannidis, G., et al. (2017). A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends. *IEEE Journal on Selected Areas in Communications, 35*, 2181–2195.
15. Metropolis, N., Rosenbluth, A.-W., Rosenbluth, M.-N., et al. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics., 21*, 1087–1092.
16. Eliodorou, M., Psomas, C., Krikidis, I. & Socratous, S. (2019). *Energy efficiency for MEC offloading with NOMA through coalitional games*. Paper presented at IEEE global communications conference (GLOBECOM), Waikoloa, HI, 9–13 December 2019.
17. Wang, F., Xu, J., & Ding, Z. (2017). *Optimized multiuser computation offloading with multi-antenna NOMA*. Paper presented at the IEEE global communications conference, Singapore, 4–8 December 2017.
18. Grant, M., & Boyd, S. (2013). CVX: MATLAB software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx. Accessed July 2019.
19. Chen, K., & Duan, R. (2011). C-RAN the road towards green RAN. China Mobile Research Institute (white paper). October 2011.
20. He, Y., et al. (2019). Cross-layer resource allocation for multihop V2X communications. Wireless Communications and Mobile Computing, Hindawi. Article ID 5864657. https://doi.org/10.1155/2019/5864657
21. Irmer, R., et al. (2011). Coordinated multipoint: Concepts, performance, and field trial results. *IEEE Communications Magazine., 49*, 102–111.
22. Politis, I., et al. (2017). *On optimizing scalable video delivery over media aware mobile clouds*. Paper presented at the IEEE International Conference on Communications (ICC), Paris, 21–25 May 2017.
23. Peng, M., et al. (2014). Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies. *IEEE Wireless Communications., 21*, 126–135.
24. Hossain, E., & Hasan, M. (2015). 5G cellular: Key enabling technologies and research challenges. *IEEE Instrumentation Measurement Magazine., 18*, 11–21.
25. Georgakopoulos, P., et al. (2018). Considering CoMP for efficient cooperation among heterogeneous small cells in 5G networks. Paper presented at the 23rd international workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD),

Barcelona, 17–19 September 2018.

26. Mumtaz, S., et al. (2017). *Self-organization towards reduced cost and energy per bit for future emerging radio technologies – SONNET*. Paper presented at the IEEE Globecom workshops, Singapore, 4–8 December 2017.
27. Lee, D., et al. (2012). Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges. *IEEE Communications Magazine., 50*, 148–155.
28. 3GPP TS 36.211. (2007). Evolved universal terrestrial radio access (EUTRA). Physical channel and modulation (Release 8), Technical Report, 3GPP-TSG R1. September 2007.
29. 3GPP TR 36.932. (2013). Scenarios and requirements for small cells enhancements for E-UTRA and E-UTRAN, version 12.1.0. March 2013.
30. Hossain, E., et al. (2014). Evolution towards 5G multi-tier cellular wireless networks: An interference management perspective. *IEEE Wireless Communications., 21*, 118–127.
31. Li, J. C. F., et al. (2013). SHARP: Spectrum harvesting with ARQ retransmission and probing in cognitive radio. *IEEE Transactions on Communications, 61*, 951–960.
32. Gesbert, D., et al. (2010). Multi-cell MIMO cooperative networks: A new look at interference. *IEEE Journal on Selected Areas in Communications., 28*, 1380–1408.
33. Guruacharya, S., et al. (2013). Dynamic coalition formation for network MIMO in small cell networks. *IEEE Transactions in Wireless Communications., 12*, 5360–5372.
34. Zhou, T., et al. (2014). Network formation games in cooperative MIMO interference systems. *IEEE Transactions in Wireless Communications., 13*, 1140–1152.
35. Mochaourab, R., & Jorswieck, E. A. (2014). Coalitional games in MISO interference channels: Epsilon-core and coalition structure stable set. *IEEE Transactions on Signal Processing., 62*, 6507–6520.
36. Mayer, Z., et al. (2014). On the impact of control channel reliability on coordinated multi-point transmission. *EURASIP Journal on Wireless Communications and Networking., 1*, 1–30.
37. Wang, H., et al. (2015). SoftNet: A software defined decentralised mobile network architecture toward 5G. *IEEE Network, 29*, 16–22.
38. Sonia, N. M., et al. (2014). Uplink power control schemes in long term evolution. *International Journal of Engineering and Advanced Technology., 3*, 260–264.
39. Altman, E. (1999). *Constrained Markov decision processes*. Chapman and Hall.
40. Chung, W. C., et al. (2013). *A cognitive priority based resource management scheme for cognitive Femtocells in LTE systems*. Paper presented at the IEEE International Conference on Communications (ICC), Budapest, 9–13 June 2013.
41. Lien, S. Y., et al. (2011). Cognitive and game-theoretical radio resource management for autonomous Femtocells with QoS guarantees. *IEEE Transactions on Wireless Communications., 10*, 2196–2206.
42. Nash, J. (1950). Equilibrium points in N-person games. *Proceedings of the National Academy of Sciences, 36*, 48–49.
43. Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics., 1*, 67–96.
44. Ghareshiran, O. N., et al. (2013). Collaborative subchannel allocation in cognitive LTE Femtocells: A cooperative game theoretic approach. *IEEE Transactions on Communications, 61*, 325–334.
45. Chandrasekhar, V., et al. (2009). Power control in two-tier femtocell networks. *IEEE Transactions on Wireless Communications, 8*, 4316–4328.
46. Madan, R., et al. (2010). Cell association and interference coordination in heterogeneous LTE-A cellular networks. *IEEE Journal on Selected Areas in Communications., 28*, 1479–1489.
47. Sun, C., et al. (2014). A coalition game scheme for cooperative interference management in cloud radio access networks. *Transactions on Emerging Telecommunications Technologies., 25*, 954–964.
48. Ali, S. S., et al. (2012). *A novel static clustering approach for CoMP*. Paper presented at the 7th International Conference on Computing and Convergence Technology (ICCCT), Seoul, 3–5 December 2012.

49. Marsch, P., & Fettweis, G. (2011). *Static clustering for cooperative multi-point (CoMP) in mobile communications*. Paper presented at the IEEE International Conference on Communications (ICC), Kyoto, 5–9 June 2011.
50. Shimodaira, H., et al. (2016). Diamond cellular network optimal combination of small power base stations and CoMP cellular networks. *IEICE Transactions on Communications, E99*, 917–927.
51. Guruacharya, S., et al. (2013). Dynamic coalition formation for network MIMO in small cell networks. *IEEE Transactions on Wireless Communications, 12*, 5360–5372.
52. Ahmed, M., et al. (2015). Interference coordination in heterogeneous small-cell networks: A coalition formation game approach. *IEEE Systems Journal., 12*, 604–615.
53. Arshad, R., Elsawy, H., Sorour, S., et al. (2016). Handover management in 5G and beyond: A topology aware skipping approach. *IEEE Access., 4*, 9073–9081.
54. Rodriguez, J., et al. (2017). *SECRET – Secure network coding for reduced energy next generation Mobile small cells*. Paper presented at the IEEE internet technologies and applications (ITA) conference, Wrexham Glyndŵr University, Wales, 12–15 September 2017.
55. Tayyab, M., Koudouridis, G. P., Gelabert, X., & Jäntti, R. (2020). Uplink reference signals for energy-efficient handover. *IEEE Access., 8*, 163060–163076.
56. Tayyab, M., Gelabert, X., & Jäntti, R. (2017). A survey on handover management: From LTE to NR. *IEEE Access., 7*, 118907–118930.
57. 3GPP TS 36.300. (2017). E-UTRA and E-UTRAN; Overall description; Stage 2 (Release 15). V15.0.0, Section 10, pp. 93–143. Dec 2017.
58. 3GPP TS 38.300. (2019). NR and NG-RAN overall description; Stage 2 (Release 15). V15.5.0. Mar 2019.
59. 3GPP TS 36.211. (2018). Physical channels and modulation (Release 15). V15.1.0. Mar 2018.
60. 3GPP TS 38.211. (2020). Physical channels and modulation (Release 16). *V16.1.0.*, (Mar 2020).
61. 3GPP TS 36.213. (2018). Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures. Table 7.1.7.2.1–1. Mar 2018.
62. Tayyab, M., Koudouridis, G. P., Gelabert, X., & Jäntti, R. (2019). *Signaling overhead and power consumption during handover in LTE*. Paper presented at the IEEE Wireless Communications and Networking Conference (WCNC), Marrakech, 15–19 April 2019.
63. Tayyab, M., Koudouridis, G. P., Gelabert, X., & Jäntti, R. (2019). *Receiver power consumption during handover in LTE*. Paper presented at the IEEE 5G World Forum Conference, Dresden, October 2019.
64. 3GPP TR 36.814. (2010). Further advancements for E-UTRA physical layer aspects (Release 9). *V9.0.0.*, (Mar 2010).
65. 3GPP TS 36.331. (2010). E-UTRA Radio Resource Control (RRC). *Protocol specification (Release 9). V9.2.0.*, (Mar 2010).
66. 3GPP TS 36.133. (2013). Requirements for support of radio resource management (Release 9). *V9.15.0.*, (Mar 2013).
67. 5G Americas. (2017). LTE progress leading to the 5G Massive Internet of Things (white paper). Available at: https://www.5gamericas.org/wp-content/uploads/2019/07/LTE_Progress_Leading_to_the_5G_Massive_Internet_of_Things_Final_12.5.pdf. Accessed July 2019.
68. Akpakwu, G. A., et al. (2017). A survey on 5G networks for the internet of things: Communication technologies and challenges. *IEEE Access., 6*, 3619–3647.
69. Palattella, M. R. (2016). Internet of things in the 5G era: Enablers, architecture, and business models. *IEEE Journal on Selected Areas in Communications., 34*, 510–527.
70. Al-Fuqaha, A., Guizani, M., Mohammadi, M., et al. (2015). Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys & Tutorials., 17*, 2347–2376.

71. Torchia, M., & Shirer, M. (2017). IDC forecasts worldwide spending on the internet of things to reach $772 billion in 2018. Business Wire. Available at: https://www.businesswire.com/news/home/20171207005963/en/IDC-Forecasts-Worldwide-Spending-Internet-Things-Reach. Accessed July 2019.
72. 5G Americas. (2019). 5G – The future of IoT (white paper) Available at: https://www.5gamericas.org/wpcontent/uploads/2019/07/5G_Americas_White_Paper_on_5G_IOT_FINAL_7.16.pdf. Accessed July 2019.
73. 3GPP TR 23.720. (2016). Study on architecture enhancements for CIoT. *V13.0.0.*, (Mar 2016).
74. Flore, D. (2015). *Evolution of LTE in release*, 13. Available at: http://www.3gpp.org/news-events/3gpp-news/1628-rel13. Accessed July 2019.
75. ETSI TS 122368. (2016). Service requirements for MTC. *Release 13. V13.1.0*, (Mar 2016).
76. 3GPP TS 36.213. (2019). Physical layer procedures V15.4.0 Jan 2019.
77. 3GPP TS 38.213. (2019). NR: Physical layer procedures for control. *V16.0.0*, (Dec 2019).
78. 3GPP TS 23.501. (2018). System architecture for the 5G system. *V15.2.0*, (Jun 2018).
79. 5G Americas. (2019). IoT Deployment [Online]. Available: https://www.5gamericas.org/resources/deployments/deployments-iot/. Accessed July 2019.
80. Wang, Y. P. E. (2017). A primer on 3GPP narrowband internet of things. *IEEE Communications Magazine, 3*, 117–123.
81. De Andrade, T. P. C., Astudillo, C. A., Sekijima, L. R., et al. (2017). The random access procedure in long term evolution networks for the internet of things. *IEEE Communications Magazine, 55*, 124–131.
82. Jiang, N., Deng, Y., Nallanathan, A., et al. (2018). Analyzing random access collisions in massive IoT networks. *IEEE Transactions on Wireless Communications, 17*, 6853–6870.
83. Coverage Analysis LTE-M Category M1 (white paper) Jan 2017 [Online] http://www.coverageanalysisoflte-m.com/. Accessed July 2019.
84. 3GPP TS 36.104. (2018). Base Station (BS) radio transmission and reception. *V15.3.0*, (July 2018).
85. Harwahyu, R., Cheng, R. G., Wei, C. H., & Sari, R. F. (2018). Optimization of random Access Channel in NB-IoT. *IEEE Internet of Things Journal., 5*, 391–402.
86. Narayanan, S., Tsolkas, D., Passas, N., & Merakos, L. (2019). *Performance analysis of NB-IoT random access channel*. Paper presented at international conference on interactive mobile communication, technologies and learning, Thessaloniki, 31 October 2019.
87. Schlienz, J., & Raddino, D. (2016). Narrowband internet of things (white paper). Rohde & Schwarz. Available at https://cdn.rohde-schwarz.com/pws/dl_downloads/dl_application/application_notes/1ma266/1MA266_0e_NB_IoT.pdf
88. 3GPP TS 136 141. (2017). Base Station (BS) conformance testing. *V13.6.0*, (Jan 2017).

Chapter 4
Energy-Efficient RF for UDNs

Ahmed Abdulkhaleq, Maryam Sajedin, Yasir Al-Yasir, Steven Caicedo Mejillones, Naser Ojaroudi Parchin, Ashwain Rayit, Issa Elfergani, Jonathan Rodriguez, Raed Abd-Alhameed, Matteo Oldoni, and Michele D'Amico

Abstract Multi-standard RF front-end is a critical part of legacy and future emerging mobile architectures, where the size, the efficiency, and the integration of the elements in the RF front-end will affect the network key performance indicators (KPIs). This chapter discusses power amplifier design for both handset and base station applications for 5G and beyond. Also, this chapter deals with filter-antenna design for 5G applications that include a synthesis-based approach, differentially driven reconfigurable planar filter-antenna, and an insensitive phased array antenna with air-filled slot-loop resonators.

4.1 Introduction

This chapter targets the design of an energy-efficient and multi-standard RF front-end for next-generation (beyond 5G) multi-homing small cell devices based on S-band and mmWave frequencies. Next-generation ultra-dense networks (UDNs) need to be green, or in other words "energy aware," to support future emerging

A. Abdulkhaleq (✉) · A. Rayit
SARAS Technology, Leeds, UK
e-mail: a.abd@sarastech.co.uk; Ashwain.Rayit@sarastech.co.uk

M. Sajedin · I. Elfergani
Instituto de Telecommunicações, Aveiro, Portugal
e-mail: maryam.sajedin@av.it.pt; i.t.e.elfergani@av.it.pt

J. Rodriguez
Instituto de Telecommunicações, Campus Universitário Santiago, Aveiro, Portugal

Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, UK
e-mail: jonathan@av.it.pt

Y. Al-Yasir · N. Ojaroudi Parchin · R. Abd-Alhameed
University of Bradford, Bradford, UK
e-mail: Y.I.A.Al-Yasir@bradford.ac.uk; N.OjaroudiParchin@bradford.ac.uk; R.A.A.Abd@bradford.ac.uk

S. C. Mejillones · M. Oldoni · M. D'Amico
SIAE Microelettronica/Politecnico di Milano, Milan, Italy
e-mail: steven.caicedo@siaemic.com; matteo.oldoni@siaemic.com; michele.damico@polimi.it

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_4

smart services that are likely to be bandwidth hungry, as well as support multi-mode operation (5G, LTE, LTE-A, HSDPA, and 3G among others) in HetNet (heterogeneous network) environments. In this context, this chapter targets the RF front-end architecture and considers power consumption as a key design metric in the design process. The functional blocks will harness RF design standardization and spectrum policy design requirements to provide concrete solutions addressing reconfigurable antennas, tunable filters, and power amplifiers, the latter representing the key energy consumer in the RF chain. Specifically, Sect. 4.2 addresses MMIC (monolithic microwave integrated circuit) power amplifier (PA) design for 5G handsets, where MMIC technology is employed to enable compact circuit configurations and the confinement of electromagnetic fields within the semiconductor materials, while Sect. 4.3 considers the base station as an application in the form of the energy-efficient load modulation PA based on the Doherty technique, where the main advantages with this approach are lack of complex circuitry and simplicity. Sections 4.4 and 4.5 consider hybrid antenna-filter techniques. The first contribution in this category is referred to as the differentially fed reconfigurable filtering antenna for mid-band 5G applications. RF components using differentially fed ports are in high demand for 4G and 5G applications that offer promising characteristics such as multi-functional property, high common-mode suppression, high roll-off skirt selectivity, and low radiated power loss, among others. This is employed here to develop a filtering antenna with reconfigurability/tenability function at the heart of the design. The second contribution is the compact filtering antenna for phased arrays targeting 5G mmWave FDD backhaul applications. The new wireless fronthaul/backhaul network in 5G and beyond is expected to be reconfigurable to satisfy the dynamic nature of the mobile traffic. Phased array antenna (PAA) in the backhaul equipment is a promising solution not only to respond to this design requirement but also to provide the equipment with the ability to automatically recover the link when misalignments occur, increasing its availability. Finally, Sect. 4.6 covers the insensitive phased array antenna for 5G smartphone applications, where the insensitive property provides robust and consistent performance for different antenna substrate materials, including the handheld effect. In particular, an insensitive phased array antenna with air-filled slot-loop resonators is proposed for 5G mobile applications. It is designed on the FR-4 substrate and working at 21–23.5 GHz. Eight elements of the metal-ring elements are linearly arranged on the phone PCB. The proposed design showed a good performance that takes into account the handheld effect.

4.2 MMIC Power Amplifier Design for 5G and Beyond RF Front-End

The PA is a key enabler for an energy-efficient RF front-end, being the main power consumer of the RF transceiver architecture, where the overall system efficiency is approximated by PA characterizations of efficiency and linearity. In this context,

Table 4.1 Overview of semiconductor materials

Material	Si	SiGe	SiC	GaAs	GaN	InP
Electron mobility (Cm^2/V_s)	900–1100	>2000	500–1000	5500–7000	400–1600	10,000–12,000
Peak drift velocity ($10^7 cm/s$)	0.3–0.7	0.1–1.0	0.15–0.2	1.6–2.3	1.2–2.0	2.5–3.5
Band gap (eV)	1.1	<1.1	2.2	1.4	3.2	1.3
Freq. range (GHz)	<40	10–40	15–20	>75	20–30	>115
Gain	Moderate	Better	Lower	Higher	Lower	Higher
Noise figure	Moderate	Good	Poor	Good	Poor	Good
Production maturity	12″ wafer	8″ wafer	4″ wafer	6″ wafer	4″ wafer	2″ wafer

heterogeneous PAs in terms of minimizing energy consumption and offering high gain that can prolong battery lifetime are key design requirements. The Doherty technique is generally perceived as one of the most popular efficiency enhancement solutions for high-speed and low-power applications that has been extensively researched to meet the upcoming front-end and base station requirements, where more details will be explained in Sect. 4.2.3. As a first step, we need to choose an appropriate technology to address the substrate material and possible geometry of passive components, where we then build on this by reviewing the latest key developments in MMIC DPAs, and explore innovative approaches via a number of design examples that can minimize the PA die size and cost without compromising the gain and linearity performance.

4.2.1 MMIC Active Device Technologies

In order to meet the required specifications in terms of output power, efficiency, bandwidth, and chip area, the first and most important step is the technology selection that covers the substrate material and possible geometry of passive components. The compound substrates are typically characterized by thermal conductivity, cutoff frequency, breakdown voltage, integration level, and cost. Several basic properties of MMIC semiconductors including gallium arsenide (GaAs), silicon (Si), silicon carbide (SiC), silicon-germanium (SiGe), and gallium nitride (GaN) that provide solutions for a wide range of applications are listed in Table 4.1. SiGe HBT features good linearity and low cost for cellular handset PAs, even though it is affected by heating runaway issue and low efficiency. Inherent high band gaps of both the SiC and GaN offer the possibility of operating with high voltages, which minimizes the losses of matching and improves the thermal conductivity. Since higher breakdown voltage provides more robust devices, GaN HEMT on Si or SiC can maximize the power density and supply up to 180 W output power with efficiency up to 70% at the base station [1].

GaAs is widely accepted as a superior technology with excellent features for small-dimension and high-frequency coverage handset applications. The semi-

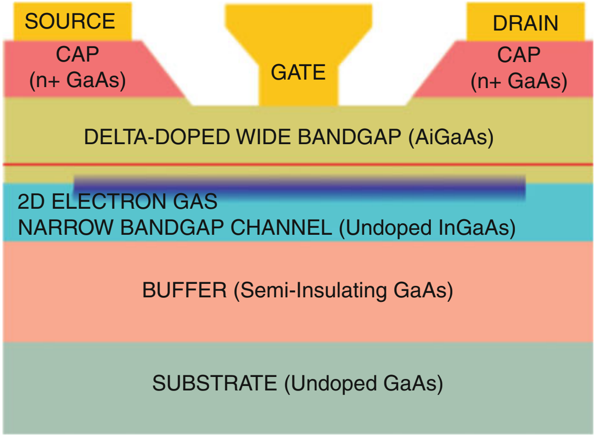

SOURCE
GATE
DRAIN
CAP
(n+ GaAs)
CAP
(n+ GaAs)
DELTA-DOPED WIDE BANDGAP (AiGaAs)
2D ELECTRON GAS
NARROW BANDGAP CHANNEL (Undoped InGaAs)
BUFFER (Semi-Insulating GaAs)
SUBSTRATE (Undoped GaAs)

Fig. 4.1 Cross-sectional representation of pHEMT structure including several layers of ion-implanted GaAs

insulating substrate property of GaAs provides high carrier mobility that enables fast switching in lower intensity and contributes to long-term operation as well. InGaP HBT benefits from high current gain and single power supply polarity and can downsize die and module area within the cell phone front-end PAs. However, GaAs HBT device's nonlinearity influences the ideal gain characteristic and reduces its optimum operation. In general, the imbalance of the thermal region due to the mismatch of the entire HBT geometry for small-scale emitter size causes heat dissipation. The heating effects can be compensated using a ballast resistor as an equalizer with the penalty of gain reduction [2].

GaAs E-pHEMT technology is the industry workhorse for monolithic integration of low-noise and low-loss active and passive components up to mmWave for handset devices. In fact, it is more reliable in low-voltage operation for thermal runaway, possessing the advantage of high gain, good linearity, and great transition frequency. Even though very thin GaAs substrate thickness (<75 μm) has a low thermal resistance, thicker substrate (>100 μm) can be used for lower loss [3]. The cross section of a GaAs MESFET is depicted in Fig. 4.1. A thin layer of low-energy band gap InGaAs is sandwiched between both sides of higher-energy band gap material layers of un-doped GaAs and doped ALGaAs. Hence, the confined electrons with a heterojunction have much higher drift velocity and move to energy levels within the thin layer that yields performance improvement.

4.2.2 Monolithic IC Technology and Design Methodology

Advancement in microelectronics technology has enabled compact circuit configurations and formed the confinement of electromagnetic fields within the semi-

Table 4.2 MMIC features vs hybrid MICs

Feature	Monolithic	Hybrid
Substrate	Semi-insulator	Insulator
Interconnections	Deposited	Wire-bonded/deposited
Solid-state device	Integrated	Discrete
Controlled parasitics	Yes	No
Equipment cost	High	Low
Design flexibility	Very good	Good
Broadband performance	Relatively good	Limited
Integration with digital ICs	Possible	N/A
Reliability	Excellent	Fair

conductor materials. MMICs drive PAs to be implemented on the surface of a semi-insulating substrate, providing much higher level of IC integration. In fact, MMICs have been pursued to outpace the discrete devices for critical targets of multi-octave operation that offers new opportunities and challenges in RF front-end PA design and realization. MMIC offers potential advantages such as thin metal thickness, multi-stage design for higher gain, better amplitude and phase tracking, compact size, higher operating frequency, broadband performance, circuit design flexibility, and higher reliability in contrast to the hybrid microwave integrated circuit (MIC)-packaged transistor counterpart provided in Table 4.2. While the discrete transistors are mounted on alumina substrate, the well-characterized MMIC devices are implemented by foundry processes. MMIC foundries design a mask-set that consists of multiple geometries of active transistor features and passive distributed components of metal-insulator-metal (MIM) capacitors, spiral inductors, and resistors that simulate the actual fabricated elements' characteristics on wafer.

In contrast to hybrid microwave, the post-fabrication tuning for monolithically fabricated circuits is no longer available since it severs the design. In fact, the chip design flow is based on using process design kits which introduces the foundry-specific components with more tolerance to process variation and automatic RF IC testing on wafer for sufficient statistical characterization data.

MMIC design flow starts with a realistic circuit scheme realized by a matching network design process to justify the assumptions. The design synchronization aims to follow the layout rules dictated by the foundry process via hole spacing, as well as ensure proximity between layers. Additionally, the statistical analysis on the effects of metal interconnection within the physical device has to be performed using a full electromagnetic EM simulator to minimize coupling effects and to verify the layout standards. Subsequently, the circuit parameters should be optimized for improving yield for smaller chip sizes or stepping back to adapt stronger robust topology. Furthermore, electrothermal analysis to model the channel temperature as a function of gate width, dissipated power, and number of fingers should not be trivialized. Finally, the design can pass to the manufacturing step for final check for packaging requirement and test. An accurate modeling of microstrip discontinuities and acceptable circuit performance can speed up the wafer fabrication process (Fig. 4.2).

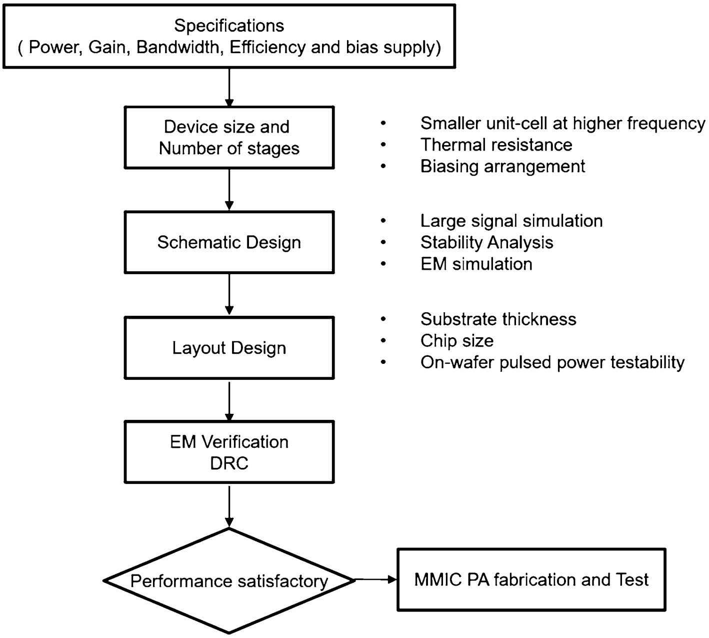

Specifications
(Power, Gain, Bandwidth, Efficiency and bias supply)
Device size and Number of stages
• Smaller unit-cell at higher frequency
• Thermal resistance
• Biasing arrangement
Schematic Design
• Large signal simulation
• Stability Analysis
• EM simulation
Layout Design
• Substrate thickness
• Chip size
• On-wafer pulsed power testability
EM Verification
DRC
Performance satisfactory
MMIC PA fabrication and Test

Fig. 4.2 Complete MMIC design flow

4.2.2.1 MMIC Transistor Size Selection

In the mobile handset device and on every portable battery-operated transmitter, where the actual size and weight of the overall apparatus are dictated by the choice of lightweight and small battery pack, the available bias voltage, voltage swing, and device's maximum current are all limited. In this regard, opting for the most appropriate MMIC process that transfers specific output power per millimeter length is of paramount importance. In particular, the total dimensions of the device can be estimated by sufficient output power specification and available output power of a unit transistor cell. The latter factor is based on the semiconductor technology selection. While enlarging device size improves the output power, it has an inverse effect on the delivered gain, which indicates the trade-off between output power and gain over the frequency range. In fact, the output power is proportional to the device size because it is a function of drain voltage, drain current, and efficiency, while the drain voltage is restricted by breakdown voltage specified by substrate material.

One possibility to achieve the maximum output power of device is to enlarge the gate periphery up to the optimum useful device size that can afford an adequate level of gain [4]. This technique depends upon the process selection and the power budget, which is carried out by enlarging the transistor gate width or increasing the number of gate fingers. It should be noted that by expanding the gate periphery, higher output power can be obtained up to a certain size; however, after exceeding the size restriction, the power densities and device gain start to drop due to saturation scaling and parasitics of output capacitance, along with phase error between the gate fingers and thermal issues [5]. Therefore, to ensure the required specifications of bandwidth, gain, output power, and efficiency, the consistent combination of several gain stages must be adapted.

The multi-stage matching MMIC PAs constitute synthesis, analysis, and optimization of input, inter-stage, and output matching networks starting from the output stage and proceeding back toward the former stages. Since the interaction between stages degrades the output power overall PAE and stability across the band frequency, nonlinear models for obtaining higher PAE and wider bandwidth should be taken into consideration in all stages. It is more desirable to employ fewer stages for higher gain and wider bandwidth, since one of the challenges associated with cascade designing is the bandwidth shrinkage [6]. Besides, higher gain at output stage improves the peak efficiency that results in higher output power capacity, while any additional gain stages only enhance the overall device gain in expense of direct current power consumption. Therefore, higher PAE can be attained by a minimum number of high-gain stages.

4.2.2.2 Matching Network Design Considerations

A two-stage MMIC power amplifier is shown in Fig. 4.3. The output matching network (OMN) is designed to maximize the output power, since the MMIC PA operates at about 1 dB gain compression point of maximum output power and the gain of the last stage considerably improves the PAE; therefore, it is of paramount importance to select an appropriate bias point for the output stage. All the matching networks contribute to combine or split the power of transistor cells with lowest loss, cover the required bandwidth, and enhance the stability. Several division/combination schemes of the power signal using the same substrates as matching networks have been adapted for monolithic PAs such as Wilkinson power divider/combiner [7], Lange couplers [8], and branch-line couplers [9] that further

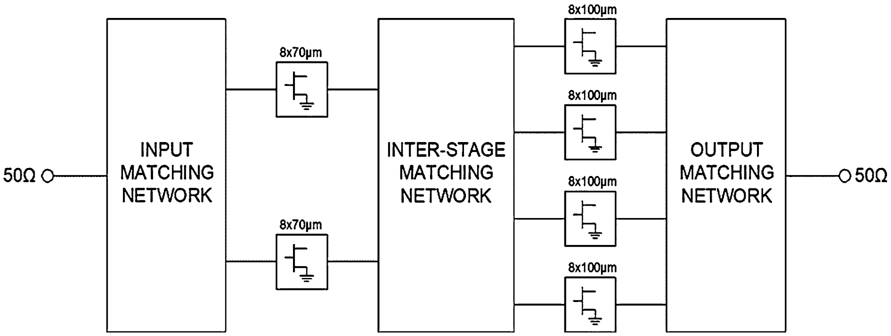


Fig. 4.3 Two-stage matching design

improve the overall output power and suppress any instability. One of the most compact techniques to combine a large number of transistor cells is the metal strip bus-bar combiner [10] that not only feeds the direct current to all the individual transistors in the array in-phase but also implements good matching to the load terminations. The multi-stage ladder [11] is one of the common solutions to combine devices in parallel with symmetrical arrangement of components; however, parallel combining decreases the linearity and bounds the bandwidth with increasing number of transistors for low-voltage technologies. Note that the number of transistors used in the output stage dictates the width of PA chip.

While the input matching network enhances the input return loss by optimizing the impedance match, the inter-stage matching networks (ISMNs) aim to compensate the gain roll-off and minimize the mismatch loss for gain flatness. Hence, the ISMNs are designed to transfer the optimal load impedance that maximizes the input power of the succeeding stage with direct current separation between bias suppliers. In fact, the ISMN should be large enough to drive the output stage into saturation for archiving the best level of power and efficiency. The optimum output power can be attained, when the input power of the output stage is high enough to compress the gain by 1 dB. Nonetheless, if the earlier stages cause the PA to reach the 1 dB compression point before the output stage is fully compressed, a lower peak efficiency level will be achieved. As a consequence, the preceding stages should be designed for linear operation. Employing high-pass networks or large series capacitors that resonates with the matching inductors are two proposed solutions to accomplished this task [12]. A broadband ISMN requires direct current isolation (blocking capacitor) between preceding and following stages, when the biasing network is generally integrated within the matching network for smaller size [13]. The gate bias contains a resistor to provide isolation between power supply and device, and the drain bias is connected to the bypass MIM capacitors, due to its influence on the maximum available gain.

4.2.3 Doherty Power Amplifier Design for Mobile Handsets

To enable a compact front-end design, significant efforts have been exerted to provide the efficient and linear handset PA when the die area has to be minimized. Dynamic load modulation, envelope tracking, and dynamic control of quiescent current techniques have been adapted individually or applied in synergy to improve the average efficiency of PAs without compromising its linearity [14]. While the envelop tracking enhances the efficiency, it requires an external wideband, small, and efficient envelope amplifier that challenges its implementation. On the other hand, at mmWave band frequencies, due to the high loss of passive components, the majority of PAs exhibit very limited back-off efficiency improvement with degraded linearity. In this context, a great interest has been on the Doherty PA (DPA) for beyond 5G multicarrier applications due to its enhanced efficiency at deep power back-offs for low supply voltage applications and wideband operation potential.

The main challenges associated with the practical DPA implementation for mmWave band applications lie in its nonlinear distortion, sensitivity of the input delay to peaking, realizing the load modulation by narrowband quarter-wavelength power combiners, as well as adding offset lines and phase compensation network. The linear operation of DPA handsets is more critical than at the base station due to the absence of digital predistortion techniques. For lower-power levels, the carrier PA should operate in a highly linear fashion, even though it matches to a high impedance, while within the Doherty region, where the level of output power with the phase shift between the two active devices can significantly vary, the optimum gate bias voltages of both PAs should contribute with an offset line at the peaking path to enable linear operation of the DPA. The linearity of a fully integrated HBT DPA has been improved [15] by employing a consistent direct power splitter based on band-pass filter (BPF) networks and a phase compensation network in the peaking path that can adjust the power dividing ratio according to the input impedance of the carrier and the peaking PAs, while the peaking PA receives more power with lower impedance for better power handling.

On the other hand, the gain compression of the carrier PA and the gain expansion of the peaking PA introduce an opposite third-order IMD phase relation between the carrier and peaking PAs that has a detrimental effect on the linearity of the DPA. The symmetric IMD3 of the carrier PA can be suppressed by proper bias condition of the peaking PA that can trade off between the flat response of the nonlinear characteristics of the amplitude modulation/phase modulation (AM/PM) variations and the efficiency enhancement. Furthermore, the precise harmonic load terminations can help to alleviate the IMD generation, since the IMD2 can be cancelled by short-circuiting the second harmonic [16]. Typically, a bias buffer is required for handset PAs to set the bias point; however, the available bias headroom of HBT devices at very low supply voltage is limited due to the summation of two series PN junctions that deteriorates the efficiency performance at a bias point less than 3 V.

Despite the fact that employing Class-AB for a carrier PA with a resistive load at the fundamental and short-circuiting all harmonics can optimize the DPA linearity, Cripps [17] indicates that both the efficiency and output power have been reduced. For better harmonic suppression, Class-F operation mode uses both arms' shunt stubs and a series tuning line to minimize the simultaneous presence of current and voltage, except at the fundamental frequency across the targeted bandwidth. In fact, the Class-F/F^{-1} improves the efficiency and output power by inserting the second and third harmonic traps realized by the extra control circuits (HCCs) [18]. A fully matched microwave saturated DPA based on Class-F PA including HCCs and offset lines to control the finite harmonic content according to the power levels is proposed in [19] that can deliver excellent efficiency of 53% PAE and 10 W output power at 2.14 GHz. However, it should be noted that the harmonic manipulation for waveform shaping at higher frequencies is more challenging due to the low impedance of the drain capacitor. On the other hand, the constant peak voltage waveform of a continuous Class-F design can be supplied by technologies with very large breakdown voltage, while the low-voltage technologies used in

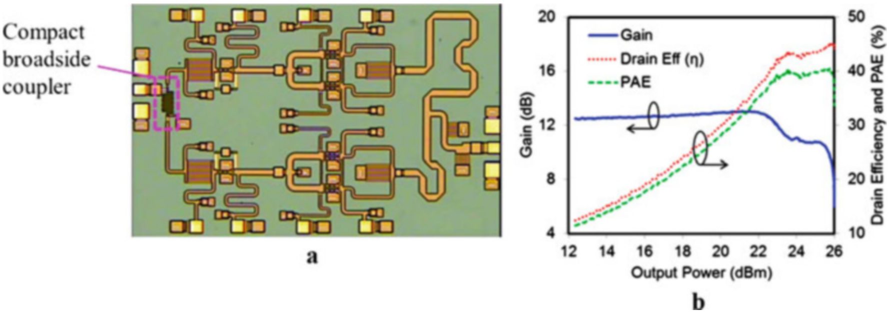


Fig. 4.4 (**a**) Layout of GaAs DPA MMIC. (**b**) Experimental results

handset PAs diminish the output power densities. Furthermore, incorporating HCC enlarges the die size of the device and narrows the band frequency response which is not desirable for the RF front-end module. The reported fully integrated 2 μm GaAs HBT DPA in [20], designed for handset application, can regulate the second harmonic for higher efficiency across the frequency band. Such a lumped DPA can provide over 27% PAE and 23.6 dBm output power across 2.2–2.8 GHz for WiMAX application.

Toward further enhancing the DPA efficiency, [21] employs HCC to resonate out the output capacitance of active devices by the inductances of the bias lines. While the short-circuited second harmonic has minor effects on the DPA linearity, it generates an in-phase second harmonic current at peaking PA that maximizes the half-sine current and results in a higher PAE. The proposed DPA employs an off-chip output matching network on the package model and a high-pass π-type lumped matching technique as a direct input power divider that also eliminates the offset lines and the quarter-wave transmission line. The implemented DPA based on the InGaP/GaAs HBT at 1.9 GHz has a chip area of $1.1\times1.2\ mm^2$ and provides 25.1 dB gain, PAE of 45%, and an output power of 27.5 dBm. However, in practical DPAs based on HBT, the maximum gain is not maintained, and the nonlinear components will vary as an exponential function. A Ka-band DPA based on E-mode GaAs process using an input broadside coupler is reported in [22] that can provide 29% PAE at 6 dB OBO and 26 dBm output power at 28 GHz. Figure 4.4a shows the chip photo of the DPA with size of 2.2 × 1.3 mm, and Fig. 4.4b shows the corresponding measurement results. The proposed integrated ultra-compact broadside coupler using stack-up as an input power divider can eliminate the offset line at the peaking path, since the coupler has an inherit phase shift.

Cripps [23] presented the conceptual analysis of even-order harmonics on improving the linearity using transistors with nonlinear transconductance that validates the superior linear performance of Class-J mode DPAs using only the second harmonic voltage enhancement technique in contrast to Class-F PAs that utilize odd-order harmonics. Small-size HCCs for second harmonic load can be realized by MIM capacitors and bond-wires at the drain of the carrier and peaking

PAs. To further reduce the MMIC DPA size, it is recommended that all the $\lambda/4$ lines at the impedance transformer give way to an equivalent low-pass π-type distributed network of series transmission line connected to shunt capacitors at each end [24]. Class-J PAs offer a linear operation of the fundamental frequency, potential wideband characteristics, and high-efficiency performance. The principle of Class-J mode of operation is to shift the phasing of the current and voltage waveforms. The overlap between the drain voltage and current waveforms introduces a pure reactive component that can be utilized to terminate the second harmonic and provide bandwidth extension as well [25]. Harmonic tuning of the Class-J PA with the purpose of engineering the voltage waveform is accomplished within the low-loss matching network, without requiring HCC. However, the overlap between waveforms degrades the efficiency improvement and can be reduced by an accurate second harmonic voltage component generation, while higher harmonics are shorted out. Employing a nonlinear gate-source capacitor at the gate node [26] or nonlinear drain-source capacitor for shaping the voltage waveform can balance the phase shift. The nonlinear capacitor changes the frequency components of the drain voltage by generating odd-order harmonics that increases the fundamental components and load impedance; subsequently, the output power and efficiency will be improved. Although this approach shapes proper current and voltage waveforms, higher-order filters increase the PA implementation complexity. Also, the phase shift of the second harmonic voltage with respect to the fundamental voltage components affects the efficiency [27] exploit the second harmonic tuning technique incorporating dual-band active load modulation. This technique uses the reduced conduction angle of Class-C current injection to the drain node of main Class-J PA at the second harmonic frequency. The injected current poses an additional phase shifter and introduces the so-called Class-J2 PA that is characterized by a band-pass filter tuned for the second harmonic and a phase offset between the main transistor current and the injected current. However, the power injection at the second harmonic provides a maximum 5% higher drain efficiency than that of Class-B PA because of the direct current power consumption of the peaking PA. Furthermore, since the Class-C PA conducts current only at the second harmonic, the output power at the fundamental frequency will be degraded. This topology has been modified by reducing the phase shift relative to the drain current of the main PA and including the fundamental frequency injection, where the current flow to the load at the fundamental will contribute to higher output power. The state of the art of MMIC DPAs implemented on different technologies is summarized in Table 4.3.

4.3 Load Modulation Amplifier for Energy-Efficient 5G Base Stations

The development of mobile generations has been driven by the users' requirements, where one of the main features is the desired data rate which has significantly

Table 4.3 Performance comparison of mobile handset DPAs

References	Year	Technology	Freq. (GHz)	PAE % OBO	Gain (dB)	Pout (dBm)	Area (mm^2)
[15]	2010	2 µm InGaP/ GaAsHBT	2.5–2.7	25	19	24.6	1.44
[22]	2017	E-mode 0.15 µm GaAs	28	29	12	26	2.85
[28]	2014	D-mode GaAs	23–25	20	12.5	30	4.29
[29]	2018	E-mode 0.15 µm GaAs	31.1	28	14	26.3	3.75
[30]	2019	0.15 µm GaAs	27–30	31	11.8	26.5	2.86
[31]	2017	0.15 µm GaAs	28.5	27	15	28	4.93
[32]	2014	0.15 µm GaAs	22.8–25.2	38	12.5	30	4.29
[33]	2017	0.13 µm SiGe	28	13.9	18.2	16.3	1.76
[34]	2018	90 nm CMOS	34	13.1	19.8	20.7	0.45
[35]	2020	130 nm SiGe BiCMOS	24–30	20	20	28	4.19
[36]	2019	130 nm SiGe BiCMOS	28	19.5	18.2	16.8	1.76

increased over the past 30 years; going beyond legacy voice services, we have data services, such as video streaming, browsing, gaming, etc., that have all been increasingly driving up the data rata requirements, where 5G KPPs target up to 20 Gbps. In this case, complex modulation schemes are used to utilize the available bandwidth effectively. However, the generated modulated signal will have an envelope which has a peak-to-average power ratio that forces the amplifier to operate in back-off from the most efficient point (usually near to the saturation region) into a region that keeps the required linearity. The following subsections will deal with the efficiency enhancement techniques in PAs, with specific emphasis toward DPAs (load modulation technique) [37–40].

4.3.1 Efficiency Enhancement Techniques

Several techniques are used for improving the efficiency of power amplifier such as the out-phasing technique (so-called linear amplification using nonlinear devices, LINC). In this context, there are two amplifiers working in their nonlinear region; however, the input amplitudes for these two amplifiers are constant, and only their phases are changing. This method is mainly depending on the following trigonometric identity, so that the signal linearity will be kept [17].

$$\cos(A) + \cos(B) = 2\cos\left(\frac{A+B}{2}\right).2\cos\left(\frac{A-B}{2}\right) \tag{4.1}$$

Another efficiency technique is the envelope tracking (ET), where the power amplifier supply voltage will be changed according to the envelope of the input signal. In this case, the envelope of the baseband signal should be known; however, this technique suffers from complexity and the bandwidth limitations.

The third technique is the envelope elimination and restoration (EER), which was invented by Leonard Kahn. In this technique, there are two amplifiers: the first one is a highly linear amplifier used to amplify the signal envelope, whereas the second amplifier is a nonlinear amplifier used for amplifying a constant level of input signal. The main difference between the ET and EER is the following: the ET approach has a supply voltage that is adjusted to reduce the dissipated power (unused power), whereas in the EER, the supply voltage is used for shaping the output waveform. The final technique is the Doherty technique which will be discussed in the following subsection, where its main advantages are lack of complex circuitry and simplicity.

4.3.2 Load Modulation Amplifier Principles

W. Doherty, in 1936, came up with a new concept for combining two amplifier outputs; Doherty used a quarter-wavelength transmission line and two tube amplifiers. The first amplifier is called the carrier amplifier, whereas the second is referred to as the peaking amplifier. Whereas the first PA operates continuously in Class AB, the second amplifier operates only during the load modulation region in Class C. For the impedance inverter, a quarter-wavelength transmission line was used to invert the impedance amount seen by the carrier amplifier as illustrated in Fig. 4.5.

The characteristic of a transistor can be explored by applying different gate voltages and checking the behavior of the drain current, as illustrated in Fig. 4.6. The power amplifier load impedance can be determined using Eq. 4.2 assuming that the drain parasitic of the transistor is disregarded:

$$R_{\mathrm{opt}} = 2\frac{V_{dd} - V_{\mathrm{knee}}}{I_{\max}} \tag{4.2}$$

The Doherty amplifier relies on the load modulation technique, where its operations can be summarized by two regions of operation assuming both amplifiers are connected to a load of 25 Ω and matched to 50 Ω.

In the low input power region, the carrier amplifier is working; nevertheless, due to the quarter-wavelength transmission line impedance inversion property, the impedance seen by the carrier amplifier is 100 Ω, which is double the optimum

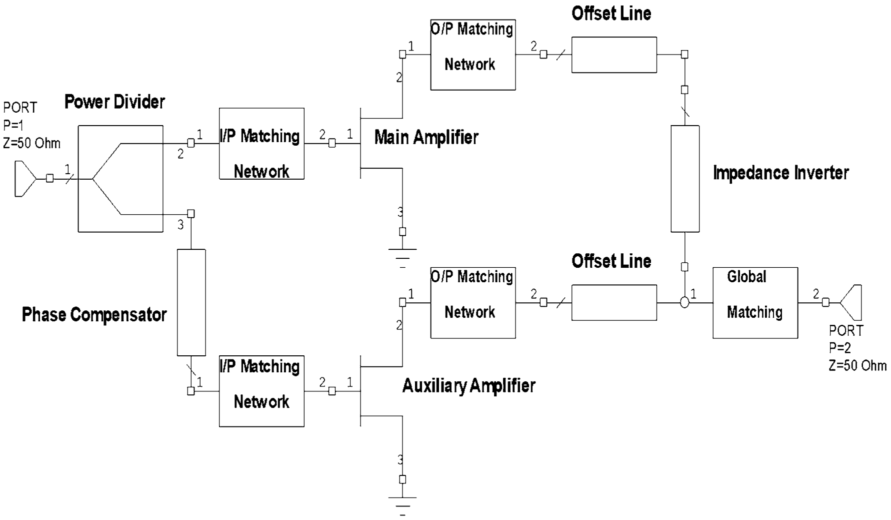


Fig. 4.5 Load modulation structure [38]

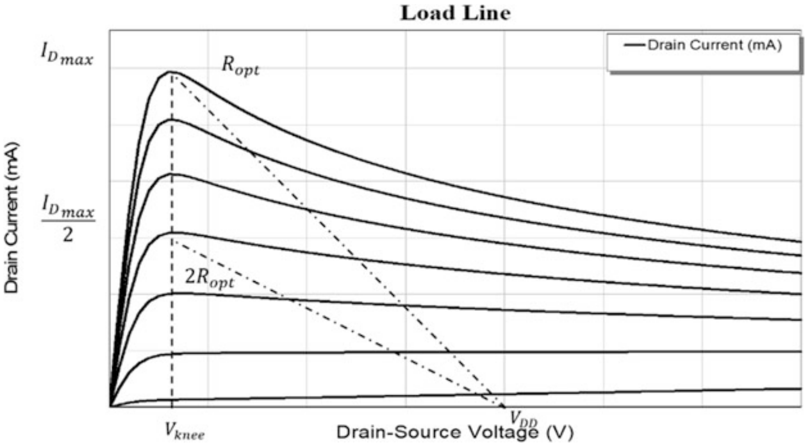


Fig. 4.6 Load line of a transistor including self-heating effect [38]

load seen by the carrier amplifier; so the first peak efficiency will be seen due to the carrier amplifier saturation.

In the load modulation region, the Doherty operation will be clear, where a current will be injected to the summing node by the peaking amplifier, so two currents contribute to the same load. There will be a reduction from a 100 Ω to 50 Ω of the impedance seen by the carrier amplifier depending on the peaking amplifier current. However, the saturation of the carrier amplifier continues until its maximum power.

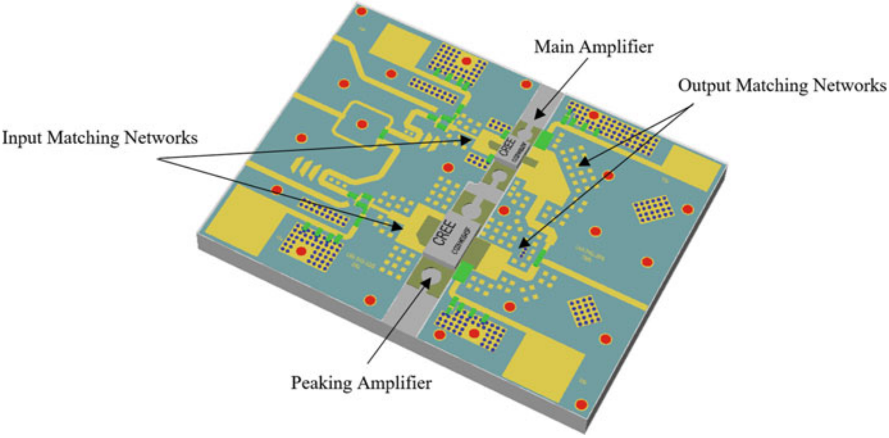


Fig. 4.7 Compact load modulation design

4.3.3 Efficient Load Modulation Amplifier Design: 5G Case Study

One of the main challenges in designing current and future power amplifiers is size, in addition to the other key requirements that are essential for each standard. In this design, a compact load modulation amplifier that can be used for 5G base station will be presented. As mentioned in the previous section, the normal Doherty amplifier can be designed using two amplifiers with the quarter-wavelength transmission line that acts as an impedance inverter. Figure 4.7 shows the designed 5G power amplifier, where this design represents a compact power amplifier, where there is no quarter-wavelength transmission line between the two amplifiers; the main idea of the design is to create a virtual transmission line that has been embedded in the design. This amplifier works for sub-6 GHz bands specifically at a center frequency of 3.6 GHz with a bandwidth of 400 MHz. The amplifier consists of two transistors, which are CG2H40025 and CG2H40045 transistors, for the main and the peaking amplifiers, respectively. The main amplifier can deliver a maximum power of 25 W, whereas the peaking amplifier can deliver a power of 45 W, so both amplifiers can provide a peak power of 70 W.

Figure 4.8 shows the performance of the designed amplifier – it can be seen that the amplifier delivered an average gain of 10.5 dB with a peak power of 48 dBm. Additionally, it can be noticed that the amplifier achieved an average peak efficiency of 70% with a back-off efficiency of 50% at 40 dBm output power. One of the main challenges in this design was the circuit design complexity, where too many parameters are required in designing each amplifier matching network. The amplifier performance is very promising, where it can be used for a 5G base station since it can provide high efficiency at the back-off region, and this amplifier can be used to amplify a modulated signal with high PAPR [41].

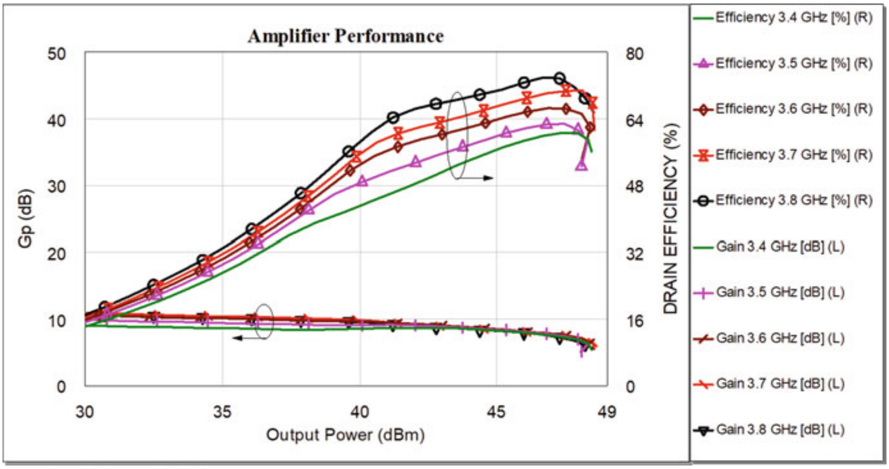


Fig. 4.8 Designed load modulation amplifier

4.4 Differentially Fed Filtering Antenna Design

With the fast development of current wireless communication systems, fourth-generation (4G) (2.6–2.7 GHz) and sub-6 GHz fifth-generation (5G) bands have been proposed for several RF planar circuits including power amplifiers, filters, and antennas [42, 43]. RF components using differentially fed ports are in high demand for 4G and 5G applications. Also, differentially fed RF structures offer promising characteristics for 5G applications and can provide some important and attractive properties [44, 45]. The main essential advantages for differentially fed RF circuits are:

- Multi-functional property
- High filtering level to the interference signals
- High common-mode suppression
- High roll-off skirt selectivity
- Wideband harmonic rejection
- Low radiated power loss
- Low cross-polarization

Nowadays, differentially fed components with dual-polarization performance have been commonly used to improve the radiation characteristics of the entire RF front-end applications [46]. Several differentially driven antenna configurations have been proposed, such as uni-planar microstrip antennas [47, 48], magneto-electric multi-dipole structures [49], and cavity-backed antennas [50]. In [46], a differentially fed microstrip antenna fed by orthogonally phased ports with 0° and 180° signals was reported. The antenna has a high realized gain of about 11 dBi and low cross-polarization of −18 dB in the E-plane. It also provides a 10 dB fractional bandwidth of 16% at 13.2 GHz for Ku-band systems. Apart from the ordinary differentially fed planar structures, the configurations presented in [47] and [48] apply a folded plate pair to construct the differentially driven signals. The reported antennas have a stable maximum gain of about 9 dB at the center frequency and over the entire bandwidth which leads to the high linearity in the radiation and symmetrical characteristics at the dual mode of operation. A differentially excited microstrip antenna with a maximum realized gain of 9 dBi and a 140 MHz impedance bandwidth was reported. In [49], the introduced antenna can work with several systems including other RF differentially fed/balanced elements, energy harvesting, and radiofrequency identification (RFID) applications.

On the other hand, planar filter-antenna integrations (filtering antennas) can also be used to improve the entire performance of the RF front-end systems [51–53]. The integrated design of a filter-antenna combination utilizing a multi-layered substrate was reported for recent and future RF front-end systems including 5G applications. The configuration is a third-order ring open-loop resonator connected to a T-shaped strip radiator. The multi-layered technology is applied to achieve a compact size structure. The filtering antenna design uses a Rogers RT5880 dielectric material in the middle layer with a permittivity of 2.3 and thickness (h) of 0.581 mm. The designed filtering antenna (filtenna) works on 2.5 GHz and has a fractional bandwidth of about 3.0% and a realized gain of about 2.1 dB at the center frequency. Although the design has a compact size, it has a complex configuration due to the use of a multi-layer technology. While in [52], the proposed filtenna was tracked with a similar design procedure and obtained good scattering parameter and radiation results with a circular polarization performance. Nevertheless, the reported design utilized different design techniques using the substrate integrated waveguide structure.

4.4.1 Reconfigurable and Tunable Filtering Antenna Structures

By integrating the filtering antenna with the reconfigurability/tenability function, more attractive combination with efficient performance can be obtained [54, 55]. Conversely, this will also lead to some more challenges in simulation, fabrication, and measurement settings. Two PIN diodes and four varactor diodes are used in the reconfigurable/tunable design in [54]. Although this requires a more complicated configuration, it also leads to a very small size structure with good filtering performance and flexible tuning capability. In [55], tunable/reconfigurable band-pass characteristics were obtained by implementing a filtering antenna structure. The design offers a high degree of freedom to control the scattering parameters and the radiation patterns using a small size layout. Therefore, this multi-function design has been proposed for several 5G front-end systems. In addition, the main types of switching techniques that can be used in the biasing circuits of the tunable/reconfigurable RF are explained in Table 4.4 [56].

Table 4.4 RF reconfiguration techniques [56]

Properties	PIN diode	Varactor	RF MEMS	Photoconductive
Speed (μs)	1–100 × 10^{-6}	0.1	1–200	3–9
Quality factor	50–85	25–55	86–165	–
Voltage (V)	3–5	0.1–15	20–100	1.8–1.9
Current (mA)	3–20	1–25	0	0–87
Power (mW)	5–100	10–200	0.05–0.1	0–50
Temperature sensitivity	Medium	High	Low	Low
Cost	Low	Low	Medium	High
Loss at 1 GHz (dB)	0.3–1.2	0.5–3	0.05–0.2	0.5–1.5
Fabrication complexity	Commercially available	Commercially available	Low fabrication complexity	Complex

4.4.2 *Differentially Fed Reconfigurable Filtering Antenna Design*

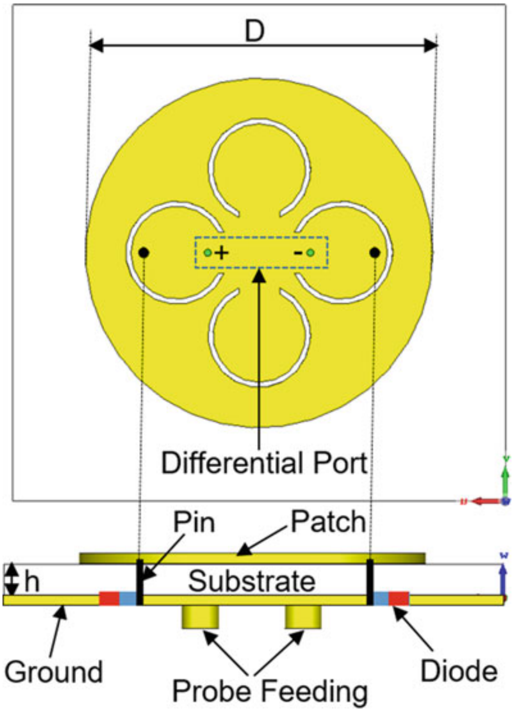
4.4.2.1 Configuration

Figure 4.9 illustrates the configuration of the designed differentially fed reconfigurable planar filter-antenna. The designed microstrip filtering antenna consists of a circular disk radiating patch with a diameter (D = 30 mm) and pair of differential feeding probes. It should be noted that loading the open ring slots with different configurations to the radiating element can reshape the distribution of the surface current or generate another frequency mode of operation. Thus, the structure size can be reduced, and the entire performance will be modified and developed.

In this design, four open-ring resonators are etched on the radiating layer to obtain broadside patterns with nulls at both sides of the passband transmission, resulting in a filtering performance with a wide-stopband suppression. The total circumference of each ring is about half the wavelength of the center frequency (λ/2). To achieve the reconfigurability property, two PIN diodes are placed on the ground layer to control the current distribution, therefore providing two resonance frequencies for 4G and 5G spectrums. Just to prove the design topology, practical PIN diode switches, SMP1320-079 from Skyworks Solutions Inc., were utilized each with a size of 1.5 × 0.7 mm^2. In computer simulation technology, CST microwave studio, these diodes are modeled with a lumped element circuit which presents 1.0 Ω as the resistance value of the PIN diode in the ON configuration (forward basing) and 0.5 pF as the capacitance value in the OFF configuration (reverse biasing).

Moreover, in the OFF configuration, the PIN diode can be designed as a series capacitance with a parasitic inductance as shown in Fig. 4.10, while, in the ON configuration, the PIN diode can be modeled as a series resistance with a parasitic inductance. The parasitic inductance was generated due to the package behavior of

Fig. 4.9 Top and side views of the designed reconfigurable filtering antenna



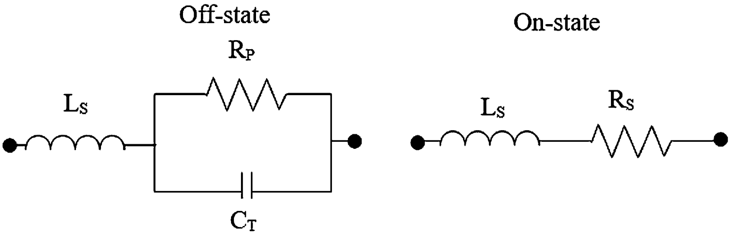


Fig. 4.10 RF equivalent circuit model for the PIN diode

the PIN diodes. The model equivalent circuit parameters can be obtained from the manufacturer's data where $L_S = 0.7$ nH, $C_T = 0.3$ pF, and RS = 1 Ω. As the value of Rp is higher than the reactance of C_T, it can be neglected in the equivalent circuit.

4.4.2.2 Filtering Antenna Analysis and Performance

Reconfigurable filter-antenna integration (filtenna) characteristics in regard to the reflection coefficients and peak realized gain for both diodes' configurations (ON and OFF) are presented in Figs. 4.11 and 4.12, respectively. The gained performance proves that the filtering antennas resonate at 2.6 GHz and 3.5 GHz with peak realized

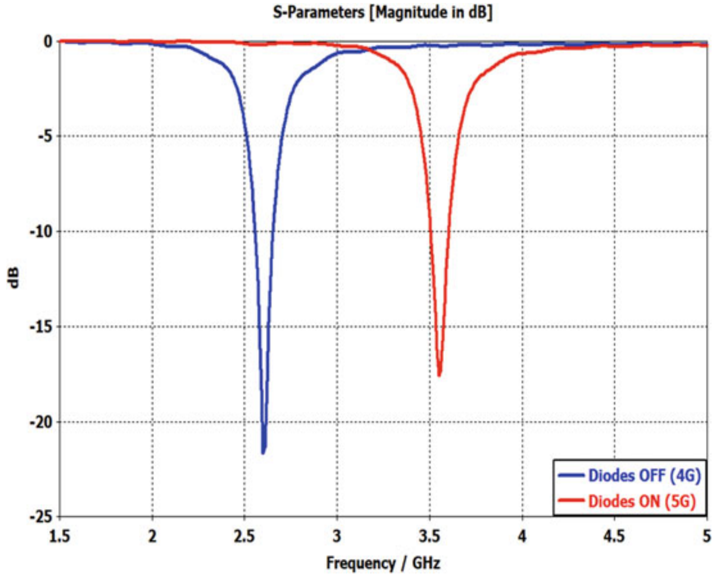


Fig. 4.11 S-parameter results of the proposed reconfigurable filtering antenna

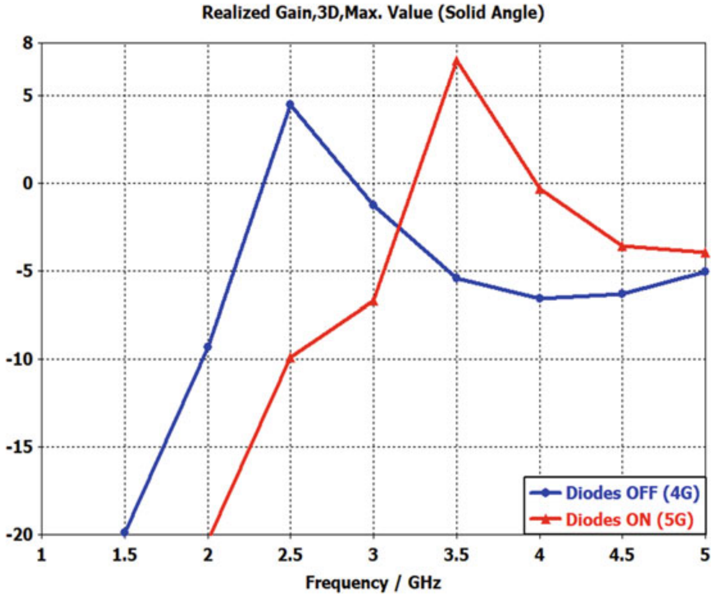


Fig. 4.12 Realized gain results of the proposed reconfigurable filtering antenna

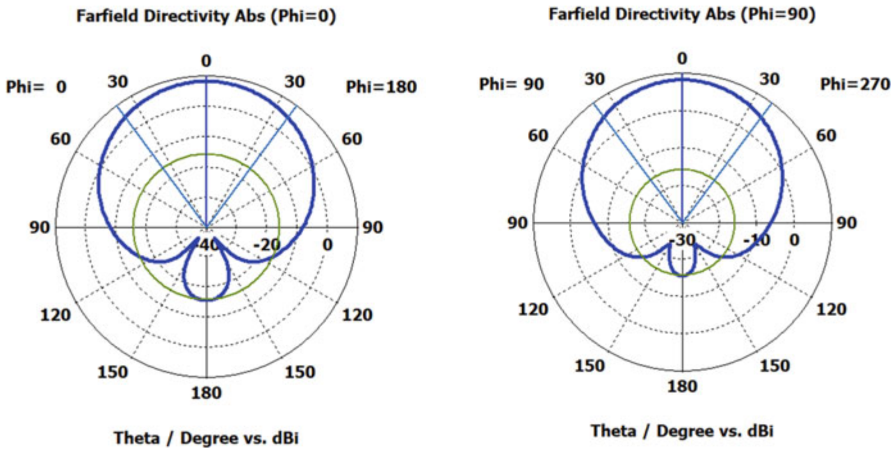


Fig. 4.13 2D radiation pattern characteristics at diodes-OFF configuration

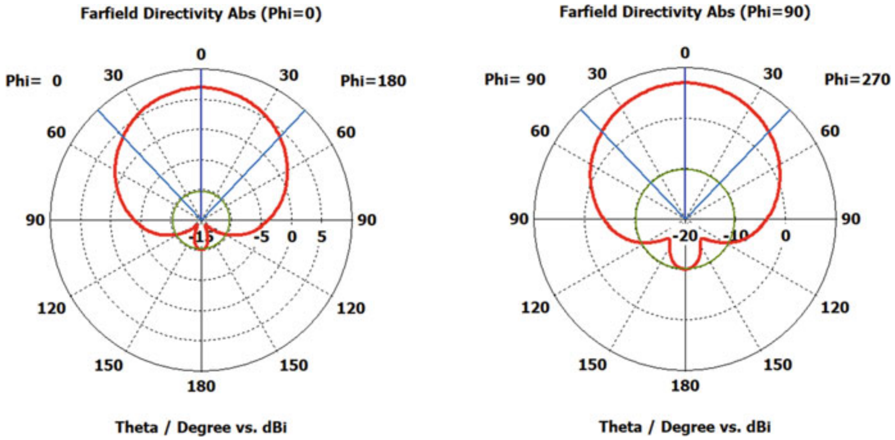


Fig. 4.14 2D radiation pattern characteristics at diodes-ON configuration

gain 5 dB and 7.5 dB at OFF and ON configurations, respectively, with impedance bandwidth 100 MHz.

Also, Figs. 4.13 and 4.14 show the radiation pattern characteristics for the proposed reconfigurable filtering antenna in diodes-OFF and diodes-ON configurations, respectively. In diodes-OFF configuration (Fig. 4.13), the operating frequency is 3.6 GHz, and the main lobe magnitude is 8.13 dBi. The angular width (3 dB) is 73.50, and the side lobe level is −23.8 dB. On the other hand, in diodes-ON configuration (Fig. 4.14), the operating frequency is 2.65 GHz, and the main lobe magnitude is 7 dBi. The angular width (3 dB) is 86, and the side lobe level equates to −17 dB. Figure 4.15 shows the 3D simulation results for the radiation pattern directivity that correspond to the two configurations of the diode switch in both forward and reverse bias.

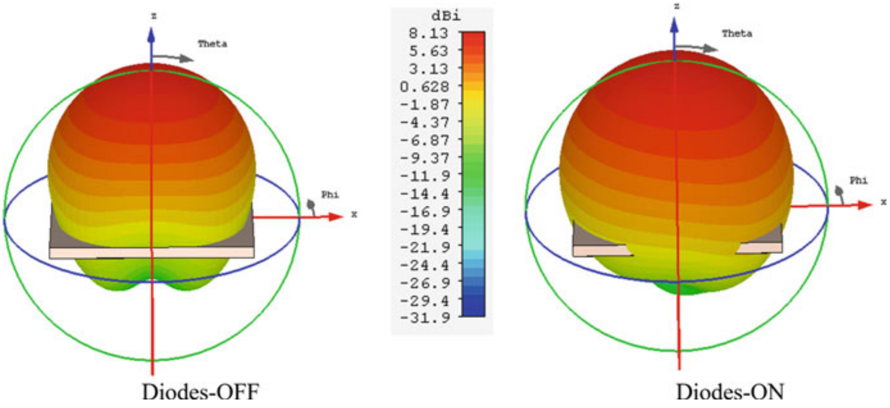


Fig. 4.15 3D radiation pattern characteristics of the proposed filtering antenna

4.5 Phased Array Antenna Design for 5G mmWave FDD Backhaul

The 5G access network targets millimeter-wave (mmWave) frequencies for enabling high data rates required by the enhanced mobile broadband services. This in turn leads to a huge densification of base stations operating at mmWave frequencies, where the cell size is vastly small in order to overcome the high propagation losses. This densification gives the name to hyperdense scenarios.

On the other hand, the 5G fronthaul/backhaul network is expected to be fiber-optic based. However, a fully optical network is not feasible in these hyperdense scenarios, first due to the high cost that such deployment would represent, as well as fiber installation restrictions. These restrictions may be natural, but also due to local environmental policies. Finally, the deployment time represents also a big disadvantage compared to its counterpart: radio links. Therefore, this leaves sufficient margin for current and future point-to-point wireless links. As an improvement over current architectures, the new wireless fronthaul/backhaul network is expected to be reconfigurable to satisfy the dynamic nature of the mobile traffic. Phased array antenna (PAA) in the backhaul equipment is a promising solution not only to respond to this need but also to provide the equipment with the ability to automatically recover the link when misalignments occur, increasing its availability. However, there are some challenges to overcome before using it in commercial equipment [57, 58], especially those that use frequency division duplexing (FDD) for transporting bidirectional data streams.

The following section focuses on a custom-made design for a compact filtering antenna to serve as the base radiation element for a PAA expected to be fitted in SIAE MICROELETTRONICA backhaul equipment [59], which uses FDD as a duplexing technique and works at 25 GHz.

4.5.1 Phased Array Antenna Architecture for FDD Applications

SIAE's equipment as well as most wireless backhauling equipment uses FDD for transporting bidirectional data streams. It means that two different frequencies are used for transmission (Tx) and reception (Rx). In the current architecture of the equipment, the Tx and Rx chains are connected to the antenna through a diplexer, ensuring 65 dB of isolation between chains, as shown in Fig. 4.16a. In the 25 GHz band, the lower-frequency band spans from 24.549 GHz up to 24.997 GHz, while the upper-frequency band ranges between 25.557 and 26.005 GHz.

In a first attempt, one could think of extrapolating this architecture to that of a PAA, as shown in Fig. 4.16b. Note that in this architecture it is still necessary to guarantee 65 dB of isolation between the Tx and Rx chains. Also note that according to the antenna array theory [60], the separation between adjacent antennas should be around half of wavelength (~ 6 mm at 25 GHz) to minimize unwanted grating lobes which appear especially when electronic beam steering is applied. Then, designing this highly selective diplexer that can easily be integrated with the rest of the PAA components, and in such a small space, represents quite a big challenge. Consider that the current diplexer is the union of two eight-pole waveguide bandpass filters with physical dimensions in the order of centimeters. For these reasons, the architecture that seems most viable is the one shown in Fig. 4.15c. With two different arrays, one for Tx and the other for Rx, a natural isolation appears. This isolation is dependent on the separation distance between the arrays and on any

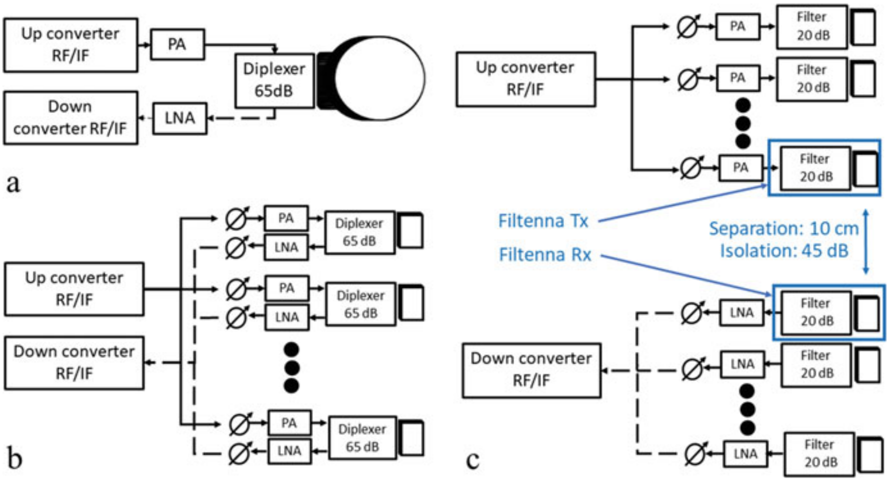


Fig. 4.16 Front-end architectures for backhaul equipment using FDD duplexing: (**a**) current architecture, (**b**) architecture with a PAA using a single antenna array, and (**c**) architecture with a PAA using two antenna arrays, one for Tx and the other for Rx. PA is power amplifier in the Tx chain. LNA is low-noise amplifier in the Rx chain

absorbing material or structures placed between them. By initially neglecting the latter possibility, full-wave simulations of two arrays of patch antennas at different distances show that a reasonable separation distance of 10 cm guarantees a natural isolation of around 45 dB between the Tx and Rx chains. However, 20 dBs are still missing to comply with equipment specification. Consequently, a filter with lower order is still required between the antenna and the active component (i.e., LNA or PA) in both Tx and Rx with the aim to provide the remaining 20 dBs of isolation. For better integration, two filtennas (filter + antenna) have been designed to provide this isolation, one for Tx and the other for Rx chain. In an FDD system, a filter is required on the Tx chain to remove the unwanted wideband noise from the Tx signal that would cancel out the low-level received signal on the Rx band. On the other hand, in the Rx chain, the filter is required to protect the LNA from the high-power Tx signal, which could saturate it and consequently distort the received signal. Thus, a filtenna in each chain is needed. Note that the filtennas are now the base radiation element of the PAA, and since the design procedure is the same for both, the following section will focus only on the Tx frequencies.

4.5.2 Filtenna Design Leveraging Filter Synthesis

There are different methods to design a filtenna [61]. First, we have the traditional one, in which the antenna and filter are designed separately and then a matching network is designed to connect them, like the work in [62]. Second, a co-design approach can be followed, in which a filter response is obtained by embedding filtering structures into different parts of the antenna (e.g., slots, split-ring resonators), trying not to alter the radiation performance, like the work in [63, 64]. The main disadvantage of the first approach is that the matching network could introduce additional losses to the circuit and increase the footprint of the filtenna. In the second approach, the main issue is that the design procedure is mainly based on full-wave optimization, which is time-consuming. Finally, a filter synthesis-based approach can be also followed, in which the antenna can behave like the last stage of the filter (i.e., resonator and load conductance or only the load conductance), like the works detailed in [65] and [66]. This last approach has the advantage that the antenna is part of the filter; therefore, there is no need for a matching network. Also, mostly all the design procedure is systematic, also supported with full-wave optimization at the end. Additionally, it allows to know in advance the expected S_{11} with a fair precision. For these reasons, this work is dedicated to this approach.

Within the last approach, two techniques can be followed: a lossless synthesis or a lossy one. The first one assumes that all the resonators of the filter are lossless, which fits quite well in air-filled waveguide filters. The second instead considers the lossy nature of the resonators when they are implemented in dielectrics. In this work, the filtenna was designed in the dielectric (PCB technology) because of the available space for the design and for a better integration with the rest of the equipment. Thus, a lossy technique seems the most adequate.

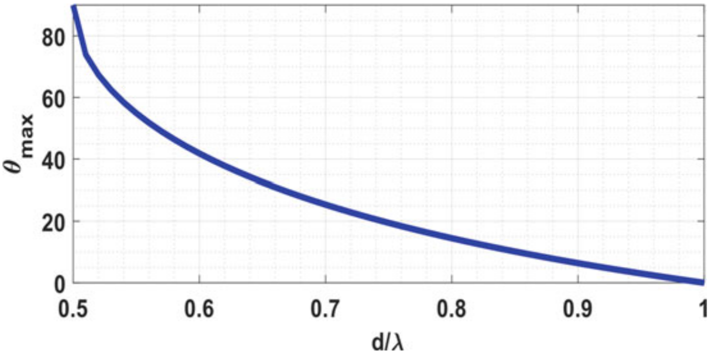


Fig. 4.17 Maximum steering angle ($\theta_{\max}$) without grating lobes versus normalized separation distance (d/λ) between radiation elements in a uniformly spaced linear antenna array. $\theta_{\max} = a\sin\left(\frac{\lambda}{d} - 1\right)$ [60]

In general, lossless techniques work very well for filtennas; in fact, most of the previous synthesis-based works were done with this technique, even in PCB. However, since the lossless assumption does not agree with the real resonators (lossy), the flatness of the transmission parameter (gain in the filtenna case) in the passband is lost. In contrast, when lossy synthesis is used, flatness of the gain in passband is preserved. To show the differences, the two techniques are discussed in this work. To the authors' knowledge, the work presented here is the first lossy synthesis-based design of a filtenna in the literature.

Finally, let us summarize the design requirements and constraints. Since the design is focused on Tx, the passband extends from 24.549 GHz to 24.997 GHz. In the stopband, extending from 25.557 GHz to 26.005 GHz, the gain must be 20 dB less than in the passband. The return loss (RL) should be better than 10 dB. Since the target is a PAA, the size of the filtenna directly affects the spacing between the radiation elements; this in turn affects the maximum angle at which the main beam can be steered without the appearance of undesirable grating lobes [60], as shown in Fig. 4.17. Note that filtenna size is constrained to be less than a wavelength (λ~12 mm at 25 GHz); otherwise, grating lobes would appear even without beam steering. Of course, the closer to λ/2, the better.

4.5.2.1 Lossless Filter Synthesis Technique

The procedure starts by synthesizing a lossless filter prototype that fulfills the specifications described above by using any of the methods available in the literature [67]. A third-order all-poles Chebyshev filter fulfills these specifications with RL = 20 dB in the passband. The de-normalized synthesized circuit is shown in Fig. 4.18a. Because of the small space available, a vertical stacked architecture was exploited [68, 69] as shown in Fig. 4.18b. Each resonator was physically

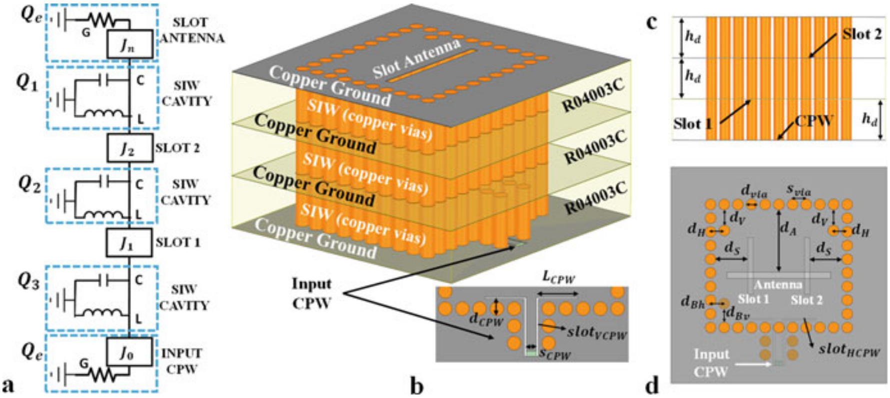


Fig. 4.18 (**a**) Synthesized lossless filter highlighting the relations with the designed filtenna. (**b**) Designed filtenna, 3D view. (**c**) Designed filtenna, 2D front view. (**d**) Designed filtenna, 2D top view. Circuital values: $C = 6.425\ pF$, $L = 6.425\ pH$, $G = 1\ \Omega^{-1}$, $J_{0,\ n} = 0.1456$, $J_{1,\ 2} = 0.01863$, $Q_e = 47.19$, $Q_{1,\,2,\,3} = \infty$. Physical dimensions in mm: $d_{via} = 0.4$, $s_{via} = 0.506$, $d_V = 0.565$, $d_H = 0.512$, $d_S = 1.188$, $d_A = 2.303$, $d_{Bv} = 0.684$, $d_{Bh} = 0.5$, slot$_{VCPW} = 0.06$, slot$_{HCPW} = 0.042$, $s_{CPW} = 0.3$, $L_{CPW} = 1.16$, $d_{CPW} = 0.491$, $h_d = 1.524$, $h_{\text{ground}} = 0.017$. Antenna slot: 3.85 x 0.235. Coupling slots: 2.1 x 0.23

implemented with a substrate integrated waveguide (SIW) cavity. Note from Fig. 4.18a that all the cavities must have the same resonant frequency (F = 24.772 GHz) and were initially dimensioned by using the well-known design formulas for SIW cavities [70]:

$$\begin{gathered} W_{eff} = W_{siw} - \frac{d^2}{0.95\ s}, \quad L_{eff} = L_{siw} - \frac{d^2}{0.95\ s}, \\ F_{TEmp0} = \frac{c}{2\sqrt{e_r}}\sqrt{\left(\frac{m}{W_{eff}}\right)^2 + \left(\frac{p}{L_{eff}}\right)^2} \end{gathered} \tag{4.3}$$

where d is the via diameter, s is the via separation, and $Wsiw$, $Lsiw$ are the cavity dimensions and $m = p = 1$ is used to address the TE110 resonant mode.

The coupling between resonators (inverter J_i in the equivalent circuit) were physically implemented with slots. The closer the coupling slots are to the edge of the SIW cavity, the greater the coupling. To determine the position of the slots, the relationship between the dimension d_S (Fig. 4.18d) and the coefficient of coupling J_i (Fig. 4.18a) was computed. This relationship is shown in Fig. 4.19a. This curve was made by coupling two SIW cavities, varying d_S, and calculating the coefficient J_i using Eq. 4.4 for each variation, where f_1 and f_2 are the two peak frequencies in the S_{21} parameter. The coupling with the source and load must be as low as possible, for this procedure to work properly [71].

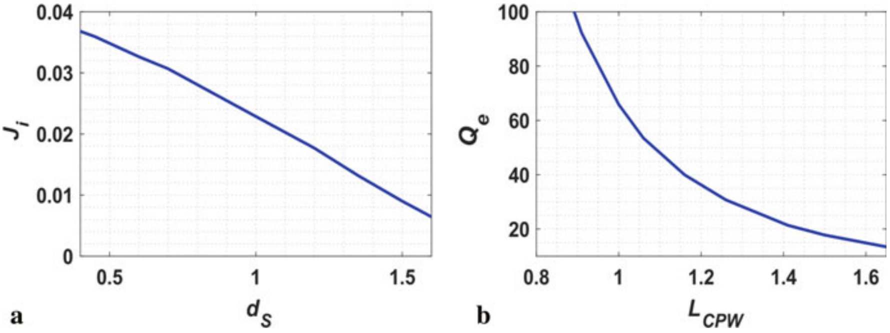


Fig. 4.19 (**a**) Relation between the dimension d_S and the coupling J_i. (**b**) Relation between the dimension L_{CPW} and the quality factor Q_e

$$J_i = \frac{{f_2}^2 - {f_1}^2}{{f_2}^2 + {f_1}^2} \tag{4.4}$$

To design the external quality factor (Q_e) at the input port, a 50 Ω coplanar waveguide (CPW) with two perpendicular slots was designed as shown in Fig. 4.18b. By varying the dimensions L_{cpw} and d_{CPW} in a short-circuited SIW, the Q_e can be computed by [71]:

$$Q_e = \frac{2.\pi.f_0.\tau_{s11}\,(f_0)}{4} \tag{4.5}$$

where f_0 is the peak of the group delay of the S11 parameter and $\tau_{s11}(f_0)$ is the group delay at f_0. The relation between the dimension L_{CPW} and Q_e is shown in Fig. 4.19b. The dimension d_{CPW} does not affect greatly Q_e, but it can be used for fine-tuning.

Note from Fig. 4.18a that the circuit is symmetric; therefore, the physical filtenna must also share this symmetric nature. This implies that the scattering parameter response of the input part (CPW + SIW) should be the same as the output part (SIW + antenna slot). Therefore, since the input part is already designed, the output part must be optimized to behave in frequency as the input part. Also note that in both input and output SIWs, there are vias inside the cavity. These vias help to tune the cavities in the central frequency. This is because the slots (antenna, CPW) make these cavities electrically bigger in comparison with the central cavity.

To the best of the authors' knowledge, the above procedures provide a very good initial design that would require only a few full-wave optimization steps. Final dimensions of the filtenna are written in the caption of Fig. 4.18a. The frequency response of both equivalent circuit and the physical filtenna is shown in Fig. 4.20c. Note the good correlation between the S_{11} of the synthesized circuit and the full-wave simulations of the physical filtenna.

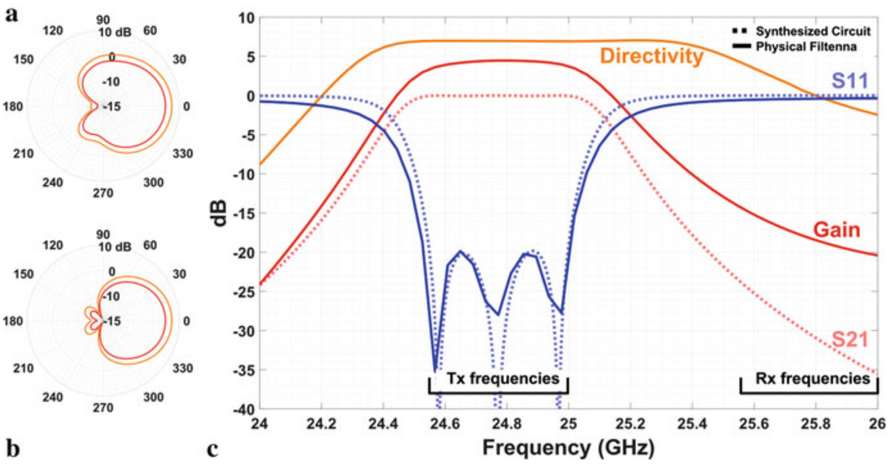


Fig. 4.20 Circuit and full-wave simulations of the designed filtenna. Radiation pattern at central frequency 24.772 GHz: (**a**) (cut $\varnothing = 0^{\circ}$), (**b**) (cut $\varnothing = 90^{\circ}$), and (**c**) scattering parameters of the synthesized lossless filter in dotted lines. S11, gain, and directivity response of the designed filtenna in solid lines

Focusing instead on the transmission parameter, note the flatness of the S_{21} of the synthesized lossless circuit throughout the band-pass, in contrast with the filtenna gain with 4.46 dBi in the center frequency and 3.73 dBi at passband edges. This is because of the losslessness assumption at the beginning of the design, which is not true in the physical filtenna. The difference between the gain at the central frequency of the Tx and at the beginning of the Rx band is 19 dB approximately, close enough to the desired (20 dB). To increase the isolation, the filter order could be increased, but this would also imply an increase in the complexity of the filtenna (additional layer in the stack-up) and a reduction in the antenna gain due to the increase in insertion loss. A better option would be to add complementary split-ring resonators between the Tx and Rx antenna arrays to get further isolation [72].

Figure 4.20a, b shows the $\varnothing = 0^{\circ}$, $\varnothing\ = 90^{\circ}$ cuts of the antenna radiation pattern, where the directivity and gain are about 6.96 and 4.46 dBi at boresight, respectively. Finally, according to the filtenna size, it is possible to get 6 mm ($\lambda/2$) of separation between radiation elements in the PAA. Therefore, according to Fig. 4.15, it would be possible to perform beam steering up to $\pm 90^{\circ}$ in both azimuth and elevation planes without the appearance of grating lobes.

4.5.2.2 Lossy Filter Synthesis Technique

The main difference with respect to the previous design lies in the synthesis technique. Here we have used the lossy synthesis described in [73]. The procedure starts by computing the rational polynomials defining a desired transmission and

Table 4.5 Polynomials throughout the process of the lossy filter synthesis

	$N_{11}(s) = N_{22}(s)$	$N_{21}\ (s)$	*Denominator* (s)
Lossless Chebyshev Polynomials ($RL = 22dB$)	$s^3 + 0.75s$	-3.1374	$s^3 + 2.5881s^2 + 4.0990\,s$ $+3.1374$
Scaling $-$ 2 dB	$0.7943s^3 + 0.5957s$	-2.4921	
Force $S_{11}(\infty) = S_{22}(\infty) = 0$ ($RL = 12$ dB)	$s^3\ (1 + 0.002i)+$ $s^2\ (0.532 + 0.006i)+$ s $(1.439 + 0.009i)+$ $(0.645 + 0.007i)$	-2.4921	

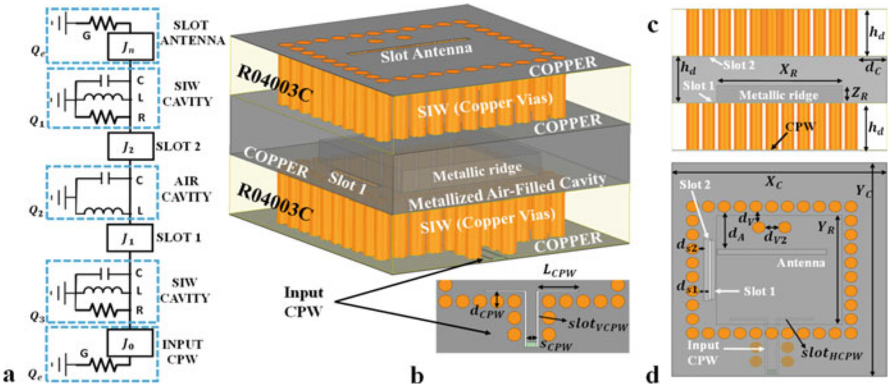


Fig. 4.21 (**a**) Synthesized lossy filter highlighting the relations with the designed filtenna. (**b**) Designed filtenna, 3D view. (**c**) Designed filtenna, 2D front view. (**d**) Designed filtenna, 2D top view. Circuital values: $C = 6.425\,pF$, $L = 6.425\,pH$, $R = 207.8\ \Omega$, $G = 1\ \Omega^{-1}$, $J_{0,\ n} = 0.1363$, $J_{1,\ 2} = 0.01991$, $Q_2 = \infty$, $Q_e = 53.8$, $Q_{1,3} = 207.8$. Physical dimensions in mm: $d_{via} = 0.4$, $s_{via} = 0.5$, $d_V = 0.439$, $d_{V2} = 0.42$, $d_A = 1.317$, $d_{S1} = 0.3$, $d_{S2} = 0.166$, $slot_{VCPW} = 0.06$, $slot_{HCPW} = 0.056$, $s_{CPW} = 0.3$, $L_{CPW} = 1.099$, $d_{CPW} = 0.276$. $X_R = 3.968$, $Y_R = 3.872$, $Z_R = 0.57$, $h_d = 1.524$, $h_{cooper} = 0.017$. Antenna slot: 3.451 *x* 0.187. Coupling slot 1: 2.1 *x* 0.205. Coupling slot 2: 2.194 *x* 0.225

reflection characteristic $S_{11}(s)$, $S_{21}(s)$, and $S_{22}(S)$, using, for example, the procedure of [74] to get Chebyshev filtering response. Then, all the numerators N_{11}(s), $N_{21}(s)$, and $N_{22}(s)$ are scaled with a suitable constant, according to the desired or expected insertion loss (−2 dB in the current design). Then, by using the even- and odd-mode decomposition, $N_{11}(s)$ and $N_{22}(s)$ are recomputed so that $S_{11}(\infty) = S_{22}(\infty) = 0$, which imposes a degradation to 12 dB of return loss in the passband. All the polynomials of all these three steps are detailed in Table 4.5.

With these new polynomials, the normal coupling matrix synthesis described in [67] can be applied. The final de-normalized synthesized circuit is shown in Fig. 4.21 b,c,d. Note that in the resulting circuit, the first and last resonators are lossy, but the middle one is lossless. It is possible to make lossy the central resonator (to have all the cavities in the same dielectric) by matrix rotations, but this leads to obtaining lossy couplings, which are physically implemented by adding resistors which are not desired in our design for integration reasons (Fig. 4.21).

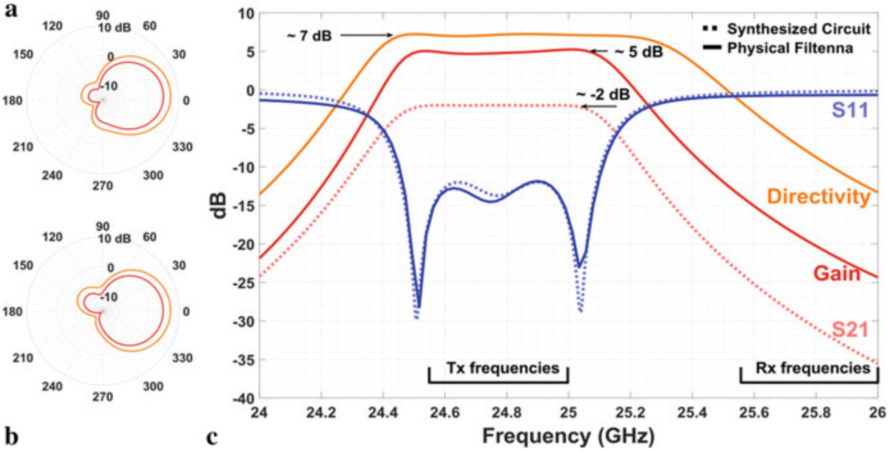


Fig. 4.22 Circuit and full-wave simulations of the designed filtenna. Radiation pattern at central frequency 24.772 GHz: (**a**) (cut $\varnothing = 0^{\circ}$), (**b**) (cut $\varnothing = 90^{\circ}$). (**c**) Scattering parameters of the synthesized lossy filter in dotted lines. S11, gain, and directivity response of the designed filtenna in solid lines

To physically implement this circuit as faithfully as possible, the middle cavity needs to be lossless: it can be achieved by an air-filled cavity (approximately lossless). New techniques have emerged to manufacture these cavities in PCB technologies, such as the one described in [75]. According to [75], the air-filled cavity can be modeled by a traditional waveguide cavity since the manufacturing procedure consists of digging the dielectric and then metalizing the walls. The first issue that arises in the design procedure is that the size of the air cavity is in the order of λ which considerably limits the beam-steering capabilities as detailed above. Therefore, to reduce the size, a metallic ridge is inserted inside the cavity. This ridge makes the cavity electrically bigger, which allows to reduce the size. Note that, by maintaining the cavity dimensions, as the ridge height increases, the cavity resonance frequency decreases. By using this rule and with the help of an eigenmode solver, the cavity can be tuned to our desired center frequency and with acceptable dimensions according to the constraints imposed by the antenna array theory. The first (CPW + SIW) and last part (SIW + antenna slot) of the synthesized lossy circuit can be physically dimensioned with the same design rules of the previous section, again finding that the closer the coupling slots are to the edge of the SIW or air-filled cavity, the greater the coupling.

Different views of the designed filtenna using a lossy synthesis technique are Fig. 4.22b-c, while the final dimensions are shown in the captions. Note in Fig. 4.22c the very good agreement between S_{11} response of the synthesized lossy filter and the full-wave simulations of the designed filtenna. Also note the flatness of the transmission parameters throughout the band-pass in both circuit (S_{21}) and filtenna (gain). Also note the coherence between the directivity D_{Fil} and gain G_{Fil} of the filtenna and the S_{21} of the synthesized lossy filter: $G_{Fil} \sim D_{Fil} + S_{21}$.

Figure 4.22a–b shows the $\varnothing = 0^{\circ}$, $\varnothing\ = 90^{\circ}$ cuts of the antenna radiation pattern. The directivity and gain are about 7 and 5 dBi at boresight, respectively. As in the previous case, the difference in dB between the gain at the central frequency of Tx and at the beginning of the Rx frequencies is 19 dB approximately, and complementary split-ring resonators can be placed between the Tx and Rx filtenna arrays to obtain further isolation [72].

The size of the designed filtenna is $6.8\ x\ 7.6\ mm^2$ approximately, but according to [75], via holes must surround the air cavity to ensure connectivity of the upper and lower ground planes. Despite the negligible impact of these vias in the frequency response of the design, they increase the total footprint of the filtenna by up to $7.8\ x\ 8.6\ mm^2$ ($0.72\ \lambda\ x\ 0.65\ \lambda$) approximately. This in turn limits beam steering up to $\pm 32^{\circ}\ x\ \pm 25^{\circ}$ in azimuth and elevation, respectively. This is achieved without the appearance of grating lobes according to Fig. 4.17.

It is important to remark that this design based on a lossy filter synthesis technique overcomes in terms of gain flatness in the passband the previous one based on a lossless technique. Finally, the same procedure can be applied to the design of the filtenna for the Rx frequencies.

4.6 Insensitive Phased Array Antenna for 5G Smartphone Applications

To support the increasing demand of high transmission rate for various fixed and mobile services, phased arrays with multiple antenna elements have been attracting much attention for next-generation (5G) networks [76]. Apart from the sub-6 GHz spectrum, 5G devices are also expected to cover the higher frequencies (beyond 10 GHz) where it can be carried out by employing phased array antennas [77, 78]. One of the challenges in designing phased arrays is the implementation and arranging of compact antennas with improved performances [79, 80]. In this study, a high-efficiency phased array antenna is proposed for 5G smartphone applications. Sufficient and quite good outputs have been achieved for the presented design. It also offers good performance in data-mode accounting for the hand phantom.

4.6.1 Beam-Steerable and Phased Array Antennas for 5G Smartphones

In cellular environments, the angle of arrival is likely to be distributed across the sphere. Therefore, high-gain phased array with wide scanning and improved radiation coverage is desirable for universal applications. The phased array contains multiple compact radiators arranged in planar or linear from and fed with different phase shifts [81]. The antenna radiation pattern can be steered electronically to

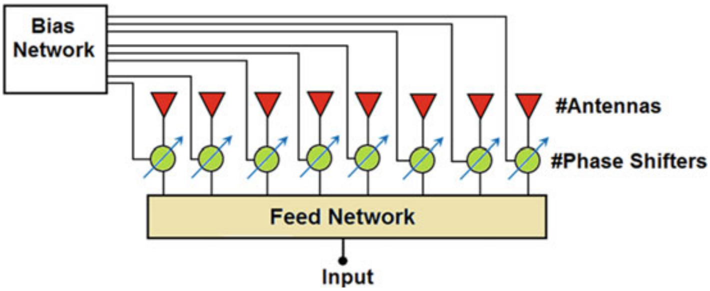


Fig. 4.23 Phased array architecture

Table 4.6 The values of the design parameters

Parameter	W_{sub}	L_{sub}	h_{sub}	W	W_1
Value (mm)	55	110	0.787	5.25	1.5
Parameter	W_2	W_3	W_4	L	L_1
Value (mm)	0.5	3.125	3.125	8.3	7.8

different scanning angles. Figure 4.23 shows a phased array architecture that can be used for the beam-steering purpose in linear array 5G smartphone antennas. The feeding network of the antenna arrays can be accomplished by employing cheap phase shifters. This might increase the complexity of the system but would improve the coverage and performance of the antenna array [82].

Compact antennas can be arranged in linear or planar array form to be used in phased array structures with high-gain characteristics for 5G wireless communications. Different from the conventional antennas (patch, monopole, and planar inverted-F antenna (PIFA) antennas) with omnidirectional radiation, the end-fire resonators, such as Vivaldi, Yagi, and linear tapered slot *antenna* (LTSA), are more suitable for the communication between user and base station in 5G mobile communications [83–85].

4.6.2 *Antenna Design Details*

The introduced phased array smartphone antenna is arranged on a cheap FR-4 dielectric with characteristics of $h_{\text{sub}} = 0.8$ mm, permittivity (ε_r) = 4.3, and loss tangent (δ) = 0.025. As illustrated in Fig. 4.24, eight substrate-insensitive antenna elements are based on a linear array arrangement that are placed in the top edge of the board with size of $W_{\text{sub}} \times L_{\text{sub}} = 55 \times 110$ mm^2. The arranged array has a low profile. For the beam-steering purpose, the distance among the antenna resonators ($W + W_2$) is chosen to be λ/2 of the resonance frequency. The EM simulation CST software was used for the investigation. The dimensions of the design parameters are specified in Table 4.6.

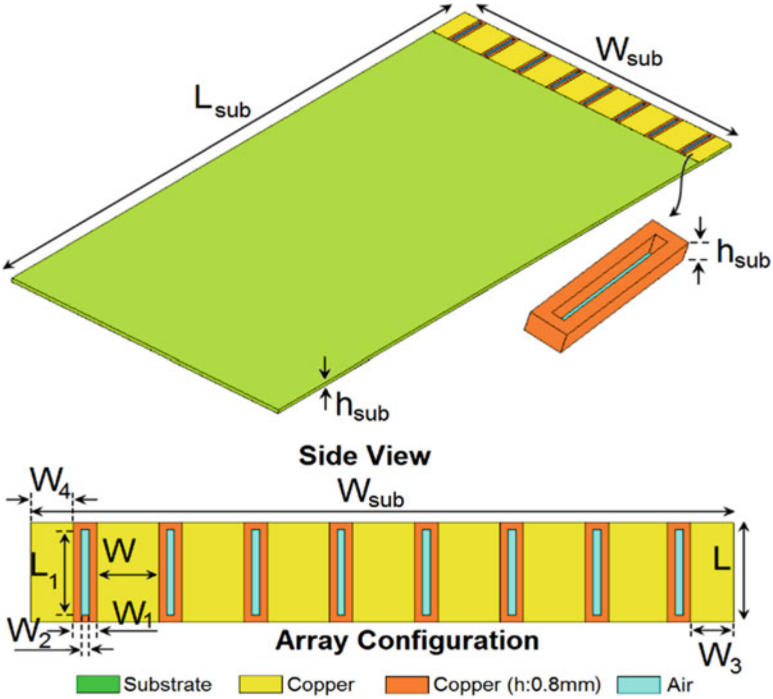


Fig. 4.24 Proposed 5G smartphone antenna configuration

4.6.3 *Characteristics of the Antenna Element*

Conventionally, the printed slot antenna contains a radiation element which is arranged by cutting a rectangular slot in a copper layer. Its length is about λ/2 and the width is a small fraction of a wavelength. The slot antenna is a complementary element of a dipole antenna with a different polarization mode. It has a low profile with a simple structure and flexible in nature and low cost in fabrication [86]. In the presented antenna design technique, as illustrated in Fig. 4.25, the resonator of a conventional slot structure with an operation band of 22–23.5 GHz is converted to a metal-ring loop resonator with the same thickness of the substrate. This not only improves the performance of the antenna but also eliminates the effect of the FR-4 dielectric which is lossy for higher spectrums. The discrete-port feeding technique is employed for the antenna excitation. The current density for the metal-ring slot resonator at its resonant frequency (22.25 GHz) is plotted in Fig. 4.26a. As expected, the employed metal-ring radiator has maximum densities and behaves highly active [87]. The 3D radiation of the metal-ring design is represented in Fig. 4.26b. It is found that the design offers a well-defined radiation covering both sides of the FR-4 dielectric. Besides, it provides a high potential gain of 5 dB.

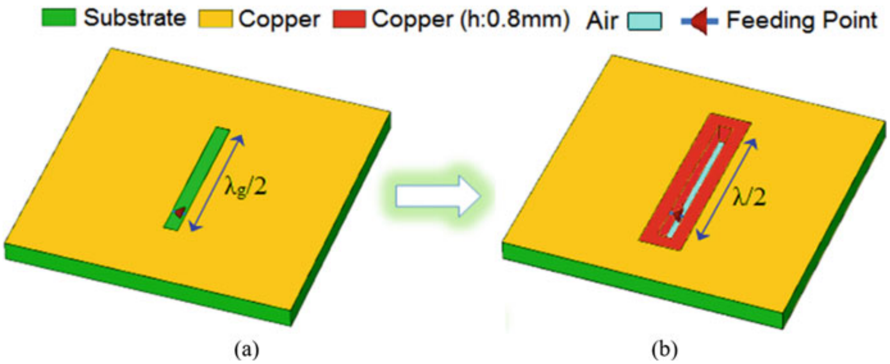


Fig. 4.25 (**a**) Conventional 22.25 GHz slot resonator, (**b**) the proposed insensitive antenna

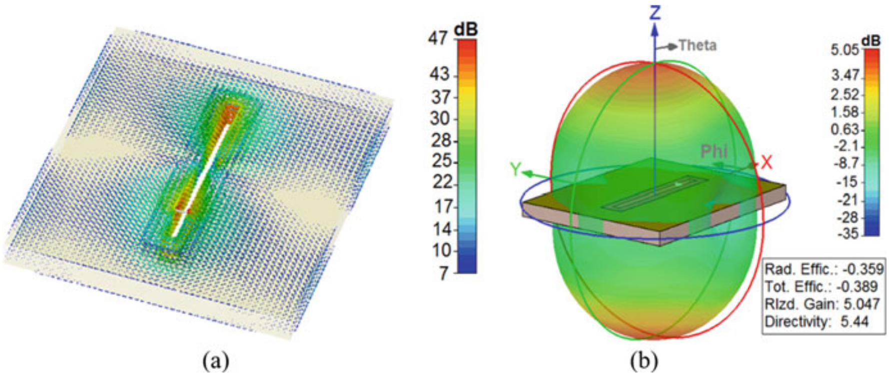


Fig. 4.26 (**a**) The surface current, (**b**) transparent radiation pattern at 22.25 GHz

4.6.4 Fundamental Properties of the Insensitive Phased Array Design

Fundamental properties of the proposed beam-steerable 5G smartphone antenna are described in this section. Figure 4.27 illustrates the scattering parameters (S_{11} ~ S_{81}). As illustrated, the design exhibits quite good S_{11} ~ S_{81} characteristics around 22.25 GHz. Moreover, low coupling ($S_{mn} < -15$ dB) is observed for the introduced array design. Figure 4.28 illustrates the Cartesian realized gains of the array different scanning angles. As seen, the antenna has good gain levels and covers a wide scanning range of ±70°.

Figure 4.29 illustrates the 3D beams of the designed array 5G antenna for various angles, where excellent radiation beams over the 0–70 scanning angles are exhibited. As shown, the design provides well-defined radiation beams at 0°, 15°, 30°, 45°, 60°, and 70° which could cover half the space of the radiation

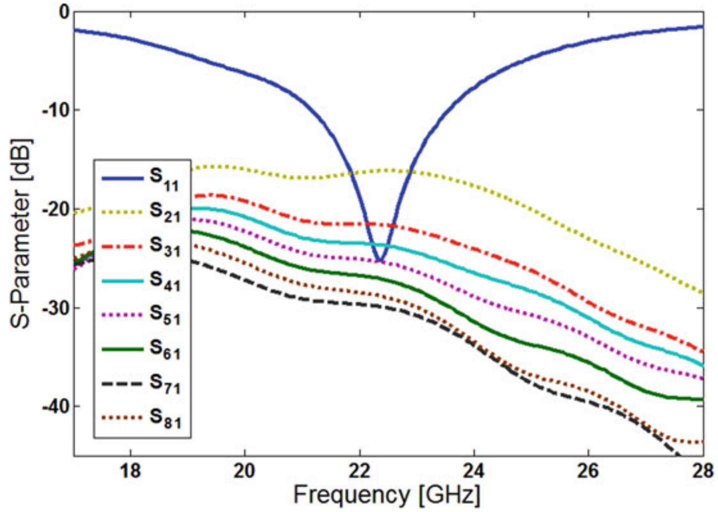


Fig. 4.27 S-parameters of the designed 5G phased array

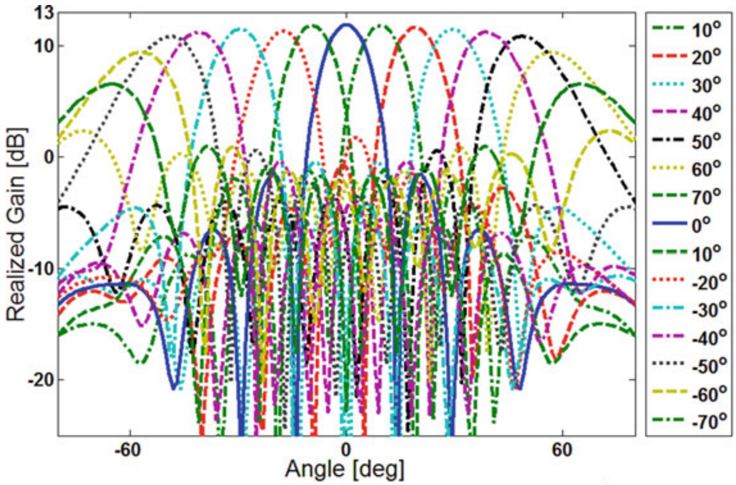


Fig. 4.28 Beam steering of the 5G phased array with 2D Cartesian gain values

coverage for the smartphone mainboard [88]. Fundamental properties of the design including directivity and efficiencies for the steered beams of the mobile-phone array at 22.25 GHz design are presented in Fig. 4.30. Across the range of 0°–60°, the efficiencies are greater than 90% (−0.5 dB), as well as provide sufficient maximum gain levels.

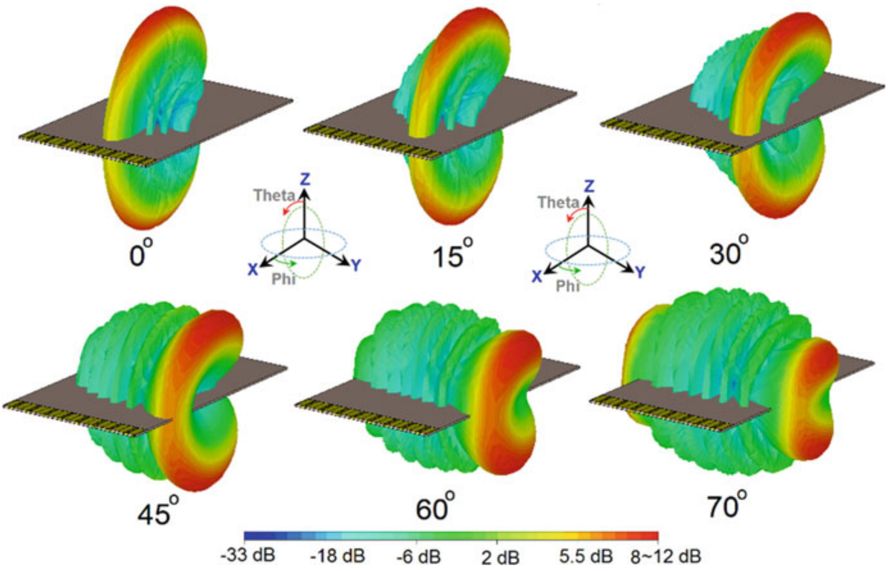


Fig. 4.29 3D radiation beams for 0–70 degrees

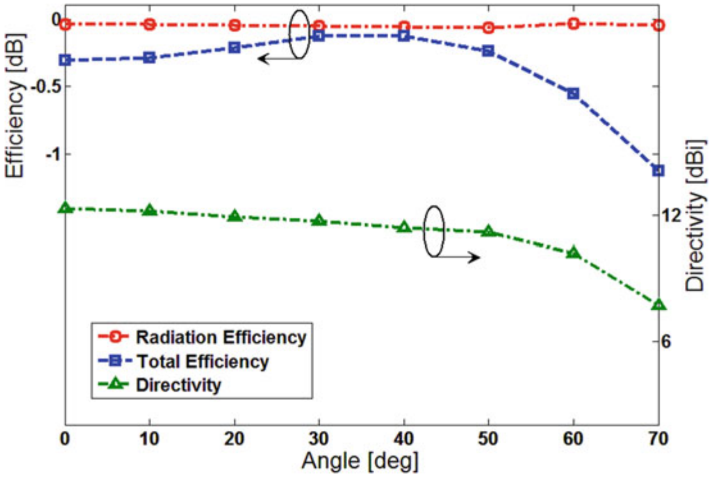


Fig. 4.30 Fundamental radiation characteristics

4.6.5 Insensitivity Characteristic of the Proposed Antenna

The design characteristics of the proposed phased array are insensitive to various substrate's properties. To understand this function, the coefficient reflection (S_{11}) results for different dielectric constant (epsilon: ε) are investigated in Fig. 4.31. In addition, the antenna gain and efficiencies (radiation and total) for different loss

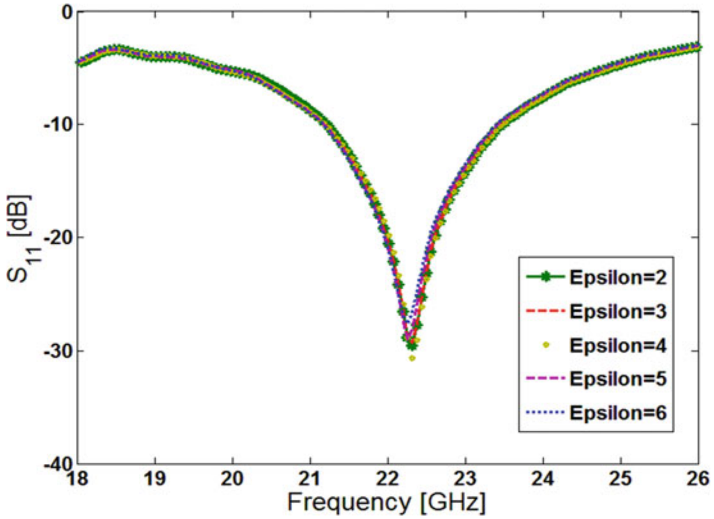


Fig. 4.31 The coefficient reflection (S_{11}) for various substrate types

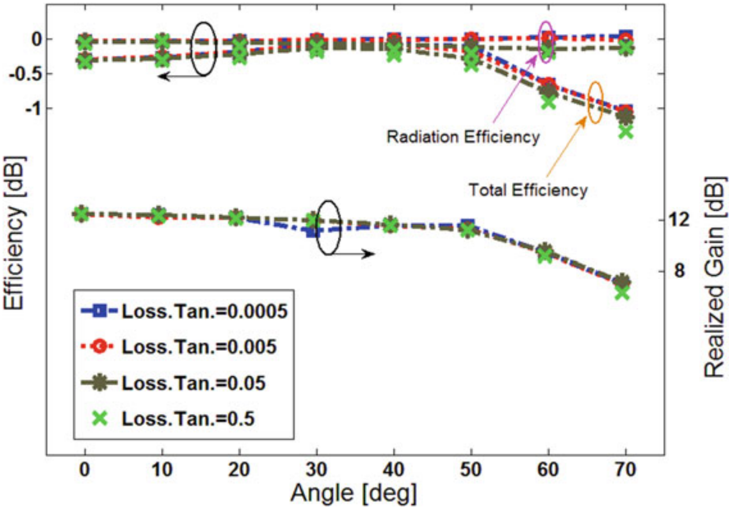


Fig. 4.32 Efficiency and gain results for different loss tangent (δ) values

tangent (δ) values are studied in Fig. 4.32. It should be noted the studied substrates have different loss tangent values which could affect the efficiency and gain of an antenna [89]. According to the obtained results in Fig. 4.31, it can be observed that the designed phased array is insensitive for different substrate types and exhibits similar behavior for different dielectric constants and loss tangent values.

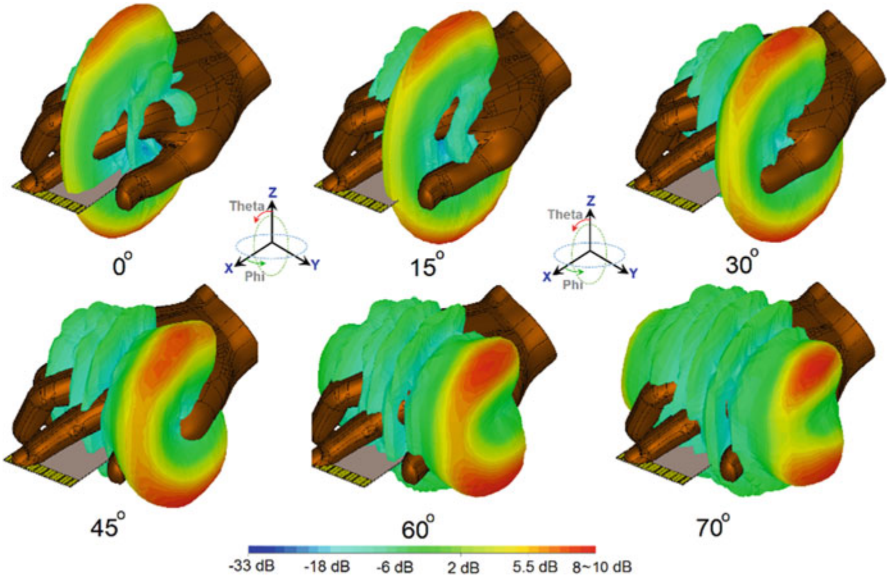


Fig. 4.33 Radiation beams of the design in the presence of the user-hand

4.6.6 User-Hand Impact on Antenna Performance

The user-hand is a body part that most frequently touches handheld devices and usually has negative impacts on antenna performance [90]. Figure 4.33 represents the 3D beams at various angels (0° ~ 60°) in the presence of a hand phantom (data-mode). As plotted, the beam-steerable phased array antenna provides well-defined radiation beams at various angels. This might be due to compact sizes and the insensitivity function of the employed elements, which are not highly affected by the user-hand.

4.7 Conclusion

This chapter has investigated the RF front-end for a typical 5G and beyond transceiver architecture focused on designing RF functional blocks with energy efficiency at the heart of the design. In this context, this chapter provided insights into the latest works on this topic to provide design recommendations on the PA, filter, and antenna units that are constituent components of an integrated architecture.

In this context, Sect. 4.2 provided the current state of the art in energy-efficient DPA MMIC design techniques for massive MIMO systems. The GaAs pHEMT process is a mature technology offering fast switching advantage that has led to widespread adoption in the cellular communication market. To address the limitation associated with the practical DPA implementations, we adopted a MMIC-based PA design to offer a compact approach taking a step toward reducing the circuit space in future emerging handsets. We proposed a novel GaAs pHEMT-based DPA based on incorporating harmonic manipulation. The newly presented Class-J DPA has demonstrated the potential not only to obtain high efficiency over wide bandwidths but also to achieve good linearity and back-off efficiency. It is worthy to note that the compact Class-J load modulation can be developed for 5G handset application without employing complicated circuits or predistortion linearization. Moreover, the authors investigated the load modulation technique in terms of a Doherty power amplifier design approach. The PA design targeted the 3.4–3.8 GHz bandwidth, where the PA was able to provide high efficiency of >75% at 48 dBm output power and back-off efficiency of 50% at 40 dBm; the design of this highly efficient compact power comes at the expense of circuit complexity for both the input and output matching networks, but still a good candidate for 5G base stations.

The design of a differentially driven reconfigurable planar filter-antenna integration with high-gain and high-common mode suppression is proposed for 4G and 5G RF front-end systems. Along with the frequency-reconfigurable characteristics, the design offers additional advantages including high-gain and radiation efficiency and low cross-polarization level due to the differentially fed terminals, as well as the accurate symmetry of the structure. The design constituted four open-loop ring structures that are used to realize the required nulls at both edges of the passband for the filtering performance. The introduced reconfigurable filtenna has a height of 0.82 mm and operates at the center frequencies of 4G and sub-6 GHz 5G. The presented filtenna is simulated, analyzed, and optimized using the CST tool. In addition, two filtenna synthesis techniques have been discussed, lossless-based and lossy-based, the latter being a better approach to preserve the flatness in the gain of the filtenna. Filtenna are ideal for phased arrays in commercial FDD backhauling radios while supporting several other positive features at the application level such as increased network flexibility, thanks to beam steering, and compatibility with software-defined network paradigms involving dynamic reconfiguration.

Moreover, an insensitive phased array antenna with air-filled slot-loop resonators is introduced for 5G mobile applications. It is designed on the FR-4 substrate and working at 21–23.5 GHz. Eight elements of the metal-ring elements are linearly arranged on the top of the phone PCB. The proposed design has shown good performance that takes into account the handheld effect due to its insensitive attribute.

Acknowledgments This research work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement Nos. 722424 (SECRET) and 722429 (5GSTEPFWD).

References

1. Bahl, I., & Blass, B. (2003). *Microwave solid state circuit design*. John Wiley & Sons.
2. Sechi, F., & Bujatti, M. (2009). *Solid-state microwave high-power amplifiers*. Artech House Inc.
3. Robertson, I., Somjit, N., & Chong, M. (2016). *Microwave and millimeter-wave design for wireless communication*. Wiley.
4. Haigh, D., Soin, G., & Wood, R. S. (2001). *RF IC and MMIC design and technology, IET circuits, devices and system*. Institution of Electrical Engineers.
5. Marsh, S. (2006). *Practical MMIC design*. Artech House Inc.
6. Walker, J. (2012). *Handbook of RF and microwave power amplifiers*. Cambridge University Press.
7. Sajedin, M. et al. (2020). A Doherty power amplifier based on the harmonic generating mechanism. In *14th European conference on antennas and propagation (EuCAP)*, Copenhagen, Denmark, 1–5, https://doi.org/10.23919/EuCAP48036.2020.9135416.
8. Tsai, J., & Huang, T. (May 2007). A 38–46 GHz MMIC doherty power amplifier using post-distortion linearization. *IEEE Microwave and Wireless Components Letters, 17*(5), 388–390. https://doi.org/10.1109/LMWC.2007.895726
9. Das, N., & Bertoni, H. (1999). *Directions for the next generation of MMIC devices and systems*. Plenum Press.
10. Sajedin, M. et al. (2020). *Design of a broadband frequency response class-J power amplifier*. International Multi-Disciplinary Conference Theme, Sustainable Development and Smart Planning.
11. Grebennikov, A., Kumar, N., Binboga, S., & Yarman, S. (2016). *Broadband RF and microwave amplifiers*. Taylor & Francis Group, LLC.
12. Carey, E., & Lidholm, S. (2005). *Millimeter-wave integrated circuits*. Springer.
13. Giannini, F., & Leuzzi, G. (2004). *Nonlinear microwave circuit design*. Wiley.
14. Sajedin, M., Elfergani, I., Rodriguez, J., Abd-Alhameed, R., & Barciela, M. (2019). A survey on RF and microwave Doherty power amplifier for mobile handset applications. *Electronics, 8*(717), 1–15.
15. Kang, D., Kim, D., Moon, J., & Kim, B. (December 2010). Broadband HBT Doherty power amplifiers for handset applications. *IEEE Transactions on Microwave Theory and Techniques, 58*(12), 4031–4039. https://doi.org/10.1109/TMTT.2010.2086070
16. Refai, W. Y., & Davis, W. A. (2015). A linear, highly-efficient, class-J handset power amplifier utilizing GaAs HBT technology. In *2015 IEEE 16th annual wireless and microwave technology conference (WAMICON), Cocoa Beach, FL* (pp. 1–4). https://doi.org/10.1109/WAMICON.2015.7120353
17. Cripps, S. (2006). *RF power amplifiers for wireless communications*. Artech House.
18. Sajedin, M., et al. (2020). A Doherty power amplifier based on the harmonic generating mechanism. In *2020 14th European conference on antennas and propagation (EuCAP), Copenhagen, Denmark* (pp. 1–5). https://doi.org/10.23919/EuCAP48036.2020.9135416
19. Kim, J., et al. (February 2008). Analysis of a fully matched saturated Doherty amplifier with excellent efficiency. *IEEE Transactions on Microwave Theory and Techniques, 56*(2), 328–338. https://doi.org/10.1109/TMTT.2007.914361
20. Cho, Y., Kang, D., Moon, K., Jeong, D., & Kim, B. (September-October 2017). A handy dandy Doherty PA: A linear Doherty power amplifier for mobile handset application. *IEEE Microwave Magazine, 18*(6), 110–124. https://doi.org/10.1109/MMM.2017.2712040
21. Cho, Y., Moon, K., Park, B., Kim, J., Jin, H., & Kim, B. (2015). Compact design of linear Doherty power amplifier with harmonic control for handset applications. In *2015 10th European microwave integrated circuits conference (EuMIC), Paris* (pp. 37–40). https://doi.org/10.1109/EuMIC.2015.7345062
22. Nguyen, D. P., Pham, B. L., & Pham, A. (2017). A compact 29% PAE at 6 dB power back-off E-mode GaAs pHEMT MMIC Doherty power amplifier at Ka-band. In *2017 IEEE MTT-*

S international microwave symposium (IMS), Honolulu, HI (pp. 1683–1686). https://doi.org/10.1109/MWSYM.2017.8058964

23. Cripps, S. C., Tasker, P. J., Clarke, A. L., Lees, J., & Benedikt, J. (October 2009). On the continuity of high efficiency modes in linear RF power amplifiers. *IEEE Microwave and Wireless Components Letters, 19*(10), 665–667. https://doi.org/10.1109/LMWC.2009.2029754
24. Chen, W., Lv, G., Liu, X., Wang, D., & Ghannouchi, F. M. (May 2020). Doherty PAs for 5G massive MIMO: Energy-efficient integrated DPA MMICs for sub-6-GHz and mm-wave 5G massive MIMO systems. *IEEE Microwave Magazine, 21*(5), 78–93. https://doi.org/10.1109/MMM.2020.2971183
25. Pedro, J. C., Carvalho, N., Fager, C., & Garcia, J. (2004). Linearity versus efficiency in mobile handset power amplifiers: A battle without a loser. In *Microwave engineering Europe*, EENEWS EUROPE (pp. 19–26).
26. Alizadeh, A., & Medi, A. (August 2017). Investigation of a class-J mode power amplifier in presence of a second-harmonic voltage at the gate node of the transistor. *IEEE Transactions on Microwave Theory and Techniques, 65*(8), 3024–3033. https://doi.org/10.1109/TMTT.2017.2666145
27. Alizadeh, A., Hassanzadehyamchi, S., Medi, A., & Kiaei, S. (October 2020). An X-band class-J power amplifier with active load modulation to boost drain efficiency. *IEEE Transactions on Circuits and Systems I: Regular Papers, 67*(10), 3364–3377. https://doi.org/10.1109/TCSI.2020.2991184
28. Pereira, A., Parker, A., Heimlich, M., Weste, N., Quay, R., & Carrubba, V. (2014). X-band high-efficiency GaAs MMIC PA. In *Proceedings of WAMICON* (pp. 1–4).
29. Lv, G., Chen, W., & Feng, Z. (2018). A compact and broadband Ka-band asymmetrical GaAs Doherty power amplifier MMIC for 5G communications. In *2018 IEEE/MTT-S international microwave symposium – IMS, Philadelphia, PA* (pp. 808–811). https://doi.org/10.1109/MWSYM.2018.8439219
30. Nguyen, D. P., Pham, B. L., & Pham, A. (January 2019). A compact Ka-band integrated Doherty amplifier with reconfigurable input network. *IEEE Transactions on Microwave Theory and Techniques, 67*(1), 205–215. https://doi.org/10.1109/TMTT.2018.2874249
31. Nguyen, D. P., Pham, T., & Pham, A. (2017). A Ka-band asymmetrical stacked-FET MMIC Doherty power amplifier. In *2017 IEEE radio frequency integrated circuits symposium (RFIC), Honolulu, HI* (pp. 398–401). https://doi.org/10.1109/RFIC.2017.7969102
32. Quaglia, R., Camarchia, V., Jiang, T., Pirola, M., Donati Guerrieri, S., & Loran, B. (November 2014). K-band GaAs MMIC Doherty power amplifier for microwave radio with optimized driver. *IEEE Transactions on Microwave Theory and Techniques, 62*(11), 2518–2525. https://doi.org/10.1109/TMTT.2014.2360395
33. Hu, S., Wang, F., & Wang, H. (2017). 2.1 A 28GHz/37GHz/39GHz multiband linear Doherty power amplifier for 5G massive MIMO applications. In *2017 IEEE international solid-state circuits conference (ISSCC), San Francisco, CA* (pp. 32–33). https://doi.org/10.1109/ISSCC.2017.7870246
34. Chen, Y., Lin, Y., Lin, J., & Wang, H. (December 2018). A Ka-band transformer-based Doherty power amplifier for multi-Gb/s application in 90-nm CMOS. *IEEE Microwave and Wireless Components Letters, 28*(12), 1134–1136. https://doi.org/10.1109/LMWC.2018.2878133
35. Wang, F., & Wang, H. (2020). 24.1 A 24-to-30GHz watt-level broadband linear Doherty power amplifier with multi-primary distributed-active-transformer power-combining supporting 5G NR FR2 64-QAM with >19dBm average pout and >19% average PAE. In *2020 IEEE international solid- state circuits conference – (ISSCC), San Francisco, CA, USA* (pp. 362–364). https://doi.org/10.1109/ISSCC19947.2020.9063146
36. Hu, S., Wang, F., & Wang, H. (June 2019). A 28-/37-/39-GHz linear Doherty power amplifier in silicon for 5G applications. *IEEE Journal of Solid-State Circuits, 54*(6), 1586–1599. https://doi.org/10.1109/JSSC.2019.2902307
37. Abdulkhaleq, A. M., et al. (2020). Load-modulation technique without using quarter-wavelength transmission line. *IET Microwaves, Antennas and Propagation, 14*, 1209. https://doi.org/10.1049/iet-map.2019.0957

38. Abdulkhaleq, A. M., et al. (2019). Recent developments of dual-band Doherty power amplifiers for upcoming mobile communications systems. *Electronics, 8*(6), 638. https://doi.org/10.3390/electronics8060638
39. Abdulkhaleq, A. M., et al. (2019). A 70-W asymmetrical Doherty power amplifier for 5G base stations. In V. Sucasas, G. Mantas, & S. Althunibat (Eds.), *Broadband communications, networks, and systems* (pp. 446–454). Springer International Publishing.
40. Abdulkhaleq, A. M. et al. (2020). *A compact load-modulation amplifier for improved efficiency next generation mobile*. Presented at the 50th The European Microwave Conference (EuMC), The Jaarbeurs, The Netherlands.
41. Abdulkhaleq, A. M., et al. (2020). *Mutual coupling effect on three-way Doherty amplifier for green compact mobile communications*. Presented at the EuCAP 2020, 15–20-March-2020.
42. Al-Yasir, Y. I. A., et al. (2020). A differential-fed dual-polarized high-gain filtering antenna based on SIW technology for 5G applications. In *2020 14th European Conference on Antennas and Propagation (EuCAP), Copenhagen, Denmark*, IEEE (pp. 1–5).
43. Al-Yasir, Y. I. A., Ojaroudi Parchin, N., Abdulkhaleq, A., Hameed, K., Al-Sadoon, M., & Abd-Alhameed, R. (2019). Design, simulation and implementation of very compact dual-band microstrip bandpass filter for 4G and 5G applications. In *2019 16th international conference on synthesis, modeling, analysis and simulation methods and applications to circuit design (SMACD), Lausanne, Switzerland*. IEEE.
44. Feng, W., Che, W., & Xue, Q. (June 2015). The proper balance: Overview of microstrip wideband balance circuits with wideband common mode suppression. *IEEE Microwave Magazine, 16*(5), 55–68.
45. Al-Yasir, Y. I. A., Ojaroudi Parchin, N., Abdulkhaleq, A. M., Bakr, M. S., & Abd-Alhameed, R. A. (2020). A survey of differential-fed microstrip bandpass filters: Recent techniques and challenges. *Sensors, 20*(8), 2356.
46. Jin, H., Chin, K., Che, W., Chang, C., Li, H., & Xue, Q. (2014). Differential-fed patch antenna arrays with low cross polarization and wide bandwidths. *IEEE Antennas and Wireless Propagation Letters, 13*, 1069–1072.
47. Chin, C. K., Xue, Q., & Wong, H. (September 2007). Broadband patch antenna with a folded plate pair as a differential feeding scheme. *IEEE Transactions on Antennas and Propagation, 55*(9), 2461–2467.
48. Chin, C. h. k., Xue, Q., Wong, H., & Zhang, X. y. (February 2007). Broadband patch antenna with low cross-polarisation. *Electronics Letters, 43*(3), 137–138.
49. Luo, Y., & Chu, Q. (November 2015). Oriental crown-shaped differentially fed dual-polarized multidipole antenna. *IEEE Transactions on Antennas and Propagation, 63*(11), 4678–4685.
50. White, C. R., & Rebeiz, G. M. (November 2010). A differential dual-polarized cavity-backed microstrip patch antenna with independent frequency tuning. *IEEE Transactions on Antennas and Propagation, 58*(11), 3490–3498.
51. Cui, J., Zhang, A., & Yan, S. (2020, February). Co-design of a filtering antenna based on multilayer structure. *International Journal of RF and Microwave Computer-Aided Engineering, 30*(2), 1–6.
52. Hua, C., Liu, M., & Lu, Y. (February 2019). Planar integrated substrate integrated waveguide circularly polarized filtering antenna. *International Journal of RF and Microwave Computer-Aided Engineering, 29*(2), e21517.
53. Al-Yasir, Y. I. A., et al. (2020). A new and compact wide-band microstrip filter-antenna design for 2.4 GHz ISM band and 4G applications. *Electronics, 9*(7), 1084.
54. Majid, H. A., Rahim, M. K. A., Hamid, M. R., & Ismail, M. F. (2012). A compact frequency-reconfigurable narrowband microstrip slot antenna. *IEEE Antennas and Wireless Propagation Letters, 11*, 616–619.
55. Yassin, M. E., Mohamed, H. A., Abdallah, E. A. F., & El-Hennawy, H. S. (2019). Circularly polarized wideband-to-narrowband switchable antenna. *IEEE Access, 7*, 36010–36018.
56. Tu, Y., Al-Yasir, Y., Ojaroudi Parchin, N., Abdulkhaleq, A., & Abd-Alhameed, R. (2020, June). A survey on reconfigurable microstrip filter–antenna integration: Recent developments and challenges. *Electronics, 9*(8), 1–21.

57. Caicedo, S., Oldoni, M., & Moscato, S. (2021). Challenges of using phased array antennas in a commercial backhaul equipment at 26 GHz. In *Internet of things, infrastructures and Mobile applications. IMCL 2019* (Advances in intelligent systems and computing, vol 1192). Springer.
58. Mejillones, S. C., Oldoni, M., Moscato, S., Fonte, A., & D'Amico, M. (2020). Power consumption and radiation trade-offs in phased arrays for 5G wireless transport. In *2020 43rd international conference on telecommunications and signal processing (TSP), Milan, Italy* (pp. 112–116). https://doi.org/10.1109/TSP49548.2020.9163445
59. SIAE Microelettronica. *ALFOplus2: Wireless backhaul/fronthaul equipment*. https://www.siaemic.com/index.php/products-services/telecommunication-systems/microwave-product-portfolio/alfo-plus2. Accessed 12 Dic 2020.
60. Mailloux, R. J. (2017). *Phased array antenna handbook* (3rd ed.). Artech House, Inc.
61. Shome, P. P., Khan, T., Koul, S., & Antar, Y. (2020). Filtenna designs for radio-frequency front-end systems: A structural-oriented review. *IEEE Antennas and Propagation Magazine*. https://doi.org/10.1109/MAP.2020.2988518
62. Lee, J., Kidera, N., Pinel, S., Laskar, J., & Tentzeris, M. M. (2007). Fully integrated passive front-end solutions for a V-band LTCC wireless system. *IEEE Antennas and Wireless Propagation Letters, 6*, 285–288. https://doi.org/10.1109/LAWP.2007.891964
63. Li, R., & Gao, P. (January 2016). Design of a UWB filtering antenna with defected ground structure. *Progress in Electromagnetics Research Letters, 63*, 65–70.
64. Mishra, S., Sheeja, K., & Pathak, N. (December 2017). Split ring resonator inspired microstrip Filtenna for KU-band application. *Journal Europeen des Systemes Automatises, 50*, 391–403.
65. Hu, K., Tang, M., Li, M., & Ziolkowski, R. W. (August 2018). Compact, low-profile, bandwidth-enhanced substrate integrated waveguide filtenna. *IEEE Antennas and Wireless Propagation Letters, 17*(8), 1552–1556. https://doi.org/10.1109/LAWP.2018.2854898
66. Escobar, A. H., Tirado, J. A. V., Gomez, J. C. C., Mateu, J., Cantenys, E. R., & Gonzalez, J. L. (March 2014). Filtenna integration achieving ideal Chebyshev return losses. *Radioengineering, 23*, 362–368.
67. Cameron, R. J. (January 2003). Advanced coupling matrix synthesis techniques for microwave filters. *IEEE Transactions on Microwave Theory and Techniques, 51*(1), 1–10. https://doi.org/10.1109/TMTT.2002.806937
68. Li, T., & Gong, X. (June 2018). Vertical integration of high-Q filter with circularly polarized patch antenna with enhanced impedance-axial ratio bandwidth. *IEEE Transactions on Microwave Theory and Techniques, 66*(6), 3119–3128. https://doi.org/10.1109/TMTT.2018.2832073
69. Yusuf, Y., Cheng, H., & Gong, X. (November 2011). A seamless integration of 3-D vertical filters with highly efficient slot antennas. *IEEE Transactions on Antennas and Propagation, 59*(11), 4016–4022. https://doi.org/10.1109/TAP.2011.2164186
70. Cassivi, Y., Perregrini, L., Arcioni, P., Bressan, M., Wu, K., & Conciauro, G. (September 2002). Dispersion characteristics of substrate integrated rectangular waveguide. *IEEE Microwave and Wireless Components Letters, 12*(9), 333–335. https://doi.org/10.1109/LMWC.2002.803188
71. Jia-Sheng, & Lancaster, M. J. (2011). *Microstrip filters for RF/microwave applications*. Wiley.
72. Selvaraju, R., Jamaluddin, M. h., Kamarudin, M., Nasir, J., & Dahri, M. (January 2018). Complementary split ring resonator for isolation enhancement in 5g communication antenna array. *Progress in Electromagnetics Research C, 83*, 217.
73. Oldoni, M., Macchiarella, G., Gentili, G. G., & Ernst, C. (May 2010). A new approach to the synthesis of microwave lossy filters. *IEEE Transactions on Microwave Theory and Techniques, 58*(5), 1222–1229. https://doi.org/10.1109/TMTT.2010.2045534
74. Cameron, R. J. (April 1999). General coupling matrix synthesis methods for Chebyshev filtering functions. *IEEE Transactions on Microwave Theory and Techniques, 47*(4), 433–442. https://doi.org/10.1109/22.754877
75. Bigelli, F., et al. (February 2016). Design and fabrication of a dielectricless substrate-integrated waveguide. *IEEE Transactions on Components, Packaging and Manufacturing Technology, 6*(2), 256–261. https://doi.org/10.1109/TCPMT.2015.2513077

76. Osseiran, A., et al. (May 2014). Scenarios for 5G mobile and wireless communications: The vision of the METIS project. *IEEE Communications Magazine, 52*(5), 26–35.
77. Rappaport, T. S., et al. (2013). Millimeter wave mobile communications for 5G cellular: It will work! *IEEE Access, 1*, 335–349.
78. Rodriguez, J., et al. (2017). SECRET — Secure network coding for reduced energy next generation mobile small cells: A European training network in wireless communications and networking for 5G. In *2017 internet technologies and applications (ITA), Wrexham*, IEEE (pp. 329–333).
79. Parchin, N. O., Shen, M., & Pedersen, G. F. (2016). UWB MM-wave antenna array with quasi omnidirectional beams for 5G handheld devices. In *2016 IEEE international conference on ubiquitous wireless broadband (ICUWB), Nanjing*, IEEE (pp. 1–4).
80. Ojaroudiparchin, N., Shen, M., & Pedersen, G. F. (2016). 8×8 planar phased array antenna with high efficiency and insensitivity properties for 5G mobile base stations. In *2016 10th European conference on antennas and propagation (EuCAP), Davos*, IEEE (pp. 1–5).
81. HMC933LP4E. *Analog phase shifter*. Hittite Microwave Company. http://www.hittite.com
82. Hong, W., Baek, K., Lee, Y., & Kim, Y. G. (2014). Design and analysis of a low-profile 28 GHz beam steering antenna solution for future 5G cellular applications. In *2014 IEEE MTT-S international microwave symposium (IMS2014), Tampa, FL*, IEEE (pp. 1–4).
83. Parchin, N. O., et al. (2019). MM-wave phased array quasi-yagi antenna for the upcoming 5G cellular communications. *Applied Sciences, 9*, 1–14.
84. Parchin, N. O., et al. (2019). Frequency reconfigurable antenna array for mm-wave 5G mobile handsets. In *Broadband communications, networks, and systems, Faro, Portugal, 19–20 September 2018*. Springer.
85. Tang, M., Ziolkowski, R. W., & Xiao, S. (June 2014). Compact hyper-band printed slot antenna with stable radiation properties. *IEEE Transactions on Antennas and Propagation, 62*(6), 2962–2969.
86. Ojaroudi, N., & Ghadimi, N. (2014). Dual-band CPW-fed slot antenna for LTE and WiBro applications. *Microwave and Optical Technology Letters, 56*, 1013–1015.
87. Parchin, N. O., et al. (2019). Eight-element dual-polarized MIMO slot antenna system for 5G smartphone applications. *IEEE Access, 7*, 15612–15622.
88. Salman, J., et al. (2006). Effects of the loss tangent, dielectric substrate permittivity and thickness on the performance of circular microstrip antennas. *Journal of Engineering and Development, 10*, 1–13.
89. Rajagopal, S., Abu-Surra, S., Pi, Z., & Khan, F. (2011). Antenna array design for multi-Gbps mmWave mobile broadband communication. In *2011 IEEE global telecommunications conference – GLOBECOM 2011, Houston, TX, USA*, IEEE (pp. 1–6).
90. Ilvonen, J., Kivekas, O., Holopainen, J., Valkonen, R., Rasilainen, K., & Vainikainen, P. (2011). Mobile terminal antenna performance with the user's hand: Effect of antenna dimensioning and location. *IEEE Antennas and Wireless Propagation Letters, 10*, 772–775.

Chapter 5
Security for UDNs: A Step Toward 6G

Marcus de Ree, Reza Parsamehr, Vipindev Adat, Georgios Mantas, Ilias Politis, Jonathan Rodriguez, Stavros Kotsopoulos, Ifiok E. Otung, José-Fernán Martínez-Ortega, and Felipe Gil-Castiñeira

Abstract The next-generation mobile networks are taking advantage of small cell technology toward building the notion of ultra-dense networks (UDNs). The considered UDN within this chapter consists of virtual network coding (NC)-enabled mobile small cells (MSCs), a novel networking scenario that consists entirely out of heterogeneous mobile devices. In this networking scenario, the mobile devices benefit from high transmission speeds, low latency, and increased energy efficiency, whereas the mobile network infrastructure benefits from a reduction in traffic due to enabling traffic offloading. However, MSCs can potentially face a variety of security and privacy challenges. This chapter covers three important security infrastructures, (i) decentralized key management schemes, (ii) intrusion detection and prevention schemes, and (iii) blockchain-based integrity schemes. Decen-

M. de Ree (✉)
Instituto de Telecommunicações, Aveiro, Portugal

University of South Wales, Treforest, UK
e-mail: mderee@av.it.pt

J. Rodriguez
Instituto de Telecommunicações, Campus Universitário Santiago, Aveiro, Portugal

Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, UK
e-mail: jonathan@av.it.pt

R. Parsamehr
Instituto de Telecomunicações, Aveiro, Portugal

Universidad Politécnica de Madrid, Madrid, Spain
e-mail: parsamehr.r@av.it.pt

V. Adat
University of Patras, Patras, Greece

atlanTTic Research Center for Telecommunication Technologies, Universidade de Vigo, Information Technologies Group, Vigo, Spain
e-mail: adat@upatras.gr; xil@gti.uvigo.es

G. Mantas
Instituto de Telecomunicações, Aveiro, Portugal

University of Greenwich, London, UK

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_5

tralized key management enables the heterogeneous mobile devices to securely exchange cryptographic keys. These cryptographic keys can then be utilized by blockchain-based integrity schemes to establish secure and reliable communication channels between mobile devices, even in the presence of malicious adversaries. The intrusion detection and prevention schemes attempt to identify and remove these malicious adversaries from the network. These security infrastructures can potentially be used as stepping stones toward a security architecture for general MSC architectures and 6G communications.

5.1 Introduction

Cybercrime is a recurring theme in today's networks with recent events, such as the ransomware strain called WannaCry (May 2017) spread worldwide, attacking a multitude of targets, including public utilities and large corporations. On March 7, 2017, WikiLeaks published a data treasure trove of 8761 documents allegedly stolen from the CIA that contained information of purported spying operations and hacking tools. WikiLeaks called the dump – followed by smaller disclosures – "Vault 7" [1]. The past has taught us that networks will always be prone to new security attacks and the best we can do is to detect security threats and adopt a proactive stance. So how do we deal with security in UDNs, where clients are likely to upload/download confidential information over the small cell wireless network?

In this chapter, the considered UDN consists of virtual MSCs. The virtual MSC networking architecture was first proposed by the EU-funded H2020-MSCA project "SECRET" [2] and is illustrated in Fig. 5.1.

These virtual MSCs combine the small cell, the network function virtualization (NFV), and the software-defined network (SDN) networking technologies to realize a virtual network of heterogeneous mobile devices. The software-defined controller controls and maintains each virtual MSC through interaction with a so-called cluster-head, or hotspot. This hotspot is a heterogeneous mobile device that is selected as the local radio manager, controlling and maintaining the cluster. Through cooperation, the selected hotspots form a wireless network that has several gateways to the mobile network using intelligent high-speed connections. These virtual

I. Politis
University of Patras, Patras, Greece

University of Piraeus, Piraeus, Greece

S. Kotsopoulos
University of Patras, Patras, Greece

I. E. Otung
University of South Wales, Treforest, UK

J. -F. Martínez-Ortega
Universidad Politécnica de Madrid, Madrid, Spain

F. Gil-Castiñeira
atlanTTic Research Center for Telecommunication Technologies, Universidade de Vigo,
Information Technologies Group, Vigo, Spain

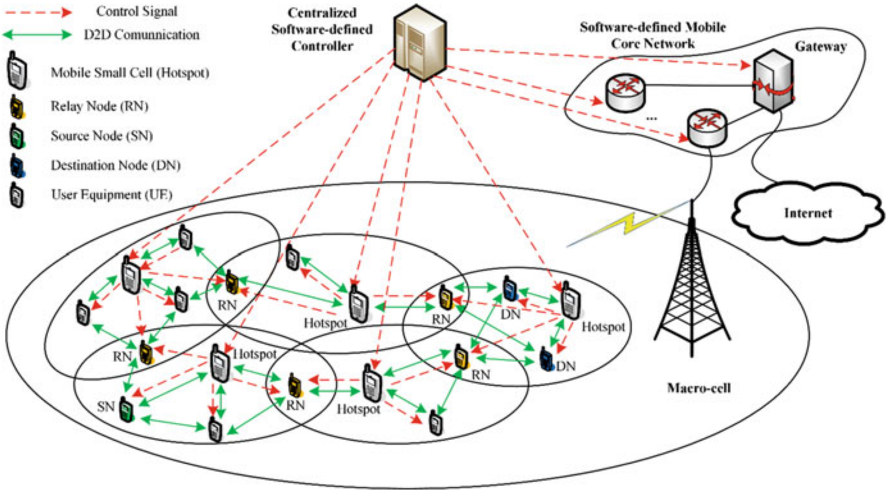


Fig. 5.1 The network architecture with NC-MSCs [3]

MSCs can be set up on the fly, based on demand, at any place and at any time. Mobile data traffic between the mobile devices is enabled through device-to-device (D2D) communications. This allows mobile devices within relative close proximity to communicate (over multiple hops if necessary) without having to rely on the network infrastructure. Moreover, data traffic flow between mobile devices is optimized with the incorporation of the NC networking paradigm. These virtual NC-enabled MSCs (from here on abbreviated to NC-MSCs) can function alongside the next-generation mobile network, effectively offloading the mobile network infrastructure while providing a high level of quality of service (QoS) to network subscribers.

This chapter tackles the security aspects of UDNs based on developing a secure architecture for NC-MSCs, capable of preventing and defending against common attacks including pollution and denial-of-service (DoS)-type attacks. It is worthy to note that the considered security approaches can be applied to general UDN networks, although we consider NC-MSCs here as a case study. As such, the chapter aims to identify the key technologies and proposes preliminary reference architectures that potentially could be used as stepping stones toward a secure architecture. Section 5.2 covers the potential of decentralized key management schemes through a fully distributed trusted third party (TTP) to provide security for highly dynamic UDNs, also compatible with NC-enabled networks such as NC-MSCs. In Sect. 5.3, a recently proposed intrusion detection and prevention scheme (IDPS) is examined. The examined IDPS had specifically been designed to support and secure NC-MSCs. A recently proposed blockchain-based integrity scheme for UDNs is examined in Sect. 5.4. This blockchain-based integrity scheme enables the possibility of verifying the authenticity of transmitted data at every link in a multi-hop network. This potential can counter data pollution and tag pollution attacks that are both considered severe network performance-damaging attacks in NC-enabled networks.

5.2 Decentralized Key Management for NC-MSCs

The virtual NC-MSCs are able to provide major benefits to the mobile networking environment. As was mentioned in the introduction, one benefit of NC-MSCs is its ability to offload the network infrastructure by allowing network users to communicate through multi-hop D2D communications. This will become increasingly more important in dense urban areas due to the constant increase in mobile device connections and mobile data demands. Moreover, the incorporation of NC-MSCs provides additional robustness of the mobile network in case of failure or during a power outage. However, the intermediate mobile devices forward private data between communicating parties and thus must be secured.

To secure multi-hop D2D communications, cryptographic security solutions (e.g., encryption schemes, signature schemes) are available. These solutions, however, assume that any pair of network users have a means to securely exchange cryptographic keys. In other words, they rely on a key management scheme that manages the establishment, exchange, verification, update, and revocation of cryptographic keys. In this section, we investigate how secure multi-hop D2D communications can be established in a self-organizing and decentralized manner.

5.2.1 State-of-the-Art Decentralized Key Management

We consider the network of NC-MSCs as a wireless mobile ad hoc network (MANET), designed to be deployed alongside future mobile networks. Due to the resemblance between NC-MSCs and MANETs, a recent survey [4] evaluated whether any of the six decentralized key management solutions proposed for MANETs can be adopted to secure NC-MSCs, and its results are shown in Table 5.1. The evaluation was based on various requirements, including:

- Security: Its ability to provide a high level of security for all network users
- Connectivity (Connect.): Its ability to establish a secure channel between any arbitrary pair of network users such that security solutions against NC-related attacks (e.g., pollution attacks) are supported

Table 5.1 Evaluation and comparison of explored decentralized key management solutions [4]

Key management solution	Security	Connect.	Scal.	Sustain.	Fairness
Certificate chaining-based [5]	✘	✘	✓	✓	✓
Mobility-based [6]	✓	✘	✓	✓	✓
Self-certification-based [7]	✘	✓	✘	✓	✓
Combinatorics-based [8]	✘	✓	✓	✘	✓
Partially distributed TTP-based [9]	✓	✓	✘	✘	✘
Fully distributed TTP-based [10]	✓	✓	✓	✓	✓

- Scalability (Scal.): Its ability to support a large number of network users
- Sustainability (Sustain.): Its ability to sustain its level of security, connectivity, and overhead during the long mobile network lifetime
- Fairness: Its ability to fairly distribute the workload of key management services

The survey found that five of the six key management solutions suffered from severe drawbacks and were found unsuitable for adoption to NC-MSCs. The certificate chaining-based [5] and mobility-based key management solutions [6] suffer from similar drawbacks. Both solutions are unable to guarantee connectivity between any arbitrary pair of network users, and the bootstrapping of trust requires real-life interactions. The self-certification-based solution [7] suffers from tremendous amounts of communication overhead in dense and highly dynamic networks. The combinatorics-based solution [8] is outright insecure against man-in-the-middle (MITM) attacks during the exchange of public keys. The partially distributed TTP (PD-TTP)-based solution [9] relies on a proper subset of network users to provide key management services and therefore does not satisfy the fairness requirement, especially in a large, dense, and highly dynamic network. The fully distributed TTP (FD-TTP)-based solution [10] does not have any of the previously mentioned or any other severe inherent drawbacks. Therefore, the FD-TTP-based key management solution was proposed as the most suitable candidate to secure multi-hop D2D communications in NC-MSCs.

5.2.2 Distributed Trusted Third Party-Based Key Management

Any public key cryptographic infrastructure relies on a TTP as a trust anchor, in possession of a master private key, to perform some kind of key management service. This key management service can be the creation and distribution of a signed certificate or (partial) private key. Thus, the master private key can be used for different purposes, but in all cases, they enable network users to establish secure communication channels with other network users. Through careful analysis, we argued that the network architecture of NC-MSCs is unable to support a single, secure, and trustworthy entity that is capable of acting as the TTP. Both the network infrastructure and individual mobile devices are susceptible to DoS-type attacks and physical compromise. Therefore, we are required to decentralize trust.

5.2.2.1 Fully Distributed Trusted Third Party-Based Key Management

There are two major forms of the distributed TTP. In the PD-TTP, the master private key is divided into shares, and these shares are then distributed to a proper subset of all the network users, also called servers. However, partial distribution of trust leads to an asymmetric workload distribution. The assigned servers that provide key management services suffer from an increase in energy consumption, effectively reducing their battery lifetime. This may lead to servers acting selfishly, denying to provide key management services to preserve their battery, even when the server roles are rotated periodically. The drawback of the asymmetric workload is significantly worsened when a limited number of servers are responsible for key management services of a large and dense network. Finally, another significant issue is related to the sustainability of the key management service. The dynamic network topology may cause certain network areas to be server-less, leading to a temporary unavailable key management service. Furthermore, servers may leave the network entirely over time, potentially leading to a permanently unavailable key management service.

The FD-TTP is similar to the PD-TTP, except that the shares are distributed to all the network users. This also means that when a new user joins the network, it is also provided with a share of the master private key. This distribution solves all the issues present in the PD-TTP. The distribution of shares in the FD-TTP yields an evenly distributed workload and maximizes the key management service availability at any place and at any time. Therefore, a key management scheme utilizing the FD-TTP is preferred to secure NC-MSCs.

5.2.2.2 Secret Sharing Techniques

The distribution of trust is equivalent to the distribution of shares of the master private key. The shares of the master private key are generated through a technique called secret sharing. In ordinary secret sharing, a piece of secret data (e.g., the master private key) is divided into a multitude of shares and distributed among a group of users. The secret data can be reconstructed by combining a certain number of shares. For the creation of a distributed TTP, we utilize threshold secret sharing. This means that any threshold number of shares are capable of reconstructing the secret data.

The threshold secret sharing technique proposed by Shamir [11] is generally used to create the distributed TTP. The master private key (represented as *MSK* in Eq. (5.1)), usually in possession of a secure, trustworthy, and centralized TTP, is encoded into a polynomial of degree $t - 1$, as shown below:

$$s_{ID} = f(ID) = MSK + a_1 ID + a_2 ID^2 + \cdots + a_{t-1} ID^{t-1} \in \mathbb{Z}_q \tag{5.1}$$

The value t represents the threshold, the number of shares necessary to recover the value of the master private key. For a network user with identity *ID*, its share s_{ID} is computed through polynomial evaluation. This polynomial allows us to distribute shares to a group of network users, becoming servers. These servers form the newly created distributed TTP. A threshold amount of these servers can collaboratively provide key management services since they collectively have enough pieces to reconstruct the master private key. Furthermore, they can provide such a service without having to disclose their individual share, thus preserving the secrecy of the master private key.

To ensure that the provided key management service is trustworthy, we can incorporate a verifiable secret sharing extension [12, 13]. Verifiable secret sharing (VSS) enables users, receiving the key management service, to verify that the servers used their share and not any other arbitrary value. If a server still provides a false key management service, the server can be accused of being malicious and be removed as a member of the distributed TTP.

5.2.2.3 Security Considerations

In a decentralized key management scheme that relies on the distribution of trust through secret sharing techniques, the most important aspect of security is the continued secrecy of the master private key. We discuss, underneath, two types of attacks that attempt to reconstruct the master private key and an important and inherent characteristic of cryptographic infrastructures and their impact on providing security for a decentralized key management scheme.

Mobile Adversary Attack
In the mobile adversary attack [14], a malicious user dynamically moves through the network and compromises user's mobile devices, one at a time, with the goal to extract and collect a threshold number of shares of the master private key. If the mobile adversary is successful in collecting a threshold number of shares, it is capable of reconstructing the master private key.

Proactive secret sharing (PSS) [15] [16], another extension of secret sharing, is generally used to prevent a successful mobile adversary attack. With the incorporation of PSS, servers periodically update their share of the master private key. Shares that are collected by a mobile adversary in between different updating phases are incompatible in the reconstruction of the master private key. Therefore, PSS limits the amount of time a mobile adversary has to launch a successful attack.

Sybil Attack
In the Sybil attack [17], a malicious user obtains a multitude of (fake) identities (e.g., mobile devices) and wishes to join the network with each one of them. If each identity is given a share of the master private key upon joining the network, a Sybil attacker with a threshold number of identities can collect enough shares to reconstruct the master private key.

Unfortunately, the Sybil attack is generally dismissed in adversarial models of FD-TTP-based key management solutions. This is mainly because different use-case scenarios require different solutions. Use-case scenarios such as rescue operations in remote areas or international military operations can utilize an offline authority (base or headquarters) where users are authenticated prior to joining the network [10]. However, with NC-MSCs, we wish to accept new network users to the network without any physical interaction. If authentication through physical interaction would be required, then mobile network users will continue to communicate through the network infrastructure and refrain from using NC-MSCs. A recent work [18] proposed an FD-TTP-based key management scheme that is suitable for NC-MSCs and would be resilient against a Sybil attack based on a cloaking technique. Unfortunately, their proposal merely formulated an outline of the scheme and therefore still lacks a proof of concept. Otherwise, network operators may have to play a role in the prevention of successful Sybil attacks since network operators have identifying information of their network subscribers. These network operators, for example, could decide to send (or not to send) some kind of access token to mobile devices for network joining purposes.

Trust Level of the Distributed Trusted Third Party

Girault [19] found that public key cryptographic infrastructures have a variety of trust levels. After careful analysis, he defined these levels as follows:

1. At level 1: the TTP knows (or can easily compute) a user' private key and, therefore, launches identity impersonation attacks without being detected.
2. At level 2: the TTP does not know (and cannot easily compute) a user's private key but is still able to launch identity impersonation attacks without being detected.
3. At level 3: the TTP does not know (and cannot easily compute) a user's private key nor is it able to launch identity impersonation attacks without being detected.

The trust level of the distributed TTP essentially defines the capabilities of a malicious user after a successful mobile adversary or Sybil attack. At the third trust level, network users are capable of detecting malicious behavior and thus detect whether the entire network is compromised. The detection of network compromise provides network operators with the ability to reboot and re-initialize the network, potentially with enhanced security parameters such as an increased threshold value or a reduced time interval in between share updating phases. This additional layer of security reduces the payoff of malicious users and should discourage them from launching such attacks. This kind of detection is vital for the B5G mobile network due to its extended lifetime.

5.2.2.4 General Key Management Structure

The general key management structure of a FD-TTP-based key management scheme exists of two main network phases, the network initialization phase and the network operational phase. Furthermore, the network operational phase can be divided into two subphases, namely, the operational subphase and the share updating subphase. The general key management structure is summarized in Table 5.2.

In the network initialization phase, either a centralized TTP is present to initialize a set of at least a threshold number of network users, or these network users initialize the network in a distributed manner [20]. A centralized or decentralized network initialization depends on the assumption whether the networking scenario could support a centralized TTP during this phase. This phase consists of three protocols, the master key creation protocol, the secret share establishment protocol (i.e., the distributed TTP establishment protocol), and the "key" establishment protocol.

Table 5.2 The general key management structure of FD-TTP-based key management schemes

Network phase	Network subphase	Associated protocols
Network initialization		Master key creation Secret share establishment Key establishment
Network operation	Operational	Secure channel establishment Key updating Key revocation Distributed secret share establishment Distributed key establishment
	Share updating	Secret share updating

The exact nature of the "key" depends on the used cryptographic infrastructure. For example, in a traditional public key infrastructure (PKI)-based system, the key represents a signed certificate, and in a traditional identity-based public key cryptographic (ID-PKC) system, the key represents a user's private key.

In the network operation phase, a centralized TTP is not online accessible which requires the network to be self-organized by the individual network users. During the operational subphase, network users can establish a secure communications channel, request the distributed TTP to have their "key" updated, accuse and convict malicious users to have their key revoked, and accept new users to the network through the distributed secret share and key establishment protocols. Periodically, the network enters the share updating subphase to execute the network-wide share updating protocol. This protocol updates every user's secret share that reduces the chances of a successful mobile adversary attack. It is important to mention that this protocol generally should not change the master private key and the associated master public key as described in Eq. (5.1).

5.2.3 *Security Analysis of Fully Distributed Trusted Third Party-Based Key Management Schemes per Cryptographic Infrastructure*

In this section, we discuss the security perspectives of previously proposed FD-TTP-based key management solutions per cryptographic infrastructure. Table 5.3 summarizes these findings.

5.2.3.1 Traditional Public Key Infrastructure

Luo et al. [10, 21] proposed a FD-TTP-based key management solution that is based on the traditional PKI cryptographic infrastructure. In this cryptographic infrastructure, every network user generates their own public-private key pair.

Table 5.3 The security evaluations of proposed FD-TTP-based key management schemes

Cryptographic infrastructure	Scheme	FD-TTP trust level	VSS	PSS	Sybil attack
Traditional PKI	Luo et al. [10] [21]	3	✓[a]	✓	✗
Traditional ID-PKC	Deng et al. [22]	1	✗[b]	✗[c]	✗
TT-ID-PKC	da Silva et al. [23]	1	✗	✗[c]	✗
	de Ree et al. (1) [18]	1+[d]	–	–	✗
CL-PKC	de Ree et al. (2) [18]	1+[d]	–	–	✓
	Zhang et al. [24]	2	✗	✗[c]	✗
	Li et al. [25]	2	✓	✓	✗
	Gharib et al. [26]	2	✓	✗[c]	✗
	Lai et al. [27]	3	✗[b]	✗[c]	✗
	de Ree et al. [28]	3	✓	✓	✗

[a] Verifiable secret sharing is incorporated in the scheme; however, it is incapable of verifying partial keying material

[b] The use of verifiable secret sharing is mentioned; however, it is not incorporated in any proposed protocols

[c] This scheme did not incorporate a share updating phase and protocol to protect the security system against mobile adversaries; however, such a phase can be added in a trivial manner

[d] An important difference between the TT-ID-PKC and other cryptographic infrastructures is that the master private key can be updated periodically. Therefore, the security system will only be compromised until the next secret share/key updating protocol is executed, limiting the payoff of malicious entities. Other cryptographic infrastructures do not allow the master private key to be updated, leading to a compromised security system until network reboot

The network user would then request the FD-TTP to certify its public key. Upon receiving a threshold amount of partially signed certificates, the network user can combine these into its complete certificate. This certificate can then be exchanged to other network users, which can verify its authenticity. The authenticated public key can then be used to establish a secure communications channel.

Clearly, the FD-TTP achieves trust level 3 in schemes that are based on traditional PKI. The FD-TTP is unable to compute a user's private key, and the existence of two (or more) different certificates for the same user would prove that the FD-TTP has cheated [19, 29]. Unfortunately, it was demonstrated in [30] that network users are unable to verify whether partial certificates and partial secret shares are correct.

5.2.3.2 Traditional Identity-Based Public Key Cryptography

Deng et al. [22] and da Silva et al. [23] proposed a FD-TTP-based key management solution that is based on traditional ID-PKC. In this cryptographic infrastructure, the user's network identity (e.g., e-mail address, phone number) is used as the user's public key. This public key is considered public knowledge and thus eliminates the need to exchange public keys. However, a user is unable to compute its private key from the public key. A user's private key can be computed using the master private key; thus, any user must request the FD-TTP to collect pieces of its private key.

It is clear that the FD-TTP following the traditional ID-PKC-based cryptographic infrastructure only achieves trust level 1 [19, 29]. Therefore, a compromised FD-TTP gains tremendous power. It has been suggested that schemes based on traditional ID-PKC are more suitable in small and closed networks with limited security requirements due to this drawback [24, 25].

5.2.3.3 Threshold-Tolerant Identity-Based Public Key Cryptography

Recently, de Ree et al. [18] proposed two versions of a FD-TTP-based key management solution that is based on threshold-tolerant ID-PKC (TT-ID-PKC) [31]. This TT-ID-PKC cryptographic infrastructure is essentially a translation of Feldman's VSS scheme [12] where the secret shares are directly used as private keys. Like ID-PKC, the public key of a network user can be computed from publicly available information, and the private key must be obtained through interaction with the FD-TTP. Therefore, the FD-TTP again reaches only a trust level of 1.

However, directly using the secret shares as private keys has significant consequences to the key management structure. As mentioned in [18], (i) the key management design can be significantly simplified since secret share-related and key-related protocols are merged; and (ii) the master private key is no longer necessary to provide a key management service; thus, it does not need to be preserved in the share updating protocol. In that case, a malicious FD-TTP will only be capable of launching malicious attacks prior to the next share updating phase. Unfortunately, the authors of [18] only provided a general outline of their key management solutions and require a proof of concept to prove that such benefits can be achieved.

5.2.3.4 Certificateless Public Key Cryptography

Zhang et al. [24], Li et al. [25], Gharib et al. [26], Lai et al. [27], and de Ree et al. [28] proposed FD-TTP-based key management solutions based on certificateless public key cryptography (CL-PKC) [32]. This cryptographic infrastructure is a hybrid between traditional PKI and ID-PKC. A network user essentially combines the self-generated public-private key pair with an identity-based public-private key pair. The self-generated public key and the user's identity are combined into the user's public key, and the self-generated private key and the identity-based partial private key (obtained from the FD-TTP) are combined into the user's private key.

Al-Riyami [32] mentioned that the TTP could reach either trust level 2 or trust level 3, depending on the key generation technique. We found that the key generation technique used by Zhang et al. [24], Li et al. [25], and Gharib et al. [26] leads to a FD-TTP trust level of 2, whereas Lai et al. [27] and de Ree et al. [28]'s key generation technique increases the FD-TTP trust level to 3.

5.2.4 Concluding Remarks

The development and design of a secure and decentralized key management solution for self-organizing networks has been a challenging task for over two decades. This also applies in the design of a decentralized key management solution that efficiently supports NC-MSCs, especially since NC-MSCs pose unique requirements. Based on these requirements, we found that the FD-TTP-based key management solution has the greatest potential, but neither of the proposed solutions have proven themselves yet to be robust against both mobile adversary and Sybil attacks.

There seem to be two main approaches remaining in developing a security system for NC-MSCs. The first approach relies on the development of an intricate access mechanism that relies on network operators deciding whether a mobile device can participate in NC-MSC-type communication and does not pose a threat relative to the Sybil attack. In such a case, this access mechanism can be combined with a key management solution that either follows the traditional PKI or CL-PKC cryptographic infrastructure. This could potentially be a redesigned scheme based on Luo et al. [10, 21]'s work or the work by de Ree et al. [28]. The second approach requires the mitigation of the Sybil attack through alternative methods. The only scheme that seemed to be capable of preventing such a Sybil attack was recently proposed by de Ree et al. [18], combining the TT-ID-PKC cryptographic infrastructure with a so-called private key cloaking technique.

5.3 Intrusion Detection and Prevention for NC-MSCs

The NC-enabled environment faces pollution attacks where malicious intermediate nodes manipulate packets in transition. These adjusted packets (i.e., polluted packets) will lead to incorrect decoding at the receivers. Therefore, identifying the polluted packets as well as the exact location of malicious users are similarly important tasks. However, many integrity schemes have been developed against pollution attacks [33–43], and only a few concentrate on identifying the exact location of malicious users [43–46].

In this section, an efficient intrusion detection and location-aware prevention (IDLP) mechanism is offered to detect pollution attacks and find the exact location of the adversary and prevent pollution attacks in NC-MSCs. The proposed IDLP mechanism is supplementary to our location-aware intrusion detection and prevention scheme (IDPS) scheme for NC-MSCs presented in [46]. For both the detection and locating schemes, the null space-based homomorphic message authentication code (MAC) scheme is applied [33], which is adjusted to the mobile small cell environment. The detection scheme enables the opportunity to detect pollution attacks effectively at the earliest possible node and drop the detected polluted packets. Still, this course of action is mostly inadequate since the adversaries can continue to pollute packets in the next transmission of coded packets of the same generation from the source to the destination node, leading to inefficient usage of network bandwidth. As a result, we focus on the identification of the adversaries' exact location and blocking them to protect the network from future pollution attacks.

5.3.1 State-of-the-Art of Intrusion IDPS for NC-Enabled Wireless Networks

Protecting against pollution attacks in NC-enabled networks chiefly depends on safeguarding the integrity of the packets in transition. Still, basic integrity schemes could not work with NC owing to the recoding of packets at intermediate nodes. Schemes that have a homomorphic property are necessary to guarantee the integrity of packets in NC-enabled networks. In this section, three topics are discussed: (i) secure NC, (ii) locating schemes, and (iii) IDPS schemes.

5.3.1.1 Secure Network Coding

Many different detection schemes have been developed against pollution attacks in NC-enabled networks, including information-theoretic schemes and cryptographic schemes such as homomorphic signature schemes or homomorphic MAC-based schemes. We concentrate on the homomorphic MAC-based schemes whose scope is to guarantee integrity in network coded packets, as introduced by Agrawal et al. [35]. Still, the schemes based on homomorphic MACs are susceptible to tag pollution attacks.

In [47], Zhang et al. studied the application of orthogonality property creating tags. Additionally, they solved the issue of tag pollution attack by combining a homomorphic signature to the MAC scheme leading to the proposed MacSig approach. Esfahani et al. enhanced the performance of these schemes over a sequence of works [33, 48, 49]. The work in [33] is focused on null space-based scheme where tags are mixed with the original packets according to a randomly produced swapping vector; this also reduces the probability of a successful tag pollution attack without additional overheads.

5.3.1.2 Locating Schemes

Recognizing the location of a malicious user is as important as the detection of the security attack, so that other participating nodes can be informed about the presence of an adversary. Therefore, an additional location scheme or verification of adversaries is necessary for maintaining a fair network environment. Siavoshani et al. [44] proposed an integrity scheme that also locates the adversary using a central controller. Another integrity scheme which discusses locating the adversary is SpaceMac [43]. In SpaceMac, a cooperative environment between parent and child nodes is considered. Lastly, a location-aware IDPS being able to not only detect and drop pollution attacks but also spot the attacker's exact location was suggested by Parsamehr et al. [46, 50]. The proposed IDPS is made up of detection and locating schemes according to null space homomorphic MAC.

5.3.1.3 IDPS Schemes

What mostly concerns IDPSs is the detection of potential security incidents, followed by blocking or preventing malicious activity. As far as detection is concerned, IDPSs apply a signature-based detection to recognize known adversaries in the networks that are not linked to legitimate users [51–54]. In our previous IDPS scheme [55], we offered for the first time innovative IDPS for network mobile small cells that are coding-enabled.

5.3.2 Energy-Efficient Intrusion Detection and Prevention for NC-MSCs

5.3.2.1 System Model

The IDLP mechanism consists of a detection scheme and a locating scheme which are both based on the null space homomorphic MAC scheme [33], and they are described in the following sections. This mechanism is divided into two phases for improving its efficiency in terms of resource consumption.

- ***Phase 1: Identification of the MSC where pollution attack occurred.*** In the first step, the detection scheme of the proposed IDLP mechanism is applied to all relay nodes (RNs) and destination nodes (DNs). When a pollution attack is detected by an RN or DN, it drops the polluted packet and sends a report to the hotspots of the MSCs that is associated to the reporter. The hotspot will forward the report to the SDN controller, which is responsible for identifying the MSC where a pollution attack occurred based on the received reports.
- ***Phase 2: Identification of the adversary node's location within the polluted MSC.*** The detection and locating schemes are applied to all mobile devices in the identified polluted MSC in phase 1. When a mobile device within the polluted MSC detects any pollution, they will drop the polluted packet and will send a report based on the locating scheme to the hotspot. The hotspot will forward it to the SDN controller to decide the most appropriate preventive action (e.g., block adversary mobile device(s) from accessing the network). Otherwise, the mobile device will create an expanded coded packet that is based on the received coded packet and the key shared between each mobile device and the SDN controller.

Then, the mobile device sends the expanded coded packet to the next node and the local hotspot. The hotspot then forwards this packet to the SDN controller.

5.3.2.2 Detection Scheme

According to [55] and [33], in the detection scheme of the proposed IDLP mechanism, the message is divided into a generation of native packets denoted as $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_m$ by the source node (SN), where m is the generation size and each packet $\boldsymbol{b}_i$ consists of n symbols (i.e., $b_{i,1}, b_{i,2}, \ldots, b_{i,n}$) in the finite field F_p^n. Therefore, the SN will generate a coded packet $\boldsymbol{b}_i$ according to Eq. 5.2 and send it to the next intermediate nodes.

$$\boldsymbol{b}_i = \left(\overbrace{\underbrace{0, \ldots, 0}_{i-1}, 1, 0, \ldots, 0}^{m}, b_{i,1}, b_{i,2}, \ldots, b_{i,n} \right) \in F_p^{m+n} \tag{5.2}$$

For simplicity, (5.2) can also be written as follows:

$$\mathbf{b}_\mathrm{i} = \left(\mathrm{b}_{\mathrm{i},1}, \mathrm{b}_{\mathrm{i},2}, \ldots, \mathrm{b}_{\mathrm{i},\mathrm{m}+\mathrm{n}} \right) \in \mathrm{F}_\mathrm{p}^{\mathrm{m}+\mathrm{n}} \tag{5.3}$$

As shown in (5.4), each intermediate node creates a new coded packet x which is a linear combination of h received coded packets ($\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_h$) and sends it to its neighbors. β_i is the coding coefficient which is chosen randomly from F_p, and all arithmetic operations are performed over the finite field F_p.

$$x = \sum_{i-1}^{h} \beta_i \boldsymbol{b}_1 \tag{5.4}$$

The L tags are generated based on null space properties [47] by the SN, for detecting pollution attacks. The following five steps are used to create the tags as well as to verify the orthogonality of the received coded packets:

1. Key distribution to the SN: A set of keys ($C_1, C_2, \ldots, C_L$) are created by the key distribution center (KDC) in the finite field F_p^{m+n+L}, and they are distributed in the SN.
2. The L tags (i.e., $t_1, t_2, \ldots, t_L$) are created using L keys for each coded packet by the SN, according to (5.5). Each coded packet is composed of $m + n$ symbols and L generated tags (i.e., t_{SN}).

$$\begin{bmatrix} C_{1,1} & \cdots & C_{1,m+n} \\ \vdots & \ddots & \vdots \\ C_{L,1} & \cdots & C_{L,m+n} \end{bmatrix}_{L\times(m+n)} \cdot \begin{bmatrix} \boldsymbol{b}_{i,1} \\ \vdots \\ \boldsymbol{b}_{i,m+n} \end{bmatrix}_{(m+n)\times 1} + \begin{bmatrix} C_{1,m+n+1} & \cdots & C_{1,m+n+L} \\ \vdots & \ddots & \vdots \\ C_{L,m+n+1} & \cdots & C_{L,m+n+L} \end{bmatrix}_{L\times L} \cdot \begin{bmatrix} t_1 \\ \vdots \\ t_L \end{bmatrix}_{L\times 1} \tag{5.5}$$

3. To avoid tag pollution attacks, the L tags are swapped based on the shared secret key (SV) between the SN and DNs according to (5.6).

$$\overline{\boldsymbol{b}_i} = Swap(\boldsymbol{b}_i)_{SV} \tag{5.6}$$

4. A set of new keys are created by the KDC using the swapping vector SV and based on the set of keys that were distributed to the SN in step 1 according to (5.7). Then, these keys are distributed to the intermediate nodes and DNs to verify the received coded packets.

$$C\prime_1 = Swap(C_i)_{SV} \tag{5.7}$$

5. Finally, the received coded packet is verified by each intermediate node and DN based on (5.8).

$$\delta = Swap(C_i)_{SV} \cdot Swap(\boldsymbol{b}_i)_{SV} = \sum_{j=1}^{m+n+L} C'_{i,j} \cdot \overline{\boldsymbol{b}}_{i,j} \tag{5.8}$$

If $\delta = 0$, then the received coded packet is verified and acceptable to transmit the next nodes. Otherwise, it is dropped.

5.3.2.3 Locating Scheme

The locating scheme identifies the exact location of the adversary mobile node within the polluted MSC. In this step, each mobile node is responsible for a) generating an expanded coded packet, based on the received coded packet, and transmitting it to the next node and hotspot as well and b) sending a report to the hotspot when a polluted packet is detected through the detection scheme within the polluted MSC. Both the expanded coded packet and the report are forwarded to the SDN controller which is responsible for identifying the exact location of the adversary.

Expanded Coded Packet

An extra tag is added to each coded packet by each intermediate node for verifying itself to the SDN controller. This tag is created based on the pre-distributed shared key between each node and the SDN controller. This tag is calculated based on the following equation:

$$\begin{bmatrix} C''_{1,1} \\ \vdots \\ C''_{1,m+n} \\ \vdots \\ C''_{1,m+n+L} \end{bmatrix}^{T}_{1\times(m+n+L)} \cdot \begin{bmatrix} \boldsymbol{b}_{i,1} \\ \vdots \\ \boldsymbol{b}_{i,m+n} \\ t_1 \\ \vdots \\ t_L \end{bmatrix}_{(m+n)\times 1} + C''_{1,m+n+L+1} \cdot s_i = 0 \tag{5.9}$$

The vector $\left[C''_{1,1} \cdots \; C''_{1,m+n} \cdots C''_{1,m+n+L} \right]_{1\times(m+n+L)}$ is the pre-shared key distributed by the KDC, and s_i is the properly calculated tag.

The SDN controller verifies the received expanded coded packet $\{\boldsymbol{b}_i||t_{SN}||s_i\}$ based on the following formula, where $\boldsymbol{b}_i$ is the coded packet, t_{SN} represents the set of appended tags by SN, and s_i is the appended tag by the given intermediate node. If $\delta = 0$, then the received expanded coded packet is verified.

$$\delta = \sum_{j=1}^{m+n+L+1} C''_{i,j} \cdot \overline{\left\{\boldsymbol{b}_{i,j} \,||t_{SN}||\, s_i\right\}} \tag{5.10}$$

Report

When a polluted packet e signed by the previous mobile device's key ($\{e\|s_{i-1}\}$) is detected, a report is generated by the intermediate node or a DN, who detects pollution. The generated report is the received polluted packet ($\{e\|s_{i-1}\}$) signed by the given node and is represented as $\{e||s_{i-1}||s_i\}$.

In the following equation, if $\delta = 0$, then the sender is verified by the SDN controller.

$$\delta = \sum_{j=1}^{m+n+L+1} C'{}'_{i,j} \cdot \overline{\{e\,||s_{i-1}||\,s_i\}} \tag{5.11}$$

Then the signature of the adversary node is verified if $\delta = 0$.

$$\delta = \sum_{j=1}^{m+n+L+1} C'{}'_{i,j} \cdot \overline{\left\{e \middle\| s_{i-1}\right\}} \tag{5.12}$$

5.3.3 Implementation

Throughout this section, we discuss the process of implementation related to the proposed IDLP mechanism, and we compare it with our previous IDPS scheme [55], being the first time that a new scheme for detection and prevention of intrusions was proposed for NC-enabled mobile small cells. Firstly, 3 butterfly topologies were implemented, consisting of 18 normal nodes and 1 opponent node (see Fig. 5.2), and then the random linear network coding (RLNC) approach was applied. Furthermore, in our implementation, the adversary node was programmed to adjust its received packets in order that it could demonstrate a pollution attack.

The implementation is based on the recoding library of Kodo, which made it possible to encode at the SN, recode at the intermediate nodes, and decode at the destination nodes [56]. Kodo has some restrictions with creating a customized generation of packets and keys and also with tag generation. Thus, MATLAB was used in our implementation to generate the packets, their proper tags, and the required keys at the source node and intermediate nodes (these were included manually in Kodo in order to achieve the desired functionality of the implemented scenario).

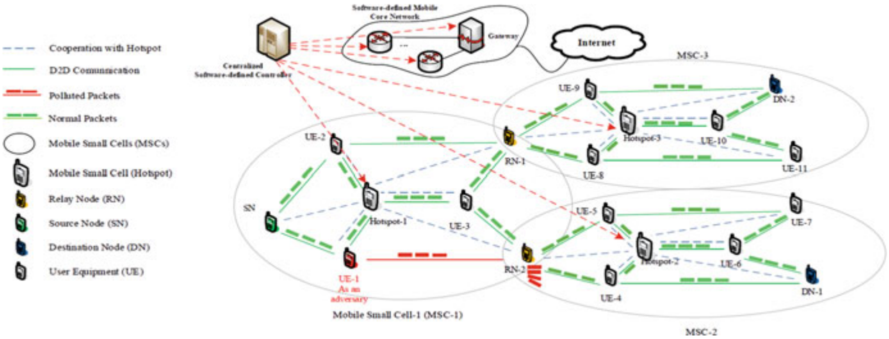


Fig. 5.2 Implemented three butterfly topologies

The size of packet generation has been designated to be 64 symbols, and the symbol size is fixed between 1,000 and 10,000 bytes. Additionally, the quantity of tags attached to the end of each packet is L, which can only be 27, 42, or 54 [47], where the Galois field is $GF2^8$. Lastly, it should be considered that the machine being used for running the entire implementation comes with the following characteristics: a 2.7 GHz Core i7 CPU with 8GB of physical memory.

5.3.4 Performance Evaluation

Throughout this section, we provide the performance evaluation of the proposed IDLP mechanism based on computational and communication overheads, along with the successful decoding probability. It is worthy to reiterate that the proposed IDLP mechanism along with detection of pollution attacks also detects the exact site of the attacker(s) and selects the most suitable preventive approach (e.g., blocking the mobile device being at risk from gaining access to the network) to stop and protect network resources. This will be compared with our baseline IDPS [55] that only detects and drops the polluted packets, where intruders continue to create pollution attacks that result in wastage of network resources.

5.3.4.1 Computational Overhead

It must be noted that the overall timeline from when the packet is generated until the packet is confirmed and decoded at the destination nodes is shown in the following equation:

$$T_{\text{total}} = T_{\text{enc}} + T_{\text{rec}} + T_{\text{dec}} + T_{\text{ver}} \tag{5.13}$$

In this equation, the encoding time at the source node is called T_{enc}, the recoding time at each intermediate node is called T_{rec}, the decoding time at the destination node is called T_{dec}, and the verification time at the intermediate and destination nodes is called T_{ver}.

The T_{total} for the baseline IDPS [55] and the proposed IDLP mechanism are demonstrated in Fig. 5.3.

This figure contains three curves based on the quantity of tags (i.e., $L \in \{27,42,54\}$) for each method. As can be observed, through increasing the quantity of tags, the T_{total} increases almost linearly. Nevertheless, the T_{total} for a different number of tags in the proposed IDLP mechanism (e.g., $T_{\text{total}} = 0.20$ for $L = 54$ when the length of the packet is 10,000 bytes) is below the T_{total} of the baseline IDPS scheme (e.g., $T_{\text{total}} = 0.22$ for $L = 54$ when the length of the packet is 10,000 bytes) [55].

It is worth to note that the reason why the IDLP T_{total} drops below that of [55] is that in addition to the novel IDLP mechanism delivering not only detection and

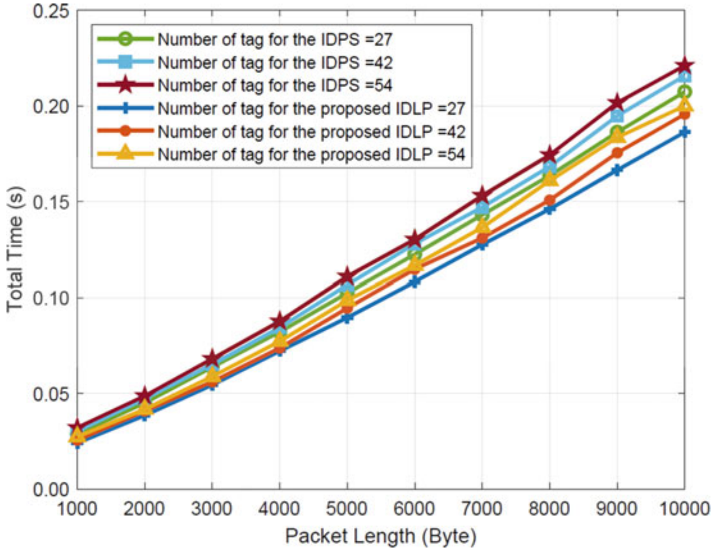


Fig. 5.3 T_{total} for different number of tags in [55] and the IDLP

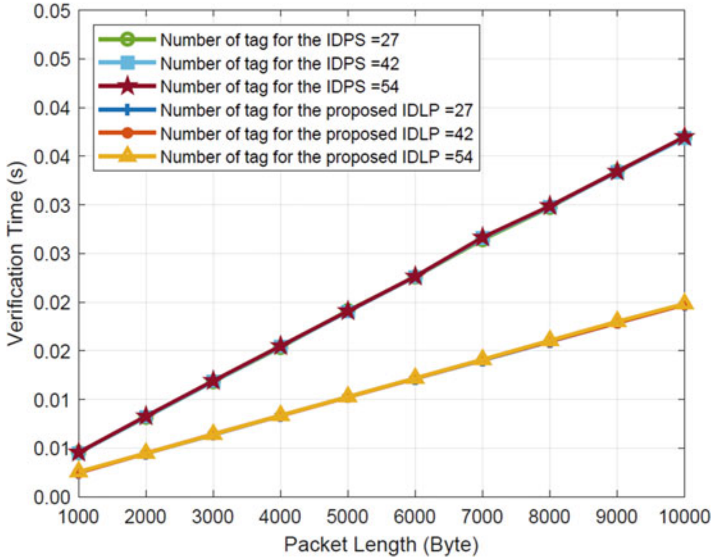


Fig. 5.4 The T_{ver} for different number of tags in [55] and the IDLP

location capability, there are fewer operational costs since it is not applied at every intermediate node.

Furthermore, the verification and detection time for any corrupted packet in the network for both IDLP schemes is given by Fig. 5.4.

As it can be seen, the proposed location-based scheme is more competitive than the baseline. On the other hand, it should be mentioned that the IDLP mechanism inherently detects and drops the polluted packet as well as detects the exact site of the attacker(s) and blocks them from the network.

5.3.4.2 Communicational Overhead

The communication time, T_{comm}, is defined as the communication overhead of the proposed IDLP mechanism. Figure 5.5 displays the T_{comm} according to the various numbers of tags being used for both IDPS schemes.

The results again substantiate that T_{comm} for the proposed IDLP is below the baseline IDPS value [55]. The difference is due to the fact that the proposed IDLP mechanism blocks the opponents, and therefore they are no longer capable of adjusting the packets in transit. Thus, the SN is not required to resend packets.

5.3.4.3 Decoding Probability

The probability that a corrupted packet is not detected in the verification phase is called P_r. The P_r for the proposed and baseline IDPS based on three different number of tags ($L \in \{27,42,54\}$) is shown in Fig. 5.6 As can be seen, the proposed IDLP mechanism exhibits a P_r value of almost 0. Nevertheless, the IDPS proposed in [55] is very close to 0.

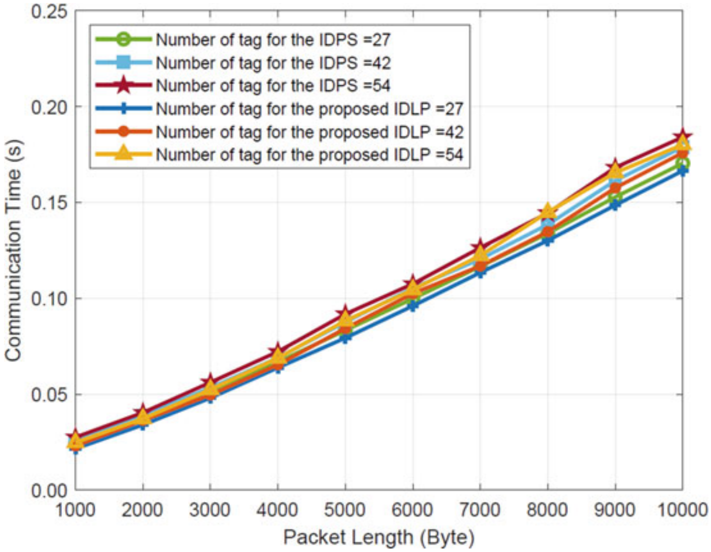


Fig. 5.5 The T_{comm} for different number of tags in [55] and the IDLP

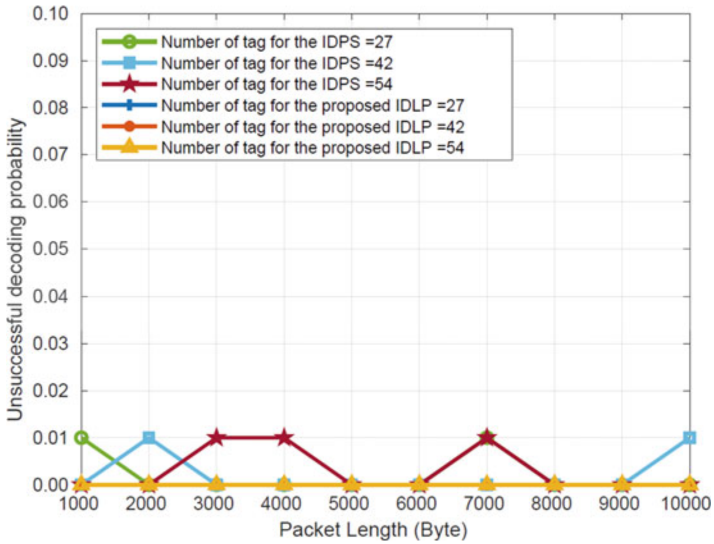


Fig. 5.6 The P_r for different number of tags in [55] and the IDLP

In other words, the proposed IDPS approach does not allow the adversary the opportunity to distribute the corrupted packet in the network. While the baseline scheme can still inject pollution in the next transmission of the coded packet from the SN to DNs in the network, in the most novel approach, the detected adversaries are blocked altogether from gaining access to the network.

5.3.5 Concluding Remarks

This study offered an effective IDLP mechanism for NC-MSCs. The proposed IDLP mechanism builds on our previous effort [46] that now is not only able to detect the pollution attack but is also context aware and able to remove the enemy from the network. The null space-based homomorphic MAC scheme [33] for both the detection and locating schemes is used, which is adjusted for UDNs, that is pivotal for next-generation networks. The proposed IDLP mechanism is able to detect the attacker's precise site and selects the best preventive approaches (e.g., blocking compromised mobile device from gaining access to the network) to defend the network resources. It is worth mentioning that the proposed IDLP mechanism is more effective compared to the baseline IDPS scheme proposed in [55], since it omits the need to operate on all mobile devices to protect the NC-MSCs from depleting their resources. In particular, simulation results have shown that the proposed IDPS approach is more effective than the baseline counterpart in terms of lower computational complexity, communicational overhead, and unsuccessful decoding probability.

5.4 Blockchain-Based Integrity Scheme for NC-MSCs

5.4.1 Introduction

Pollution attacks are considered as one of the major security challenges in NC-enabled networks and raise concerns regarding the adaptation of NC to practical use in beyond 5G networks. Cryptographic-based integrity schemes are proposed to detect and prevent pollution attacks in NC-enabled networks. In this section, we describe a blockchain-based integrity scheme against pollution attacks.

5.4.1.1 Pollution Attacks

Allowing intermediate nodes to code or recode the packets is the key feature of NC. However, this ability of intermediate nodes to change the packets in transition introduces the security challenge regarding pollution attacks. An adversary node can inject a corrupted packet instead of a genuine packet, and this will pollute the entire information flow to which the corrupted packet is introduced or mixed with [57]. Thus, a single polluted packet can significantly reduce the throughput of the network. Further, if the polluted packet is not detected, it will be used while recoding at a genuine node which will pollute more packets in transition. Thus, identifying pollution attacks at the earliest possible node is an important requirement. On the other hand, challenges in detecting pollution attacks are manifold. Since the polluted packets can be identical to the original packet in packet size and characteristics, pollution attacks can only be identified by verifying the integrity of packets. However, as per the principles of NC, the packets are coded at the intermediate nodes, and these coded packets are sent over the network. This rules out the possibility of using generic integrity schemes in NC-enabled scenarios. However, integrity schemes with homomorphic property over NC-related operations are developed to detect pollution attacks.

5.4.1.2 Integrity Schemes

Integrity schemes with the homomorphic property over NC-related operations are widely used to detect and prevent pollution attacks in NC-enabled environments [58]. Such integrity schemes using homomorphic MACs are first proposed in [35]. The homomorphic MAC-based integrity schemes are computationally less complex compared to the homomorphic signatures and hash functions [35]. However, MAC-based integrity schemes require a set of shared secret keys to be available for all the participating nodes. The source node will create the MACs using the keys accessible to it and attach them as tags to the packets. A receiving node with at least one of those keys is used to create the tags and can verify the integrity of the packets. If a node could not verify the received tag, then it will discard the packet as a polluted packet since the integrity cannot be verified. This also leads to another version of pollution attack called tag pollution. In tag pollution attack, the adversary node intentionally attaches a non-verifiable tag to a genuine packet so that it will be discarded at the next genuine node and thus reduces the throughput. An efficient integrity scheme should be able to prevent both data pollution and tag pollution attacks. However, this MAC-based integrity scheme introduces some computational complexity and bandwidth overhead to the system. Furthermore, proper key distribution is mandatory to ensure the security of these schemes. However, emerging technologies, such as blockchain, conceptually engineered as a type of distributed and immutable ledger [59] can be an approach to offer integrity services using MACs.

5.4.1.3 Blockchain Applications

The concept of blockchain evolved as a research area after it was used in the Bitcoin cryptocurrency. However, this immutable distributed ledger is being studied and used in a variety of other applications in the current digital era [60]. In our proposed scheme, we use blockchain as an immutable, distributed, and decentralized ledger for tag sharing. There are multiple blockchains with different characteristics and requirements. One of the major categorizations of blockchain is based on the method of verifying the block and achieving consensus among the blockchain nodes. Initial blockchains were employing a block verification scheme called proof of work (PoW) where all nodes will compete to verify a block by achieving a very hard cryptographic hash. However, this approach is highly resource-consuming. Proof of stake (PoS)-based block verification schemes are introduced by different blockchains [61, 62] to reduce the energy and computational requirements associated with the blockchain.

5.4.2 Blockchain-Based Integrity Scheme

As we progress toward the 5G and beyond networks, small cells will be an integral part of the network architecture to provide quality broadband services for remote and isolated devices. Furthermore, D2D communication using side-link channels are already being discussed as the part of 5G in [63]. These future networks are expected to serve a very dense network of heterogeneous devices with high data rates and low energy. Considering these requirements, integrity schemes for NC-enabled small cells need to be scalable and maintain a low computational and bandwidth requirement. Toward this extent, a blockchain-based integrity scheme was proposed in [64].

5.4.2.1 Enabling Technologies for Integrity Schemes

Most of the integrity schemes in the pre-5G era were having some level of dependency on the network size for the number of tags and the security that can be achieved with a specific number of tags [47, 48]. The number of SNs and the number of neighboring or intermediate nodes in the network are important parameters in defining the security level and key distribution of these approaches and discourage scaling up of these integrity schemes to the small cell environments. [65] presented an integrity scheme that addressed these challenges by presenting one of the initial integrity schemes for NC-enabled small cell environments. However, this work considers a secure central controller connected to all participating nodes. This proposed approach creates tags at the SN and shares it with the centralized entity. Moreover, these tags are also attached to the packets before transmission, and a DN can verify the integrity of the packets by verifying the tag against the packet payload data, as well as verifying the authenticity of the tags by comparing the tags stored at the central entity. In other words, the tags are shared through a secure secondary channel with the receiving nodes.

This approach addresses the challenges of scalability by ensuring that, even if all keys are available to an adversary, the adversary still cannot modify the tags registered at the central authority by the SN. Furthermore, the number of tags required to achieve sufficient and equivalent security compared to the existing integrity schemes was smaller, providing a lower bandwidth and computational overhead. However, this integrity scheme still suffered from other challenges like a single point of failure and requirement of a secure channel from the controller to all participating nodes. The integrity scheme was highly dependent on the security and trustworthiness of the central controller and on the control channel between this central controller and the participating nodes. Moreover, assigning the central controller a major role in the security framework will attract more attacks (honeypot syndrome), and if the adversary can compromise this single entity, the whole system will collapse. To address this challenge, a distributed and decentralized tag sharing scheme was proposed by the authors [64]. This integrity scheme uses a blockchain as the immutable distributed data ledger to share the tags. The blockchain is inherently secure against modifications and distributed in nature and allows all nodes to fetch the required information from the verified blocks.

5.4.2.2 System Model

The blockchain-based integrity scheme uses MACs to ensure the integrity of the packets. It shares MACs through the blockchain such that the receiving nodes can verify the authenticity of the MACs received with the packets by comparing them with the MACs retrieved from the blockchain. However, creating a blockchain network involving all the participating nodes may not be feasible in a dense heterogeneous network [66]. Thus, the proposed system model considers a blockchain overlay of small cells, where only the small cell heads will be part of the blockchain as full nodes. Other end-devices can fetch the verified blocks from its corresponding small cell head and send candidate transactions (in our case, the tags from the SN). Only these full nodes will participate in the block verification process and store the blockchain entirely. The proposed system architecture is presented in Fig. 5.7.

This blockchain-based integrity scheme uses a MAC which is homomorphic to the RLNC operations. We consider a small cell scenario where devices are capable of D2D multi-hop communication using RLNC principles. Thus, the SN will consider a generation of m packets where each packet P_i can be considered as a vector of n elements. In our scheme, the MACs are created over the native packet differing from the previous integrity schemes where the MACs are created over the augmented (coded) packets. Thus, the key size for our integrity scheme depends only on the packet size and is independent of the generation size. If K_i is one of the keys from the key set, it shall have $n + 1$ elements in it. The tag T_{ij} on packet P_i using the key K_j is created as follows:

$$T_{ij} = \frac{\sum_{l=1}^{n} P_{il} \times K_{jl}}{K_{jl+1}} \tag{5.14}$$

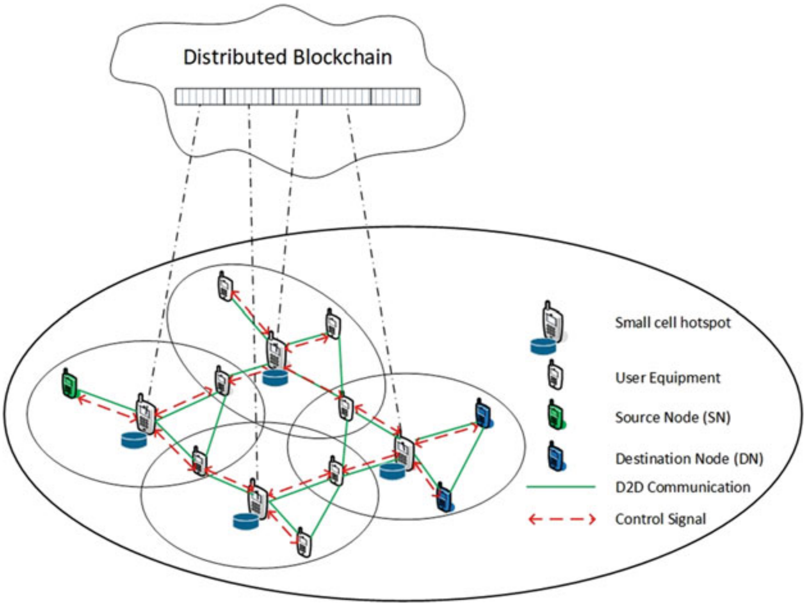


Fig. 5.7 Blockchain-based small cell architecture

These tags are sent to the blockchain as a candidate transaction and also attached to the packets before encoding. These augmented packets will be considered as a normal packet for encoding. At a receiver node, the authenticity of the tags can be verified by comparing the tags received through the communication channel along with the packets and the tags retrieved from the blockchain. Furthermore, the integrity of the packets is verified by recreating the tags using the key set available with the nodes on the received packet. It is to be noted that our scheme enables the intermediate nodes to simply recode the packet as a normal RLNC packet with no specific algorithm required for tag combining. This is because of the homomorphic property of the proposed tag creation scheme over RLNC operations, which also reduces the computations required at the intermediate node.

5.4.2.3 Security Approach

The level of security against pollution attacks using MACs will depend upon the field size used for the operations. In most of the practical applications of RLNC, a Galois field of size 2^8 is used. If q is the field size, then a single tag attached to the packet will ensure that the probability of an adversary successfully introducing a polluted packet to the network and passing through the tag verifications is $1/q$. Practically, this security may not be sufficient for all the applications, and we generally use multiple tags to increase the security of the integrity scheme. If l number of tags are attached to the packets, then we can achieve a security level of $1/q^l$. Thus, in a Galois field of size 2^8, a single tag attached to the packet provides a security level of $1/2^8$ against pollution attack.

Another main factor that affects the security of the integrity scheme and the number of tags is the probability of colluding adversaries. Multiple adversaries may cooperate to successfully bypass the integrity check. Most of the previously existed integrity schemes addressed this challenge by employing specific key distribution schemes so that the participating nodes may not have all the keys used by the SN to create tags. For example, a c-cover free set system-based key distribution is presented in [48, 67] where the SN should have at least c times the number of keys than any other participating node to achieve security against c colluding attackers. However, such methods have multiple drawbacks and scalability issues. In such cases, the number of tags attached to each packet will also depend on the probability of colluding attackers.

In a dense environment, the probability of colluding attackers is very high, and it will also increase the number of tags required to achieve a high level of security. Furthermore, the overhead due to this increased number of tags will not provide equivalent security to the number of tags attached, but only equal to the number of tags that can be verified at a particular node. In other words, even if L number of tags are attached to the packets by the source node and transmitted across the network, a receiving node that holds only l keys can verify only that many tags and thus provide a security level of $1/q^l$ only. This results in a mismatch between the bandwidth overhead of the system and the security level. Our proposed approach addresses these problems by sharing the tags not only through the communication channel but also at the blockchain such that an adversary cannot modify the packets and create valid tags even if they have all the secret keys. In our scheme, we consider a strong adversary that can possess all the secret keys that are used for tag creation. However, since the original tags are shared directly by the SN to the blockchain, an adversary cannot modify the tags that are stored in the blockchain. Thus, even if it creates valid tags for a polluted packet, the next genuine receiver will discard the packet since it will not match with the corresponding tags stored in the blockchain. This situation does not differ even if multiple adversaries are colluding to bypass the integrity check. Since our integrity scheme does not require any specific key distribution protocol that depends on the number of adversaries in the network, we can allow all the users to have the complete key set and verify all the tags that are attached to the packets. Thus, in our integrity scheme, the bandwidth overhead due to the tags for integrity check is proportional to the security level of the scheme. This will also allow the system to achieve a high security level solely depending on the number of tags attached to it. An analysis of the number of tags and security level is presented in the performance evaluation section.

5.4.2.4 Lightweight Scheme

As we discussed in the previous section, the security of our integrity scheme depends completely and solely on the number of tags attached to it. Increasing the number of tags will also increase the computational and bandwidth overhead of the system. A trade-off between the level of security and the overheads should be considered before any practical implementations. Furthermore, the security requirements of the system will vary depending on the applications and strictness of authentication schemes in the network. For example, in a restricted company network where all users are authenticated with strict verifications, the probability of an adversary node is lower, and we can reduce the number of tags (thereby reducing the security level and overheads). However, in a public Wi-Fi network, we may have to use a higher number of tags to provide strict integrity checks at every node involved in the transmission. Considering these aspects, we propose a lightweight version of our proposed integrity scheme for the applications where the security requirements are lower.

This lightweight scheme is proposed in two parts. In the first part, we reduce the bandwidth overhead over the communication channel by sharing the tags only through the blockchain. In this way, the SN creates tags over the packet and sends it only to the blockchain. When a node receives this packet, it can fetch the corresponding tags from the blockchain and verify the integrity of the packet. This approach removes any extra overhead due to the security scheme from the communication channel without compromising the security of the scheme. It also eliminates the probability of tag pollution attacks since the intermediate nodes do not send the tags with the packet. In the second part, we compromise on the security level that can be achieved at an intermediate node, but with the advantage of reduced computational complexity by reducing the number of tags verified at that node. Here, we choose an adaptive scheme considering the security requirements of the network and then decide on the number of tags to be verified at each node. Even though every node will have access to all the keys and tags, it will only verify a random number of tags to reduce the computational requirements.

5.4.2.5 Performance Evaluation

In this section, we evaluate the performance of both the proposed blockchain-based integrity scheme and its lightweight version by comparing its computational complexity and security level. The schemes are simulated in a hybrid environment using the KODO RLNC library [56] for NC-related operations and BigchainDB [62] as the blockchain environment. The integration of these environments is enabled by Postman, a cross-platform integration environment [68]. We consider the field size $q = 2^8$, packet size of 1000 bytes, and generation size of 32 for simulations unless specified otherwise in the overhead analysis.

The computational complexity of the proposed approach depends on the tag creation process. As defined by Eq. 5.14, the creation of a single tag requires $n + 1$ multiplications where n is the number of vector elements in a packet. Since the receiving nodes also perform the same operation to verify the integrity, the complexity of verifying a tag is also similar. Thus, if L tags are created at the source, then the total complexity of tag verification is $L \times (n + 1)$ for the basic integrity scheme. In the lightweight version, if we verify only l tags, the complexity is reduced to $l \times (n + 1)$. On the other hand, increasing the number of tags will also result in increased security levels. Figure 5.8 shows the trade-off between security level and computational complexity in the specified simulation scenario. This evaluation shows that using eight tags can be considered as an optimal situation for the specified simulation environments. Figure 5.9 shows the comparison of computational complexity against packet size with eight tags per packet for the full version and a lightweight version where only one tag is verified at every node.

5.4.3 Concluding Remarks

Detecting and preventing pollution attacks in the NC environment is one of the main obstacles in practical NC implementations. NC-MSC proposed for the 5G and beyond era expects to serve UDNs as well. Efficient integrity schemes with scalability and minimum overhead requirements are required to address this scenario. In this section, we discussed a blockchain-based integrity scheme using homomorphic MACs tailor-made to address the challenges in NC-enabled small cells for highly dense network environments. The proposed approach is independent of the size of the network and the probability of attackers in the network, as well as achieves a high level of security using a small number of tags. Moreover, a lightweight

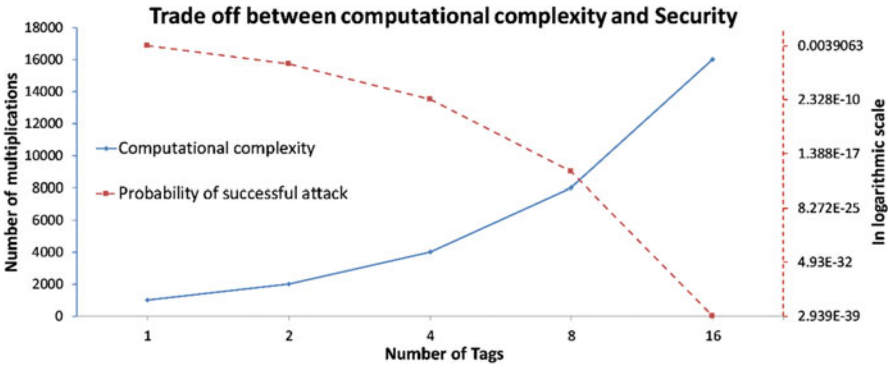


Fig. 5.8 Trade-off between complexity and security

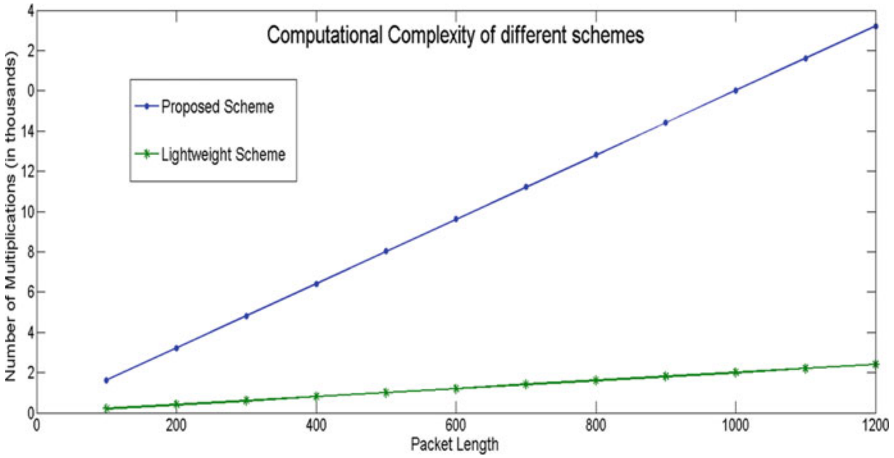


Fig. 5.9 Computational complexity of proposed schemes

version of the integrity scheme, adaptable to the network security requirements, is also presented. The performance analysis shows the computational efficiency of the proposed schemes. Furthermore, it also presents a trade-off comparison between the computational complexity and security levels to identify the optimal number of tags that can be used in practical applications. This integrity scheme can be considered as a baseline for secure NC against pollution attacks in the 5G and beyond networks.

5.5 Conclusion

The next-generation mobile network is adopting small cell technology to provide network subscribers with a high QoS, including high data transmission speeds and low latency. However, increasing demands of mobile data in a dense urban environment may still face issues in the future. A cost-effective solution to this problem is provided by the EU-funded H2020-MSCA project "SECRET" in the form of a networking architecture that utilizes NC-MSCs. However, a vast array of security challenges must be addressed prior to its adoption. This chapter addressed the security challenges from the perspective of key management, intrusion detection and prevention, and data integrity.

Various cryptographic security solutions (i.e., encryption schemes, integrity schemes) rely on a key distribution mechanism. However, it was pointed out that UDNs may be incapable of relying on a centralized TTP to organize the secure distribution of cryptographic keys. The recent survey [4] pointed out that a decentralized key management based on the FD-TTP approach is the most suitable solution. Additionally, the FD-TTP-based approach is agnostic to NC-enabled networks. The exploration of existing FD-TTP-based key management schemes shows that the majority of schemes follow the same key management structure, except for two recently proposed solutions [18]. These recently proposed solutions seem capable of having a significant impact on key management performance and security. However, these benefits are still to be validated in practical systems. An important piece missing from many key management solutions is the ability to counter a Sybil attack and has to be addressed to provide a robust security system. If the Sybil attack can be mitigated through the network operator involvement, then the solution by de Ree et al. [28] is also a suitable candidate.

Recently, Parsamehr et al. [50, 55] designed novel intrusion detection and prevention schemes. These schemes are designed for UDNs, in particular NC-MSCs, that aim to remove malicious users that launch potential data pollution attacks. These works have a twofold objective, (i) the rapid detection of polluted data packets and (ii) the identification of the malicious user. Furthermore, they are capable of selecting the best approach to prevent that user from future malicious behavior. Both of the schemes have been implemented in Kodo, where the latest IDLP scheme [50] demonstrated the best performance in terms of computational complexity, communicational overhead, and successful decoding. Therefore, the IDLP scheme is recommended as the most suitable intrusion detection and prevention scheme to secure NC-MSCs.

An alternative approach to preventing pollution attacks is through data integrity schemes. This chapter discussed a blockchain-based integrity scheme that is capable of effectively preventing pollution attacks in UDNs. Furthermore, the blockchain-based integrity scheme is scalable, meaning that the number of attackers or the size of the network has no effect on its performance. It was demonstrated that the security requirement is directly related to the computational complexity. Therefore, a lightweight version of the blockchain-based integrity scheme can be adopted for UDNs with a reduced security requirement. It was evident from the results that their integrity performance can lend this solution to be an effective approach to secure next-generation UDNs, utilizing the NC paradigm against pollution attacks.

Acknowledgments This research work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 722424 (SECRET) and the NATO Science for Peace and Security Programme under grant SPS G5797 (PHY-SEC).

References

1. WikiLeaks. *Vault 7* [Online]. Available: https://wikileaks.org/ciav7p1/
2. Rodriguez, J., Radwan, A., Barbosa, C., Fitzek, F. H. P., Abd-Alhameed, R. A., Noras, J. M., Jones, S. M. R., Politis, I., Galiotos, P., Schulte, G., Rayit, A., Sousa, M., Alheiro, R., Gelabert, X., & G. Koudouridis (2017). SECRET – Secure network coding for reduced energy next generation mobile small cells: A European training network in wireless communications and networking for 5G. In *Proceedings of the 7th International Conference on Internet Technologies and Applications (ITA)*, Wrexham, UK.
3. de Ree, M., Mantas, G., Radwan, A., Rodriguez, J., & Otung, I. E. (2018). Key management for secure network coding-enabled mobile small cells. In *Proceedings of the 9th International Conference on Broadband Communications, Networks and Systems (BROADNETS)*, .
4. de Ree, M., Mantas, G., Radwan, A., Mumtaz, S., Rodriguez, J., & Otung, I. E. (2019). Key management for beyond 5G mobile small cells: A survey. *IEEE Access, 7*, 59200–59236.
5. Capkun, S., Buttyán, L., & Hubaux, J.-P. (2003). Self-organized public-key management for mobile Ad Hoc networks. *IEEE Transactions on Mobile Computing, 2*(1), 52–64.
6. Capkun, S., Hubaux, J.-P., & Buttyán, L. (2006). Mobility helps peer-to-peer security. *IEEE Transactions on Mobile Computing, 5*(1), 43–51.
7. Li, X., Gordon, S., & Slay, J. (2004). On demand public key management for wireless Ad Hoc networks. In *Proceedings of the Australian Telecommunication Networks and Applications Conference (ATNAC)*, Sydney, NSW, Australia.
8. He, W., Huang, Y., Nahrstedt, K., & Lee, W. C. (2009). SMOCK: A scalable method of cryptographic key management for mission-critical wireless Ad-Hoc networks. *IEEE Transactions on Information Forensics Security, 4*(1), 140–150.
9. Zhou, L., & Haas, Z. J. (1999). Securing Ad Hoc networks. *IEEE Network, 13*(6), 24–30.
10. Luo, H., Kong, J., Zerfos, P., Lu, S., & Zhang, L. (2004). URSA: Ubiquitous and Robust access control for mobile Ad Hoc networks. *IEEE/ACM Transactions on Networking, 12*(6), 1049–1063.
11. Shamir, A. (1979). How to share a secret. *Communications of the ACM, 22*(11), 612–613.
12. Feldman, P. (1987). A practical scheme for non-interactive verifiable secret sharing. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science (SFCS)*, Los Angeles, CA, USA.
13. Pedersen, T. P. (1991). Non-interactive and information-theoretic secure verifiable secret sharing. In *Proceedings of the 11th Annual International Cryptology Conference (CRYPTO)*, Santa Barbara, CA, USA.
14. Ostrovsky, R., & Yung, M. (1991). How to withstand mobile virus attacks. In *Proceedings of the 10th ACM Symposium on Principles of Distributed Computing (PODC)*, Montreal, QC, Canada.
15. Herzberg, A., Jarecki, S., Krawczyk, H., & Yung, M. (1995). Proactive secret sharing OR: How to cope with perpetual leakage. In *Proceedings of the 15th Annual International Cryptology Conference (CRYPTO)*, Santa Barbara, CA, USA.
16. Jarecki, S. (1995). *Proactive secret sharing public key cryptosystems*. MS thesis. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.
17. Douceur, J. R. (2002). The Sybil attack. In *Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS)*, Cambridge, MA, USA.

18. de Ree, M., Mantas, G., Gao, J., Rodriguez, J., & Otung, I. E. (2020). Public key cryptography without certificates for beyond 5G mobile small cells. In *Proceedings of the 12th IEEE/IET International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP)*, Porto, Portugal.
19. Girault, M. (1991). Self-certified public keys. In *Proceedings of the 10th Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, Brighton, UK.
20. Gennaro, R., Jarecki, S., Krawczyk, H., & Rabin, T. (2007). Secure distributed key generation for discrete-log based cryptosystems. *Journal of Cryptology, 20*(1), 51–83.
21. Luo, H., & Lu, S. (2000). *Ubiquitous and Robust Authentication Services for Ad Hoc wireless networks* (UCLA-CSD-TR-200030). Los Angeles: University of California.
22. Deng, H., & Agrawal, D. P. (2004). TIDS: Threshold and identity-based security scheme for wireless Ad Hoc networks. *Ad Hoc Networks, 2*(3), 291–307.
23. da Silva, E., & Albini, L. C. P. (2013). Towards a fully self-organized identity-based key management system for MANETs. In *Proceedings of the 9th IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Lyon, France.
24. Zhang, Z., Susil, W., & Raad, R. (2008). Mobile Ad-Hoc network key management with certificateless cryptography. In *Proceedings of the 2nd International Conference on Signal Processing and Communication Systems (ICSPCS)*, Gold Coast, QLD, Australia.
25. Li, F., Shirase, M., & Takagi, T. (2008). Key management using certificateless public key cryptography in Ad Hoc networks. In *Proceedings of the 5th IFIP International Conference on Network and Parallel Computing (NPC)*, Shanghai, China.
26. Gharib, M., Moradlou, Z., Doostari, M. A., & Movaghar, A. (2017). Fully distributed ECC-based key management for mobile Ad Hoc networks. *Computer Networks, 113*, 269–283.
27. Lai, J., Kou, W., & Chen, K. (2011). Self-generated-certificate public key encryption without pairing and its application. *Information Sciences, 181*(11), 2422–2435.
28. de Ree, M., Mantas, G., Rodriguez, J., & Otung, I. E. (2021). DISTANT: Distributed trusted authority-based key management for beyond 5G wireless mobile small cells. Computer Communications. (Accepted for publication). https://authors.elsevier.com/tracking/article/details.do?aid=6806&jid=COMCOM&surname=de%20Ree
29. Saeednia, S. (2003). A note on Girault's self-certified model. *Information Processing Letters, 86*(6), 323–327.
30. Narasimha, M., Tsudik, G., & Yi, J. H. (2003). On the utility of distributed cryptography in P2P and MANETs: The case of membership control. In *Proceedings of the 11th IEEE International Conference on Network Protocols (ICNP)*, Atlanta, GA, USA.
31. Saxena, N. (2006). Public key cryptography Sans certificates in Ad Hoc networks. In *Proceedings of the 4th International Conference on Applied Cryptography and Network Security (ACNS)*, Singapore, Singapore.
32. Al-Riyami, S. S., & Paterson, K. G. (2003). Certificateless public key cryptography. In *Proceedings of the 9th International Conference on the Theory and Applications of Cryptology and Information Security (ASIACRYPT*), Taipei, Taiwan.
33. Esfahani, A., Mantas, G., & Rodriguez, J. (2016). An efficient null space-based homomorphic MAC scheme against tag pollution attacks in RLNC. *IEEE Communications Letters, 20*(5), 918–921.
34. Parsamehr, R., Mantas, G., Radwan, A., & Rodriguez, J. (2018). Security threats in network coding-enabled mobile small cells. In *Proceedings of the 9th International Conference on Broadband Communications, Networks and Systems (BROADNETS)*, Faro, Portugal.
35. Agrawal, S., & Boneh, D. (2009). Homomorphic MACs: MAC-based integrity for network coding. In *Proceedings of the 7th International Conference on Applied Cryptography and Network Security (ACNS)*, Paris-Rocquencourt, France.
36. Fiandrotti, A., Gaeta, R., & Grangetto, M. (2019). Securing network coding architectures against pollution attacks with band codes. *IEEE Transactions on Information Forensics and Security, 14*(3), 730–742.

37. Jaggi, S., Langberg, M., Katti, S., Ho, T., Katabi, D., & Médard, M. (2007). Resilient network coding in the presence of Byzantine adversaries. In *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM)*, Barcelona, Spain.
38. Ho, T., Leong, B., Koetter, R., Médard, M., Effros, M., & Karger, D. R. (2008). Byzantine modification detection in multicast networks with random network coding. *IEEE Transactions on Information Theory, 54*(6), 2798–2803.
39. Kim, M., Lima, F., Zhao, F., Barros, J. M. M., Koetter, R., Kalker, T., & Han, K. J. (2010). On counteracting Byzantine attacks in network coded peer-to-peer networks. *IEEE Journal on Selected Areas in Communications, 28*(5), 692–702.
40. Kim, M., Médard, M., & Barros, J. (2011). Algebraic Watchdog: Mitigating misbehavior in wireless network coding. *IEEE Journal on Selected Areas in Communications, 29*(10), 1916–1925.
41. Zhao, F., Kalker, T., Médard, M., & Han, K. J. (2007). Signatures for content distribution with network coding. In *Proceedings of the 2007 IEEE International Symposium on Information Theory (ISIT)*, Nice, France.
42. Li, Y., Yao, H., Chen, M., Jaggi, S., & Rosen, A. (2010). RIPPLE authentication for network coding. In *Proceedings of the 29th IEEE International Conference on Information Communications (INFOCOM)*, San Diego, CA, USA.
43. Le, A., & Markopoulou, A. (2012). Cooperative defense against pollution attacks in network coding using SpaceMac. *IEEE Journal on Selected Areas in Communications, 30*(2), 442–449.
44. Siavoshani, M. J., Fragouli, C., & Diggavi, S. (2008). On locating Byzantine attackers. In *Proceedings of the 4th Workshop on Network Coding, Theory and Applications (NetCod)*, Hong Kong, China.
45. Wang, Q., Vu, L., Nahrstedt, K. & Khurana, H., (2009). *Identifying Malicious nodes in network-coding-based peer-to-peer streaming networks*. Urbana/Champaign: University of Illinois.
46. Parsamehr, R., Esfahani, A., Mantas, G., Rodriguez, J., & Martínez-Ortega, J.-F. (2019). A location-aware IDPS scheme for network coding-enabled mobile small cells. In *Proceedings of the 2nd IEEE 5G World Forum (5GWF)*, Dresden, Germany.
47. Zhang, P., Jiang, Y., Lin, C., Yao, H., Wasef, A., & Shenz, X. (2011). Padding for orthogonality: Efficient subspace authentication for network coding. In *Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM)*, Shanghai, China.
48. Esfahani, A., Mantas, G., Rodriguez, J., & Neves, J. C. (2017). An efficient homomorphic MAC-based scheme against data and tag pollution attacks in network coding-enabled wireless networks. *International Journal of Information Security, 16*(6), 627–639.
49. Esfahani, A., Yang, D., Mantas, G., Nascimento, A., & Rodriguez, J. (2015). Dual-homomorphic message authentication code scheme for network coding-enabled wireless sensor networks. *International Journal of Distributed Sensor Networks, 2015*, 1–11.
50. Parsamehr, R., Mantas, G., Rodriguez, J., & Martínez-Ortega, J.-F. (2020). IDLP: An efficient intrusion detection and location-aware prevention mechanism for network coding-enabled mobile small cells. *IEEE Access, 8*, 43863–43875.
51. Scarfone, K., & Mell, P. (2012). *Guide to Intrusion Detection and Prevention Systems (IDPS)* (SP 800-94 revision 1). Gaithersburg: National Institute of Standards and Technology (NIST).
52. Kent, K., Chevalier, S., Grance, T., & Dang, H. (2006). *Guide to integrating forensic techniques into incident response* (SP 800-86). Gaithersburg: National Institute of Standards and Technology (NIST).
53. Kent, K., & Souppaya, M. (2006). *Guide to computer security log managemen* (SP 800-92). Gaithersburg: National Institute of Standards and Technology (NIST).
54. Cichonski, P., Millar, T., Grance, T., & Scarfone, K. (2012). *Computer security incident handling guide* (SP 800-61 revision 2). Gaithersburg: National Institute of Standards and Technology.

55. Parsamehr, R., Esfahani, A., Mantas, G., Radwan, A., Mumtaz, S., Rodriguez, J., & Martínez-Ortega, J.-F. (2019). A novel intrusion detection and prevention scheme for network coding-enabled mobile small cells. *IEEE Transactions on Computational Social Systems, 6*(6), 1467–1477.
56. Pedersen, M. V., Heide, J., & Fitzek, F. H. P. (2011). Kodo: An open and research oriented network coding library. In *Proceedings of the 6th International Conference on Research in Networking (NETWORKING)*, Valencia, Spain.
57. Dong, J., Curtmola, R., & Nita-Rotaru, C. (2011). Practical defenses against pollution attacks in wireless network coding. *ACM Transactions on Information and System Security, 14*(1), 1–31.
58. Vasudevan, V. A., Tselios, C., & Politis, I. (2020). On security against pollution attacks in network coding enabled 5G networks. *IEEE Access, 8*, 38416–38437.
59. Crosby, M., Pattanayak, P., Verma, S., & Kalyanaraman, V. (2016). Blockchain technology: Beyond bitcoin. *Applied Innovation Review, 2*, 6–19.
60. Zheng, Z., Xie, S., Dai, H. N., Chen, X., & Wang, H. (2018). Blockchain challenges and opportunities: A survey. *International Journal of Web and Grid Services, 14*(4), 352–375.
61. Wood, G. (2014). *Ethereum: A secure decentralised generalised transaction ledger*. EIP-150 Rev (a04ea02–2017-09-30). Ethereum.
62. McConaghy, T., Marques, R., Müller, A., de Jonghe, D., McConaghy, T., McMullen, G., Henderson, R., Bellemare, S., & Granzotto, A. (2016). *BigchainDB: A scalable Blockchain database*. Berlin: ascribe GmbH.
63. Shen, X. (2015). Device-to-device communication in 5G cellular networks. *IEEE Network, 29*(2), 2–3.
64. Adat, V., Politis, I., Tselios, C., Galiotos, P., & Kotsopoulos, S. (2018). On Blockchain enhanced secure network coding for 5G deployments. In *Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, UAE.
65. Adat, V., Politis, I., Tselios, C., & Kotsopoulos, S. (2018). Secure network coding for SDN-based mobile small cells. In *Proceedings of the 9th International Conference on Broadband Communications, Networks and Systems (BROADNETS)*, Faro, Portugal.
66. Adat, V., Politis, I., & Kotsopoulos, S. (2019). On Blockchain based secure network coding for mobile small cells. In *Proceedings of the 2nd IEEE 5G World Forum (5GWF)*, Dresden, Germany.
67. Canetti, R., Garay, J., Itkis, G., Micciancio, D., Naor, M., & Pinkas, B. (1999). Multicast security: A taxonomy and some efficient constructions. In *Proceedings of the 18th IEEE International Conference on Computer Communications (INFOCOM)*, New York, NY, USA.
68. Postman API. *The collaboration platform for API development* [Online]. Available: https://www.postman.com/. Accessed 15 Aug 2020.

Chapter 6
Channel Estimation in RIS-Aided Networks

Fadil Danufane, Placido Mursia, and Jiang Liu

Abstract Reconfigurable intelligent surface (RIS) is a recently emerging transmission technology for application to wireless communications. Regarded to be an emerging solution for the next generation of communications, RIS is a nearly passive device that realizes smart radio environment with low hardware cost and energy consumption. This merit of RIS, on the other hand, imposes a major challenge to the channel estimation of RIS-aided communication systems. Recently, many protocols and algorithms are proposed to handle this challenging problem. In this chapter, we review the problem of channel estimation in RIS-aided systems and survey recent developments on this topic.

6.1 Introduction

The need for high data rates is ever increasing in the future. By 2030, it is forecast that the global data traffic will increase up to the order of thousands of exabytes [1]. In addition, future wireless communication systems such as 6G are expected to deliver these data in a distributed and intelligent way, as well as within some delay and reliability constraints that are more stringent than ever. These requirements cannot be satisfied by the existing technologies and even the newly deployed 5G communication system.

To meet these challenging demands and requirements, a new paradigm on how a communication system is designed is needed. Recently, a vision of smart radio environment (SRE) was proposed to challenge the status quo of communication system design and redefine the performance limit of communication systems. In

F. Danufane · J. Liu (✉)
Université Paris-Saclay, CentraleSupelec, CNRS, Laboratoire des Signaux et Systèmes,
Gif-sur-Yvette, France
e-mail: fadil.danufane@centralesupelec.fr; jiang.liu@centralesupelec.fr

P. Mursia
EURECOM, Sophia-Antipolis, France
e-mail: placido.mursia@eurecom.fr

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_6

particular, in SRE, the environment is no longer seen as an impairment according to which a system has to be designed, but instead as a component that can be controlled to achieve a specific performance.

RIS is a recently emerging transmission technology for application to wireless communications. Conceptually speaking, RIS is a two-dimensional surface made of metamaterials that is capable of manipulating the incident electromagnetic waves in arbitrary ways. The main selling points of RIS are its near-passive nature, since it does not require a large power source to redirect the waves, and its low cost and low complexity of large-scale deployments. Thanks to these properties, RISs are receiving major attention from the wireless community and are considered to be the key technology to realize the vision of SRE.

A RIS consists of many sub-wavelength unit elements, usually called meta-atoms, whose phase shift can be configured independently. By configuring the phase shift of each unit cell, one can manipulate the reflected wave in many ways, e.g., by manipulating the wave by reflecting an incoming beam in any desired direction or by focusing the reflected wave to maximize the electric intensity at a specific location. Therefore, the main property of the RIS is its capability of being reconfigurable even after its deployment in a wireless environment.

Due to the sub-wavelength structure of the RIS, the distance between adjacent unit cells and the size of each unit cell is much smaller than the wavelength. Therefore, the propagation or resonance effects in the direction perpendicular to the surface can be safely ignored in the process of synthesis and analysis of the surface. Thanks to this, a RIS can be modelled through appropriate continuous surface-averaged functions (e.g., susceptibilities), despite being made of discrete elements. This representation of the RIS as a continuous entity allows for convenient performance analysis through some concepts of physics, as demonstrated in [2].

Recently, there have been exciting research activities on the realization of low-cost and practical RIS. Two recent examples of these activities are illustrated in Figs. 6.1 and 6.2. In Fig. 6.1, the RFocus prototype, recently designed by researchers of the Massachusetts Institute of Technology (MIT), USA, is depicted [3]. The prototype is made of 3720 inexpensive antennas arranged on a 6-square meter surface. At scale, each antenna element is expected to have a cost of the order of a few cents or less. In Fig. 6.2, a prototype of smart glass, recently designed by researchers from NTT DOCOMO, Japan, is depicted [4]. The manufactured smart glass is an artificially engineered thin layer (i.e., a metasurface) that comprises a large number of sub-wavelength unit elements placed in a periodic arrangement on a two-dimensional surface covered with a glass substrate. By moving the glass substrate slightly, it is possible to dynamically control the response of the impinging radio waves in three modes: (i) full penetration of the incident radio waves, (ii) partial reflection of the incident radio waves, and (iii) full reflection of all radio waves.

Although RIS is generally capable to modify the impinging electromagnetic wave in any desired way, recently, there are two widely investigated functionalities within the literature of wireless communication technology.

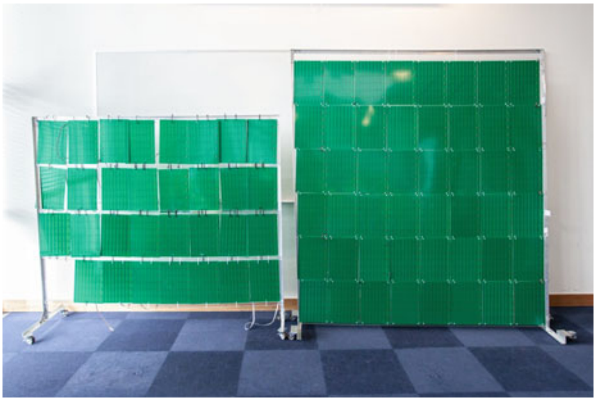


Fig. 6.1 MIT’s RFocus prototype. (Photo: Jason Dorfman, CSAIL)

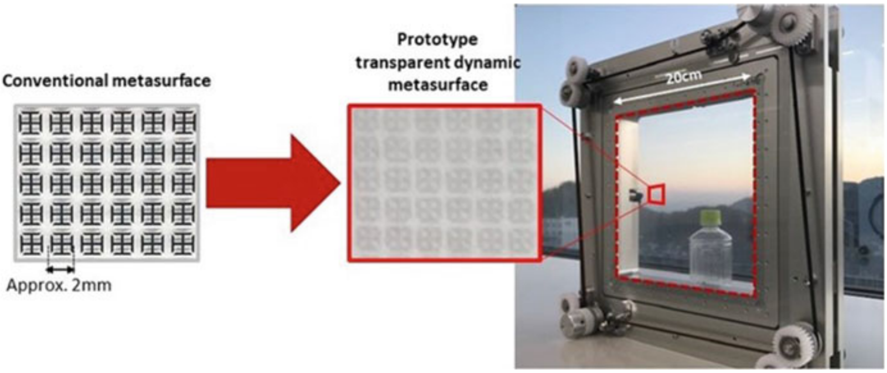


Fig. 6.2 NTT DOCOMO’s prototype. (Photo: NTT DOCOMO)

1. ***Anomalous reflection/transmission*** [5]. Under this setting, the RIS is configued to reflect or refract the impinging radios waves toward specified directions that do not necessarily adhere to the laws of reflection and refraction (i.e., the angle of incident is not necessarily equal to the angle of reflection/transmission). This setting is useful in some setups in which several users are being served simultaneously (e.g., broadcasting application) or when a single user is moving in a constant direction with respect to the RIS (e.g., vehicular application). The limitation of this setting lies on the fact that, in general, the signal-to-noise-ratio is not maximized and thus the system capacity is not achieved.
2. ***Beamforming/focusing*** [3]. Under this setting, the RIS is configured to the reflected/transmitted electromagnetic wave into a specific location such that the intensity is maximized there. Therefore, in this case, the signal-to-noise-ratio is maximized, and thus the system capacity is achieved for a single user in the designated location. The limitation of this setting lies on the potential complexity of the phase-shift reconfiguration of each unit cell of the RIS to accommodate the mobility of the user.

From an application point of view, RIS can be utilized for various use cases. Some examples include but are not limited to the following [6]:

- ***Signal engineering.*** The RIS provides an additional LOS path between a transmitter and a receiver to mitigate the non-existence of direct link between them.
- ***Interference engineering.*** The RIS is configured to minimize the signal that comes from an interfering transmitter at the intended receiver.
- ***Security engineering.*** The RIS is configured to minimize the signal containing information between a transmitter and a receiver that arrives at a malicious user.
- ***Scattering engineering.*** The RIS is configured to increase the channel rank between a transmitter and a receiver by means of creating a rich scattering environment (high rank channel) for high data rate transmission.

We end this section by mentioning that despite the study on RIS in the literature, they mostly consider a flat RIS such as internal walls of indoor environments, external facades of buildings, and the glasses of windows; in general, a RIS does not have to be planar. Some applications in wireless communications, for example, include coating several irregularly shaped objects in order to control the reflected/refracted radio waves that impinge it to enhance the overall communication performance. These functions cannot, in general, be realized by using a planar RIS.

6.2 The Channel Estimation Problem in RIS-Aided Networks

Channel estimation in a RIS-assisted wireless system is a much more challenging task than in conventional systems since the passive RIS elements are incapable of sensing and estimating channel information. Such design choice is undoubtedly more appealing due to its extremely low hardware and deployment cost. However, accurate channel state information (CSI) is critical in optimizing the RIS parameters.

Thus, the problem of estimating the channel in RIS-aided networks has gained much attention lately. In particular, the focus is on how to estimate the two cascaded channels between the transmitter and the RIS and between the RIS and the UE with purely passive reflecting elements and an affordable training overhead.

Consider a general multi-user multi-input single-output (MISO) network setup detailed in Fig. 6.3, where a base station (BS) equipped with M antennas communicates with K single-antenna user ends (UEs) with the aid of a RIS made of N reflecting elements. We assume that the transmission takes place over a total of T time slots in which the channel is assumed to be constant, following a quasi-static fading model. The channel between the BS and the RIS is denoted as $\boldsymbol{G} \in \mathbb{C}^{N \times M}$, while $\boldsymbol{h}_k \in \mathbb{C}^{N \times 1}$ denotes the channel between the RIS and UE k. Lastly, $\boldsymbol{h}_{d,k} \in \mathbb{C}^{M \times 1}$ represents the direct channel between the BS and UE k. Hence, the signal received by the k-th UE in the downlink at time t is given by

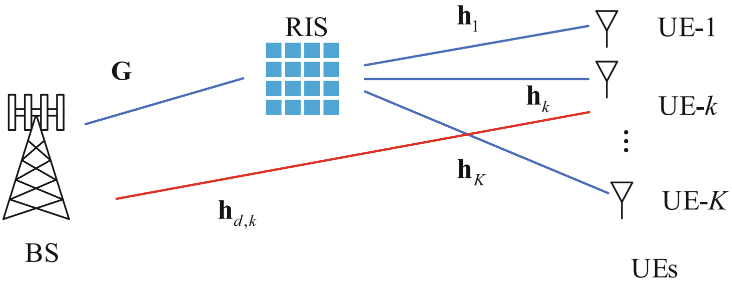


Fig. 6.3 A model of a RIS-assisted multiuser MISO system

$$y_{k,t} = \left(\boldsymbol{h}_{d,k}^H + \boldsymbol{h}_k^H \boldsymbol{\Phi}_t \boldsymbol{G}\right) \mathbf{x}_t + n \tag{6.1}$$

where $\boldsymbol{\Phi}_t = \text{diag}\left(\beta_{1,t} e^{j\phi_{1,t}}, \ldots, \beta_{N,t} e^{j\phi_{N,t}}\right) \in \mathbb{C}^{N\times N}$ is the matrix containing each RIS element absorption coefficient $\beta_{n,\, t} \in [0, 1]$ and shift $\phi_{n,\, t} \in [0, 2\pi]$ at time t, diag($\boldsymbol{x}$) represents a diagonal matrix with the entries of $\boldsymbol{x}$ on its main diagonal, $\mathbf{x}_t \in \mathbb{C}^{M\,\times\,1}$ is the signal transmitted by the BS at time t with $\text{E}[\|\mathbf{x}_t\|^2] = 1$, and$n \sim \mathcal{CN}\left(0, \sigma_n^2\right)$ is a noise coefficient. Let $\boldsymbol{y}_k = [y_{k,\,1}, \ldots, y_{k,\,T}]^T$ be the signal collected at UE k after T pilot symbols. Similarly, the receive signal at the BS at time t is expressed as

$$\boldsymbol{y}_t = \sum_{k=1}^{K} \left(\boldsymbol{h}_{d,k} + \boldsymbol{G}^H \boldsymbol{\Phi}_t \boldsymbol{h}_k\right) \mathrm{x}_{t,k} + \boldsymbol{n} \tag{6.2}$$

where $\mathrm{x}_{t,\,k}$ is the signal transmitted by UE k at time t with $\text{E}[|\mathrm{x}_{t,\,k}|^2] = 1$. Lastly, let $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$ be the receive signal at the BS after T training symbols. Equations (6.1) and (6.2) can be rewritten as

$$y_{k,t} = \left(\boldsymbol{h}_{d,k}{}^H + \boldsymbol{v}_t{}^H \overline{\boldsymbol{H}}_k\right) \mathbf{x}_t + n \tag{6.3}$$

$$y_{k,t} = \left(\boldsymbol{h}_{d,k}{}^H + \boldsymbol{v}_t{}^H \overline{\boldsymbol{H}}_k\right) \mathbf{x}_t + n \tag{6.4}$$

where $\boldsymbol{v}_t = \left[\beta_{1,t} e^{-j\phi_{1,t}}, \ldots, \beta_{N,t} e^{-j\phi_{N,t}}\right]$ contains the RIS configuration at time t and $\overline{\boldsymbol{H}}_k = \text{diag}\left(\boldsymbol{h}_k{}^H\right) \boldsymbol{G}$ represents the aggregated effective channel between UE k and the BS via the RIS.

6.3 Survey on Channel Estimation

In this section, we review the main existing solutions based on analytical optimization of the channel estimation protocol. In this respect, we identify several main categories depending on the fundamental idea behind each channel estimation procedure. For each category, we point out the main characteristics and drawbacks.

6.3.1 On/Off-Based Channel Estimation

In this section, we review a class of channel estimation protocols based on sequentially activating only one RIS element for each pilot symbol. The full channel is thus estimated in $N + 1$ training symbols where the first pilot symbol is necessary to estimate the direct channel between the BS and the UEs.

Works such as [7–9] are based on activating only one RIS element for each pilot symbol. In all such works, only the aggregated channels $\left\{\overline{\boldsymbol{H}}_k\right\}$ and $\{\boldsymbol{h}_{d,\,k}\}$ are estimated. Hence, for each UE k, the resulting aggregate channel $\overline{\boldsymbol{H}}_k$ is estimated column-wise in a total of N training symbols. An extra training symbol is necessary to estimate $\boldsymbol{h}_{d,\,k}$ with all the RIS elements deactivated. All UEs transmit such pilots concurrently, and interference among them is resolved thanks to the use of orthogonal training sequences. Indeed, we have that $\mathbf{x}_k{}^H\mathbf{x}_j = 0$ if $k \neq j$ and $\mathbf{x}_k{}^H\mathbf{x}_k = 1$.

At time $t = 1$, the received signal at the BS is given by

$$\boldsymbol{y}_1 = \sum_{k=1}^{K} \boldsymbol{h}_{d,k}\mathrm{x}_{1,k} + \boldsymbol{n} \tag{6.5}$$

while the receive signal at a generic time instant t is given by

$$\boldsymbol{y}_t = \sum_{k=1}^{K} \left(\boldsymbol{h}_{d,k} + \overline{\boldsymbol{h}}_{k,t}{}^{H} v_t\right) \mathrm{x}_{t,k} + \boldsymbol{n} \tag{6.6}$$

In [9] the channel of each UE k is estimated using least squares, i.e., the receive signal $\boldsymbol{Y}$ is multiplied by $\overline{\mathbf{x}}_k^* = \left[\mathrm{x}_{1,k}{}^*, \mathrm{x}_{2,k}{}^* v_2{}^*, \ldots, \mathrm{x}_{T,k}{}^* v_T{}^*\right]^T$. The first column of $\boldsymbol{Y}$ is used to estimate $\boldsymbol{h}_{d,\,k}$. The resulting estimate is subtracted from the signal $\boldsymbol{r} = \boldsymbol{Y}\overline{\mathbf{x}}_k^*$ in order to obtain the estimate of $\overline{\boldsymbol{h}}_k$. In [7, 8], such estimate is further refined using the minimum-mean-squared-error (MMSE) principle, i.e., by exploiting the known statistics of the channel and noise.

In practice, to implement the ON/OFF switching of the massive RIS elements is costly. Besides, as only a small portion of its elements is switched ON at each time, the channel estimation accuracy is degraded. To address this issue, [10] proposed an RIS elements-grouping method to reduce the training overhead and estimation complexity. Instead of controlling the ON/OFF states of a single element each time, the authors applied the ON/OFF method on the grouped RIS elements.

Similarly, in [11], after the superimposed channel is obtained using the least square (LS) estimation, the grouping ON/OFF method is adopted to estimate the direct channel link and the cascaded channel link.

In [12, 13], the idea of grouping ON/OFF method is extended. With the same assumption that the RIS can be divided into multiple sub-surfaces of adjacent strongly correlated reflecting elements that apply the same reflection coefficient, [12] designed the reflection pattern based on discrete Fourier transform (DFT) or Hadamard matrix based on their orthogonality, while the authors in [13] designed the pattern based on the minimum variance unbiased estimation principle, which mimics a series of discrete Fourier transforms.

In [14], the authors propose a three-phase channel estimation protocol based on the observation that each RIS element reflects the signals from all the users to the transmitter via the same channel. The first phase is similar to Eq. (6.5); all the IRS elements are switched off to estimate the direct channel. In the second phase, all the IRS reflection elements are switched on, and merely one typical user transmits nonzero pilot symbols to the BS. In this phase, the BS estimates the cascaded channel of this typical user. The construction of the reflection coefficient matrix can be based on the DFT matrix. In the last phase, the cascaded channels of other users are estimated, where the channel correlations are exploited to reduce complexity. The authors quantified the minimum time to estimate all required channels and show that massive multi-input multi-output (MIMO) may play an important role in reducing the channel estimation overhead in RIS-based communication systems.

6.3.2 Least Squares-Based Channel Estimation

In [15], the authors propose an iterative algorithm for channel estimation that is based on the parallel factor decomposition algorithm. The proposed method is based on an alternating least squares algorithm that iteratively estimates the channel between the transmitter and the RIS $\mathbf{G}$ as well as the channel between the RIS and the users $\mathbf{h}_k$. Considering the low resolution of the RIS unit elements, the RIS is assumed to have P different phase configuration. Define the $P \times N$ complex-valued matrix $\boldsymbol{\Theta}$ as the configuration matrix; the p-th row of $\boldsymbol{\Theta}$ represents the p-th RIS phase configuration. Consequently, the end-to-end RIS-based wireless channel can be given by

$$\mathbf{Z}_p = \mathbf{H}_2 \operatorname{diag}\left(\boldsymbol{\Theta}\left(\boldsymbol{p}, :\right)\right) \mathbf{G} \tag{6.7}$$

where $\mathbf{H}_2 = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_K]^T \in \mathbb{C}^{K \times N}$ is the channel between the RIS and the K users. Each (k, m)-th entry of $\mathbf{Z}_p$ with $k = 1, 2, \ldots, K$ and $m = 1, 2, \ldots, M$ is obtained as

$$\left[\mathbf{Z}_p\right]_{k,m} = \sum_{n=1}^{N} [\mathbf{H}_2]_{k,n}\ [\mathbf{G}]_{n,m}[\boldsymbol{\Theta}]_{p,n} \tag{6.8}$$

where $[\mathbf{G}]_{n,\,m}$, $[\mathbf{H}_2]_{k,\,n}$, and $[\boldsymbol{\Theta}]_{p,\,n}$ denote the (n, m)-th entry of $\mathbf{G}$, (k, n)-th entry of $\mathbf{H}_2$, and (p, n)-th entry of $\boldsymbol{\Theta}$, respectively, with $n = 1, 2, \ldots, N$.

The proposed method is based on an alternating least squares algorithm that iteratively estimates the channel between the transmitter and the RIS $\mathbf{G}$ as well as the channel between the RIS and the users $\mathbf{H}_2$. Using the PARAllel FACtor (PARAFAC) decomposition, $\mathbf{Z}_p$ can be represented using three matrix forms. These matrices form the horizontal, lateral, and frontal slices of the tensor composed of Eq. (6.8). The unfolded forms of the mode-1, mode-2, and mode-3 of $\mathbf{Z}_p$'s are expressed as follows:

$$\begin{aligned} \textbf{Mode} - \mathbf{1} : \mathbf{Z}_\alpha &= \left(\mathbf{G}^{T\circ}\boldsymbol{\Theta}\right)\mathbf{H}_2^T \in \mathbb{C}^{PM\times K} \\ \textbf{Mode} - \mathbf{2} : \mathbf{Z}_\beta &= \left(\boldsymbol{\Theta}^{\circ}\mathbf{H_2}\right)\mathbf{G} \in \mathbb{C}^{KP\times M} \\ \textbf{Mode} - \mathbf{3} : \mathbf{Z}_\gamma &= \left(\mathbf{H_2}^{\circ}\mathbf{G}^T\right)\boldsymbol{\Theta}^T \in \mathbb{C}^{MK\times P} \end{aligned} \tag{6.9}$$

where $^{\circ}$ represents the Khatri-Rao (column wise Kronecker) matrix product. Considering AWGN, we define the following three-dimensional matrix:

$$\overset{\sim}{\mathbf{Z}} = \mathbf{Z} + \overset{\sim}{\mathbf{W}} \tag{6.10}$$

where tensor $\overset{\sim}{\mathbf{W}} \in \mathbb{C}^{K\times M\times P}$ is the AWGN that incorporates all P matrices $\overset{\sim}{\mathbf{W}}_{\boldsymbol{p}}$.

The proposed iterative channel estimation is expressed as follows:

1. First step (Initialization): Initialize with a random feasible phase matrix $\boldsymbol{\Theta}$. $\hat{\mathbf{G}}^{(\mathbf{0})}$ represents the eigenvector matrix corresponding to the N nonzero eigenvalues of $\overset{\sim}{\mathbf{Z}}{}_{\boldsymbol{\beta}}^{H}\overset{\sim}{\mathbf{Z}}_\beta$, where $\overset{\sim}{\mathbf{Z}}_\beta$ is the noisy version of Mode-1 form of Eq. (6.10). Similarly, $\hat{\mathbf{H}}_2^{(0)}$ is the eigenvector matrix corresponding to the N nonzero eigenvalues of $\overset{\sim}{\mathbf{Z}}{}_{\boldsymbol{\alpha}}^{H}\overset{\sim}{\mathbf{Z}}_\alpha$, where $\overset{\sim}{\mathbf{Z}}_\alpha$ is the noisy version of Mode-2 form of Eq. (6.10). Set the algorithmic iteration $i = 1$.
2. Second and third steps (Iterative Update):

$$\hat{\mathbf{H}}_2^{(i)} = \left(\left(\hat{\mathbf{A}}_1^{(i-1)}\right)^{+}\overset{\sim}{\mathbf{Z}}{}'\right)^T$$

$$\hat{\mathbf{A}}_1^{(i-1)} = \hat{\mathbf{H}}_2^{(i-1)\circ}\boldsymbol{\Theta}.$$

$$\hat{\mathbf{G}}^{(i)} = \left(\hat{\mathbf{A}}_2^{(i)}\right)^{+} \overset{\sim\prime\prime}{\mathbf{Z}}$$

$$\hat{\mathbf{A}}_2^{(i)} = \boldsymbol{\Theta}^{\circ} \hat{\mathbf{H}}_2^{(i)}$$

where $\overset{\sim\prime}{\mathbf{Z}} \in \mathbb{C}^{PM \times K}$ is a matrix-stacked form of Eq. (6.10)'s tensor $\overset{\sim}{\mathbf{Z}}$, $\overset{\sim\prime\prime}{\mathbf{Z}} \in \mathbb{C}^{KP \times M}$ is another matrix-stacked form of $\overset{\sim}{\mathbf{Z}}$, and $(\bullet)^{+}$ denotes the pseudo-inverse matrix.

3. Fourth step (Iteration Stop Criterion): The proposed iterative algorithm terminates when either the maximum number $I_{\max}$ of algorithmic iterations is reached or when between any two algorithmic iterations $i - 1$ and i hold the following condition for ε being a very small positive real number:

$$\begin{aligned}&\left\|\hat{\mathbf{G}}^{(i)} - \hat{\mathbf{G}}^{(i-1)}\right\|_F^2 / \left\|\hat{\mathbf{G}}^{(i)}\right\|_F^2 \leq \varepsilon \\ &\text{or} \\ &\frac{\left\|\hat{\mathbf{H}}_2^{(i)} - \hat{\mathbf{H}}_2^{(i-1)}\right\|_F^2}{\left\|\hat{\mathbf{H}}_2^{(i)}\right\|_F^2} \leq \varepsilon\end{aligned} \tag{6.11}$$

Thus, the channels $\mathbf{G}$ and $\mathbf{H}_2$ are obtained using this alternate LS iteration.

6.3.3 Sparsity-Based Channel Estimation

This section deals with a class of channel estimation methods that rely on the assumption of channel sparsity. Indeed, often, the BS and the RIS are mounted on top of buildings and are in LoS with each other such that the channel $\mathbf{G}$ can be regarded as being close to rank-one, i.e., dominated by the LoS path. A similar consideration holds for each channel $\mathbf{h}_k$ especially if the latter is a mmWave or TeraHertz channel. However, even in this case, the multipath component typically carries a lower but still noticeable amount of power compared to the LoS path. Leveraging on the sparsity of $\mathbf{G}$ and the aggregated channels $\overline{\mathbf{H}}_k$, several recent works have proposed to use compressed sensing (CS) [16], beam training (BT) [17, 18], sparse matrix factorization (SMF) [19], matrix calibration [20], or orthogonal matching pursuit (OMP) [21, 22] in order to estimate the channels and reduce the training overhead compared to on/off techniques, as described in the previous section.

6.3.3.1 Compressed Sensing

The work in [16] proposes to exploit the inherent sparsity of the effective channels $\overline{\mathbf{H}}_k$ which is due to the low-scattering link connecting the BS and the RIS via CS. The training period is divided into BT symbols. For each one of the B blocks, the UEs send mutually orthogonal sequences of length T with $T \geq K$. Each UE repeats the same pilot sequence for all the B blocks. The RIS is configured following a series of mutually orthogonal sequences which are repeated for the T symbols of each block. As a result, the matrix $\mathbf{V}$ is unitary across the different blocks. Hence, this algorithm is designed to obtain diversity in the received signal across both pilot sequences and RIS configurations.

Assuming that the direct link between the BS and the UEs can be neglected due to low associated power, at each block b, the receive signal at the BS is defined as

$$\mathbf{Y}_b = \sum_{k=1}^{K} \overline{\mathbf{H}}_k{}^H \mathbf{v}_b \mathbf{x}_k{}^H + \mathbf{n}_b \in \mathbb{C}^{M \times T} \tag{6.12}$$

As a first estimate of the effective channels, the authors propose to use the least squares signal, i.e., $\mathbf{r}_{b,\,k} = \mathbf{Y}_b \mathbf{x}_k$, for each block b and UE k. Such initial estimate is then further refined by exploiting its sparsity structure. In particular, the effective channels are modelled using a virtual channel representation as

$$\overline{\mathbf{H}}_k = \mathbf{A}_R \mathbf{X}_k \mathbf{A}_B^H \tag{6.13}$$

where $\mathbf{A}_R \in \mathbb{C}^{N \times N'}$ with $N' > N$ is an over-complete array response at the RIS, $\mathbf{X}_k \in \mathbb{C}^{M' \times N'}$ is the channel coefficient matrix of UE k assumed to be sparse in which each element represents the channel gain along the associated path, and $\mathbf{A}_B \in \mathbb{C}^{M \times M'}$ with $M' > M$ is an over-complete array response at the BS. Hence, the problem of channel estimation is reduced to estimating $\mathbf{X}_k$ from the least squares signal $\mathbf{r}_{b,\,k}$ via CS. However, the authors note that the application of the standard OMP algorithm directly to the least squares signal brings substantially two disadvantages: (1) the OMP algorithm requires an accurate sampling of the angular domain to obtain good results, i.e., very large N' and M' which lead to complex matrix operations, and (2) this estimator requires an increasing training overhead in terms of pilot sequences as the channel sparsity increases. Hence, the authors propose to apply OMP on $\mathbf{X}_k \mathbf{A}_B^H$ by exploiting its row block sparsity structure. Note that this significantly reduces computational complexity since typically $M \leq N'$. Moreover, since the link connecting the BS and the RIS is common to all UEs, the aggregated effective channel of all the UEs exhibits both row and column block sparsity which can be leveraged to further enhance the performance of OMP and reduce the training overhead. Note that the column block sparsity is given by the shared $\mathbf{G}$ channel among all UEs.

6.3.3.2 Beam Training

The authors in [17] study an indoor RIS-assisted network with a massive MIMO BS serving a single receiver equipped with N_u antennas with the aid of a total of N_i RISs operating at THz frequency in the absence of the LoS path. In this case, the sparsity in the channel is given by the large-scale antenna array at the BS and the high pathloss at THz frequencies. The effective channel is thus modelled as

$$\overline{\mathbf{H}} = \sum_{i=1}^{N_i} \overline{\mathbf{H}}_{\boldsymbol{i}} \tag{6.14}$$

where $\overline{\mathbf{H}}_{\boldsymbol{i}}$ is the effective channel that is reflected by the RIS i via the reflecting coefficients in $\mathbf{v}_i$. Assuming for simplicity, a uniform linear array (ULA) at both the BS and the receiver, the effective channel relative to RIS i, is described as

$$\overline{\mathbf{H}}_{\boldsymbol{i}} = \eta_i \mathbf{a}_{N_u}\left(\theta_{UR}^i\right) \mathbf{a}_N\left(\theta_{RU}^i\right)^H \mathbf{\Phi}_i \mathbf{a}_N\left(\theta_{RB}^i\right) \mathbf{a}_M\left(\theta_{BR}^i\right)^H \tag{6.15}$$

where η_i is the overall path-loss coefficient which depends on the distance from the receiver to the RIS and from the RIS to the BS and $\mathbf{a}_{N_u}\left(\theta_{UR}^i\right)$ is the ULA response vector for the steering angle θ_{UR}^i defined as

$$\mathbf{a}_{N_u}\left(\theta_{UR}^i\right) = \frac{1}{\sqrt{N_u}}\left[1, e^{j2\pi\delta\sin\left(\theta_{UR}^i\right)}, \ldots, e^{2\pi\delta(N_u-1)\sin\left(\theta_{UR}^i\right)}\right]^T \tag{6.16}$$

with δ being the ratio between the antenna spacing and the signal wavelength. Lastly, note that θ_{UR}^i is the angle of departure (AoD) from the receiver to the i-th RIS, θ_{RU}^i is the angle of arrival (AoA) of the same link, θ_{RB}^i is the AoD from the i-th RIS to the BS, and θ_{BR}^i is the AoD of the same link. The effective channel is thus estimated via beam training, in which the BS, RIS configuration, and receiver all sweep through a codebook of beam directions, keeping as candidate estimate the direction which gives the strongest received beam power. This is done via a hierarchical search method which greatly reduces the complexity compared to brute-force exhaustive search. In a first stage, only the BS to RIS link is considered. Once the best candidate direction is found, the algorithm considers the RIS to receiver link with the BS to RIS link fixed as the result of the first stage. Note that the direct link between the BS and the UE is estimated in a prior phase via hierarchical beam search with all the RIS elements deactivated.

A similar case assuming a single RIS and both BS and receiver equipped with one antenna only has been studied in [18]. Here, the indoor network is assumed to operate at mmWave frequencies, and both the BS-RIS and RIS-UE links are assumed to be dominated by the LoS only. As in [17], the channel is modelled as depending only on distances AoA and AoD of the two separate links. Hence, in this case, the channel is completely identified by the position of the UE in space.

In order to reduce complexity, the authors propose to divide the RIS into a series of rectangular blocks of reflecting units (RUS). Each RUS is considered as an observation point that is used to estimate the position of the UE via triangulation.

Each RUS is used to sweep through a set of directions in which it is most likely to find the UE. For each RUS, the direction of maximum received power is used as an estimate of the UE position, while the corresponding RUS-UE distance is estimated using classical wideband delay estimation methods. Such estimates are then combined into one refined estimate via triangulation.

6.3.3.3 Sparse Matrix Factorization

The authors in [19] study the joint activity detection and channel estimation problem in a scenario in which a large number of UEs are present in the network, but only a small percentage of them are active in any given time instant. Hence, besides estimating the channels, the goal of this paper is to detect which UEs are actually active during each channel coherence block. In this case, the sparsity is given by the matrix $\mathbf{A} = \ diag\ [\alpha_1, \ldots, \alpha_K]$ which indicates whether each UE k is active or not, i.e., $\alpha_k \in \{0, 1\}$, where $\alpha_k = 1$ indicates that UE k is active and $\alpha_k = 0$ indicates that UE k is inactive and by a sparse design of the sequences $\mathbf{v}_t$. Indeed, at each time instant t, each RIS element is activated according to a Bernoulli distribution and with uniformly distributed phases shift. Assuming that the direct links between the BS and the UEs are blocked and do not carry a significant amount of power, the receive signal at the BS can be rewritten as

$$\mathbf{Y} = \mathbf{G}\left(\mathbf{V} \odot (\mathbf{HAX})\right) + \mathbf{N} \tag{6.17}$$

where $\mathbf{V}$ contains all the T training sequences as columns and $\mathbf{H}$ contains the channels from each UE to the RIS as columns. Equation (6.17) can be further simplified as

$$\mathbf{Y} = \mathbf{GW} + \mathbf{N} \tag{6.18}$$

where we have defined the matrices $\boldsymbol{\Theta} = \mathbf{HA}$, $\mathbf{Q} = \boldsymbol{\Theta}\mathbf{X}$ and $\mathbf{W} = \mathbf{P} \odot \mathbf{Q}$ which are all sparse and can be recovered via the following techniques: SMF is employed to estimate $\mathbf{G}$ and $\mathbf{W}$ from the observations in $\mathbf{Y}$, and matrix completion is used to complete the missing entries of $\mathbf{Q}$ given the estimate of $\mathbf{W}$ and the training sequences in $\mathbf{V}$. Lastly, multiple measurement vectors are used to estimate $\boldsymbol{\Theta}$ from the estimate of $\mathbf{Q}$ and the pilot signals in $\mathbf{X}$. Although simulations show that this method requires three times more pilot sequences than RIS elements to obtain sufficiently accurate channel estimates, it also effectively solves the activity detection problem at the same time.

6.3.3.4 Matrix Calibration

In [20], the authors model the BS-RIS channel according to the Rician fading model, i.e., as being a summation of a deterministic part which represents the LoS link and a random part which represents the fast-fading part. Using a virtual representation of the channel via a grid of sampling angles, the channel $\mathbf{G}$ can be modelled as

$$\mathbf{G} = \sqrt{\frac{\gamma}{\gamma+1}}\mathbf{G}_{LoS} + \sqrt{\frac{1}{1+\gamma}}\mathbf{G}_{NLoS} \tag{6.19}$$

where γ is the Rician factor, $\mathbf{G}_{LoS}$ represents the deterministic LoS link, and $\mathbf{G}_{NLoS}$ represents the fast-fading part modelled as

$$\mathbf{G}_{NLoS} = \mathbf{A}_B \mathbf{S} \mathbf{A}_R^H. \tag{6.20}$$

Note that $\mathbf{S} \in \mathbb{C}^{M' \times N'}$ is the channel coefficient matrix assumed to be sparse in which each element represents the channel gain along the associated path, while $\mathbf{A}_B \in \mathbb{C}^{M \times M'}$ and $\mathbf{A}_R \in \mathbb{C}^{N \times N'}$ are as defined above. A similar modelling is used for the channel from each UE k to the RIS as

$$\mathbf{h}_k = \mathbf{A}_R \mathbf{h}'_k \tag{6.21}$$

where $\mathbf{h}'_k \in \mathbb{C}^{N' \times 1}$ is a sparse channel coefficients vector. The receive signal at the BS is thus expressed as

$$\mathbf{Y} = \left(\mathbf{H}_d + \mathbf{A}_B \mathbf{S} \mathbf{A}_R^H \mathbf{A}_R\right) \mathbf{H}'\mathbf{X} + \mathbf{N} \tag{6.22}$$

where the only unknowns are the matrices $\mathbf{S}$ and $\mathbf{H}^{'}$ which are then estimated via posterior mean estimators, i.e., by studying the MMSE, and derived using a sum-product message passing algorithm. Numerical results show that this method requires a number of training symbols that scale linearly with the number of UEs in order to obtain sufficiently good estimation of the channels. Note that since the number of UEs is usually less than the number of RIS elements, this method effectively reduces the training overhead compared to on/off schemes.

6.3.3.5 Orthogonal Matching Pursuit

Lastly, we present a set of works dealing with OMP-based channel estimation. We highlight their characteristics and present a third approach which tries to counteract its limitations.

The authors in [21] study a mmWave cellular system in which the first link $\mathbf{G}$ between the BS and the RIS is assumed to be dominated by the LoS part and thus known a priori. The channels $\{\mathbf{h}_k\}$ from each UE to the RIS are assumed to be sparse and are recovered using CS. In particular, using a virtual representation of the channel as in [20], an OMP algorithm is designed to recover the sparse coefficient vector.

The authors also study the design of the RIS configurations for each one of the T pilot sequences. Such sequences comprise both the phase shifts introduced by the RIS and the baseband part which is implemented at the BS. The design choice in this case is to match the BS-to-RIS major channel directions and to uniformly spread the signal along the angular dimension for the RIS-to-UEs channels. In such a way, the authors intend to exploit the known strong channel directions from the BS to the RIS which are dictated by the LoS path and to accurately sound the channel from the RIS to the UEs.

However, as it is well-known in the literature, the OMP algorithm may fail in case the sampling grid taken to sound the signal is not precise enough, i.e., if there are not enough degrees of freedom (e.g., in the form of antennas) at both the RIS and the BS.

The authors in [22] find a sparse representation for both channels $\mathbf{G}$ and $\mathbf{h}_k$ by exploiting the properties of the Kronecker and Khatri-Rao products. In particular, the BS-RIS channel is modelled as

$$\mathbf{G} = \mathbf{A}_B \mathbf{\Sigma} \mathbf{A}_R^H \tag{6.23}$$

where $\mathbf{A}_B$ and $\mathbf{A}_R$ are the over-complete pre-discretized grids of directions at the BS and RIS, respectively, while $\mathbf{\Sigma}$ is the sparse channel coefficient matrix. Similarly, the channel between the RIS and the UE is expressed as

$$\mathbf{h}_k = \mathbf{A}_R \boldsymbol{\alpha} \tag{6.24}$$

where $\boldsymbol{\alpha}$ is the sparse channel coefficient vector. The effective channel $\overline{\mathbf{H}}_k$ can be thus expressed as

$$\overline{\mathbf{H}}_k = \mathbf{D}_U \mathbf{\Lambda} \mathbf{A}_B^H \tag{6.25}$$

where $\mathbf{D}_U$ is a matrix constructed by taking the first $N^{'}$ columns of the matrix $\mathbf{A}_R^{*} {}^{\circ} \mathbf{A}_R$ with ${}^{\circ}$ representing the Khatri-Rao product and $\mathbf{\Lambda} = (\boldsymbol{\alpha}^{*} \otimes \mathbf{\Sigma})$ with $\otimes$ representing the Kronecker product. Hence, all the relevant channel information of both the BS-RIS and RIS-UE channels is contained in $\mathbf{\Lambda}$ which can be estimated via a conventional OMP algorithm.

Again, the authors assume that the true AoAs and AoDs are contained within the pre-discretized grids $\mathbf{A}_B$ and $\mathbf{A}_R$ thus neglecting possible mismatches which may cause the OMP algorithm to fail.

To overcome the aforementioned limitations, the authors in [23] propose an iterative reweighted method where the channel estimation is performed by sending in the downlink a series of T random training matrices each of which are reflected by the RIS with a random phase-shift matrix and combined (with a random combining matrix) by a single multi-antenna user. The composite channel between the BS and the RIS and the RIS and the user is assumed to be entirely LoS. Hence, the only parameters to be estimated are the instantaneous propagation path gains and AoA and AoD of the LoS links.

The channel estimation problem is formulated as the minimization of the sum over all training symbols of the matrix norm difference between the received signal and its parametric model which depends on the product of the instantaneous path gains between the BS and the RIS and between the RIS and the user and the corresponding AoAs and AoDs plus a regularization term which ensures sparsity of the estimated channel vector. In this first step, the output of the algorithm is an estimated product of instantaneous pathloss gains and an estimated difference of directional sine between the AoA and AoD, i.e., the difference between the sine of the AoD and the sine of the AoA. In a second step, both such parameters are further refined using gradient descent.

The authors compare their method over conventional OMP-based approaches demonstrating that it guarantees a higher sum rate performance. However, the gradient descent-based second step of their proposed method may result in a slow convergence of the overall algorithm.

6.3.3.6 Machine Learning

In [24], a fully connected artificial neural network is adopted in a RIS-aided wireless system to estimate the channels and phase angles from a reflected signal received through an RIS. The proposed deep network consists of four hidden layers, each of which is a fully connected layer followed by a hyperbolic tangent (tanh) activation function. The numbers of neurons in the fully connected layers are given following a test and trial method. To avoid overfitting of the network, the channel and additive white Gaussian noise (AWGN) intensities are shuffled at each iteration. The network maps the effects of the channel and phase angles on the transmitted signal using the nonlinear function approximation in its hidden layers. The proposed deep network yields an improved performance compared with the conventional LS and MMSE estimators.

In [25], a supervised deep learning framework is used for channel estimation in a RIS-assisted massive MIMO system. The authors designed a twin convolutional neural network (CNN) for the estimation of direct (BS-user) and cascaded (BS-RIS-user) channels. The CNN is fed with the received pilot signals, and it constructs a nonlinear relationship between the received signals and the channel data. First, all of the RIS elements are turned off using the BS backhaul link, and the deep network to estimate the direct channel is trained. Then, each of the RIS elements are turned on one by one to finally estimate the cascaded channel. In the deep network, real, imaginary, and the absolute value of each entry of the received signal is fed as input, because the use of "three-channel" data ameliorates the performance by enriching the features inherited in the input data. The approach is compared against state-of-the-art deep learning-based techniques, and performance gains are shown.

In [26–29], the authors adopted a design of a small portion of active elements on the RIS. In [28], to improve the channel estimation performance, the authors proposed to utilize deep learning to reduce the angle offset rate. While in [29], a complex-valued de-noising convolution neural network is further proposed to enhance performance.

6.4 The Road Ahead

In this section, we provide a non-exhaustive list of major open research problems that we consider to be of great importance for unveiling the potential benefits of RISs.

1. ***EM-based circuit models.*** Current studies on RIS mostly rely on simplified models of RIS. To obtain accurate characteristic of RIS functionalities, it is therefore imperative to develop basic understanding of the working principles of RIS by taking a physics-based approach on the analysis. In particular, the effect of the spatial coupling among the meta-atoms needs to be taken into account.
2. ***Path-loss and channel modeling.*** In order to obtain accurate performance limits of RIS in wireless networks, realistic models for the propagation of the signals scattered by the RIS are required. Additionally, one needs to consider not only the far-field regime, which is commonly assumed in a large portion of RIS analysis, but also in the near-field regime in which the benefits of RISs deployment may arise. Along this line of research, some fundamental works such as [30] have been proposed.
3. ***Fundamental performance limits.*** Depending on how a RIS is utilized, difference performance limits may be obtained. Therefore, it is important to develop theoretical frameworks that can capture these performance limits which are still largely unknown to date.
4. ***Large-scale networks: deployment, analysis, and optimization.*** Thanks to its low cost, low energy, and low complexity of deployments, RIS has an advantage over its competing technologies to be implemented in a large-scale environment. However, unfortunately, most studies in the literature are limited to "small-size" system models where usually one or only a few RISs are considered. To investigate the potential of large-scale RIS deployments, more studies need to be conducted that take into account large-scale networks with hundreds or possibly thousands of RIS elements.
5. ***Low-complexity channel estimation.*** Due to its passive nature, RIS lacks the ability to "sense" the wireless environment, and thus channel estimation is an integral part in designing a reliable system based on RIS. In this chapter, we introduced the state of the art of channel estimation in RIS-based systems such as on/off-based algorithm and machine-learning-based methods. The complexity

of these methods increases with the number of the RIS elements. Since RIS is normally made up hundreds or thousands of elements, an improvement on low-complexity channel estimation method is essential in order to bring RIS into realization.

Acknowledgments The work of F. Danufane and J. Liu was supported in part by the European Commission through the H2020 5GstepFWD project under grant agreement number 722429.

References

1. ITU-R (2015) *IMT traffic estimates for the years 2020 to 2030*. http://www.itu.int/pub/R-REP-M.2370
2. Danufane, F. H., Di Renzo, M., De Rosny, J., & Tretyakov, S. (2020). *On the path-loss of reconfigurable intelligent surfaces: An approach based on green's theorem applied to vector fields*. https://arxiv.org/abs/2007.13158
3. Arun, V., & Balakrishnan, H. (2020). RFocus: Beamforming using thousands of passive antennas. In *17th USENIX symposium on networked systems design and implementation (NSDI 20),* pp. 104 7–1061.
4. NTT DOCOMO. (2000). *DOCOMO conducts world's first successful trial of transparent dynamic metasurface*. https://www.nttdocomo.co.jp/english/info/media_center/pr/2020/0117_00.html. Accessed 26 Aug 2000.
5. Yu, N., Genevet, P., Kats, M. A., et al. (2011). Light propagation with phase discontinuities: Generalized laws of reflection and refraction. *Science, 334*(6054), 333–337. https://doi.org/10.1126/science.1210713
6. Di Renzo, M., Ntontin, K., Song, J., Danufane, F. H., Qian, X., et al. (2020). Reconfigurable intelligent surfaces vs. relaying: Differences, similarities, and performance comparison. *IEEE Open Journal of the Communications Society, 1*, 798–807. https://doi.org/10.1109/OJCOMS.2020.3002955
7. Nadeem, Q. U. A., Kammoun, A., Chaaban, A., et al. (2019). *Intelligent reflecting surface assisted wireless communication: Modeling and channel estimation*. https://arxiv.org/abs/1906.02360
8. Lin, J., Wang, G., & Fan, R., et al. (2019). *Channel estimation for wireless communication systems assisted by large intelligent surfaces*. https://arxiv.org/abs/1911.02158
9. Mishra, D., & Johansson, H. (2019). Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp 4659–4663. https://doi.org/10.1109/ICASSP.2019.8683663
10. Yang, Y., Zheng, B., Zhang, S., et al. (2020). Intelligent reflecting surface meets OFDM: Protocol design and rate maximization. *IEEE Transactions on Communications, 68*(7), 4522–4535. https://doi.org/10.1109/TCOMM.2020.2981458
11. Zheng, B., & Zhang, R. (2019). Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization. *IEEE Wireless Communications Letters, 9*(4), 518–522. https://doi.org/10.1109/LWC.2019.2961357
12. You, C., Zheng, B., & Zhang, R. (2019). Intelligent reflecting surface with discrete phase shifts: Channel estimation and passive beamforming. In *IEEE International Conference on Communications (ICC)*, Dublin, Ireland, pp. 1–6. https://doi.org/10.1109/ICC40277.2020.9149292
13. Jensen, T. L., & De Carvalho, E. (2020). An optimal channel estimation scheme for intelligent reflecting surfaces based on a minimum variance unbiased estimator. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5000–5004. https://doi.org/10.1109/ICASSP40776.2020.9053695

14. Wang, Z., Liu, L., & Cui, S. (2020). Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis. *IEEE Transactions on Wireless Communications*.https://doi.org/10.1109/TWC.2020.3004330
15. Wei, L., Huang, C., Alexandropoulos, G. C., et al. (2020). Parallel factor decomposition channel estimation in RIS-assisted multi-user MISO communication. In *IEEE 11th sensor array and multichannel signal processing workshop (SAM)*, pp. 1–5. https://doi.org/10.1109/SAM48682.2020.9104305
16. Chen, J., Liang, Y. C., Cheng, H. V., et al. (2019). *Channel estimation for reconfigurable intelligent surface aided multi-user MIMO systems*. https://arxiv.org/abs/1912.03619
17. Ning, B., Chen, Z., Chen, W., et al. (2019). Channel estimation and transmission for intelligent reflecting surface assisted THz communications. In *IEEE international conference on communications (ICC)*, pp. 1–7. https://doi.org/10.1109/ICC40277.2020.9149153
18. Cui, Y., & Yin, H. (2019). *An efficient CSI acquisition method for intelligent reflecting surface-assisted mmwave networks*. https://arxiv.org/abs/1912.12076
19. Xia, S., & Shi, Y. (2020). Intelligent reflecting surface for massive device connectivity: Joint activity detection and channel estimation. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5175–5179. https://doi.org/10.1109/ICASSP40776.2020.9054415
20. Liu, H., Yuan, X., & Jun, Y. (2020). Matrix-calibration-based cascaded channel estimation for reconfigurable intelligent surface assisted multiuser MIMO. *IEEE Journal on Selected Areas in Communications*.https://doi.org/10.1109/JSAC.2020.3007057
21. Wan Z, Gao Z, Alouini M S (2020) Broadband channel estimation for intelligent reflecting surface aided mmWave massive MIMO systems. https://arxiv.org/abs/2002.01629
22. Wang, P., Fang, J., Duan, H., et al. (2020). Compressed channel estimation and joint beamforming for intelligent reflecting surface-assisted millimeter wave systems. *IEEE Signal Processing Letters, 27*, 905–909. https://doi.org/10.1109/LSP.2020.2998357
23. He, J., Leinonen, M., & Wymeersch, H., et al. (2020). *Channel estimation for RIS-aided mmWave MIMO channels*. https://arxiv.org/abs/2002.06453
24. Khan, S., & Shin, S. Y. (2019). *Deep-learning-aided detection for reconfigurable intelligent surfaces*. https://arxiv.org/abs/1910.09136
25. Elbir, A. M., Papazafeiropoulos, A., Kourtessis, P., et al. (2020). Deep channel learning for large intelligent surfaces aided mm-wave massive MIMO systems. *IEEE Wireless Communications Letters*.https://doi.org/10.1109/LWC.2020.2993699
26. Taha, A., Alrabeiah, M., & Alkhateeb, A. (2019). *Enabling large intelligent surfaces with compressive sensing and deep learning*. https://arxiv.org/abs/1904.10136
27. Taha, A., Zhang, Y., Mismar, F. B., et al. (2020). Deep reinforcement learning for intelligent reflecting surfaces: Towards standalone operation. In *2020 IEEE 21st international workshop on signal processing advances in wireless communications (SPAWC)*, Atlanta, GA, USA, pp. 1–5. https://doi.org/10.1109/SPAWC48557.2020.9154301
28. Jiang, F., Yang, L., da Costa, D. B., et al. (2020). *Channel estimation via direct calculation and deep learning for RIS-aided mmWave systems*. https://arxiv.org/abs/2008.04704
29. Liu, S., Gao, Z., Zhang, J., et al. (2020). Deep denoising neural network assisted compressive channel estimation for mmWave intelligent reflecting surfaces. *IEEE Transactions on Vehicular Technology, 69*(8), 9223–9228. https://doi.org/10.1109/TVT.2020.3005402
30. Danufane, F. H., Di Renzo, M., de Rosny, J., et al. (2020). *On the path-loss of reconfigurable intelligent surfaces: An approach based on green's theorem applied to vector fields*. https://arxiv.org/abs/2007.13158

Part III
PON Technology for UDNs

Chapter 7
Integrated Optical-Wireless Interface and Detection

Dimitrios Konstantinou, Lei Xue, Tanjil Shivan, Maruf Hossain, Simon Rommel, Ulf Johannsen, Christophe Caillaud, Viktor Krozer, Jiajia Chen, and Idelfonso Tafur Monroy

Abstract This chapter elaborates on the beneficial aspects and hardware implementations of incorporating ultradense WDM-PONs (UDWDM-PONs) with hybrid optical-wireless fronthaul links and fiber to the home applications. Simulation results on the synthesis of a low-cost and low-energy consumption optoelectronic unit within the future 5G base stations (BS) are presented. In addition, an advanced neural network is investigated capable of compensating for the linear and nonlinear effects induced by semiconductor optical amplifiers (SOA).

7.1 Introduction

Passive optical networks (PONs) are associated with fiber to the home (FTTH) connections providing broadband and high-speed communications. Due to the emergence of new technologies, a large amount of consumer and business use cases that need to be supported by PONs demand the migration from the existing WDM-PONs to ultradense WDM-PONs within urban areas. Moreover, in the wireless domain, the dawn of 5G had provided the impetus for future radio access networks to support 1000 times more capacity enabling high-speed connections to a huge

D. Konstantinou (✉) · S. Rommel · U. Johannsen · I. T. Monroy
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: d.konstantinou@tue.nl

L. Xue · J. Chen
Chalmers University of Technology, Göteborg, Sweden
e-mail: leixu@chalmers.se

T. Shivan · M. Hossain · V. Krozer
Ferdinand-Braun-Institut, Leibniz-Institut für Höchstfrequenztechnik (FBH), Berlin, Germany
e-mail: Tanjil.Shivan@fbh-berlin.de

C. Caillaud
III-V Lab (A Joint Laboratory Between Nokia Bell Labs, Thales R&T and CEA Leti), Palaiseau, France
e-mail: christophe.caillaud@3-5lab.fr

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_7

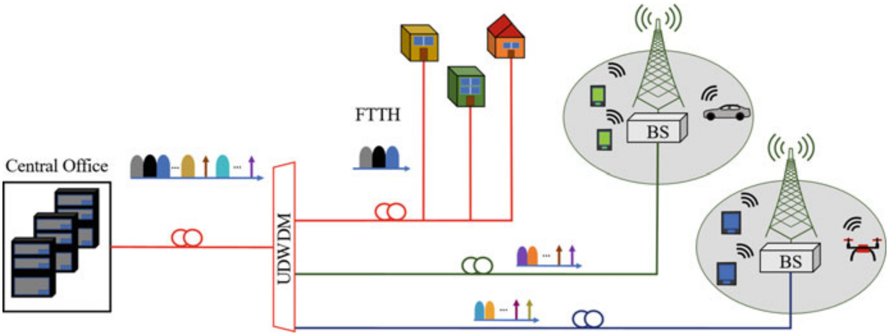


Fig. 7.1 A block diagram of the network proposed by the European ITN project 5G STEP FWD. Optical signals spaced within the UDWDM grid serve different applications such as FTTH, as well as mm-wave wireless access to various end-user applications such as autonomous vehicular communications

amount of users in a dense environment with guaranteed quality of service and exceptionally low latency [1]. A dynamic way to achieve these aforementioned requirements is by employing hybrid photonic-wireless links operating in the lightly licensed millimeter-wave (mm-wave) bands that are integrated with the UDWDM-PONs.

The core concept presented in this chapter is illustrated in Fig. 7.1. At the central office (CO), the optical signal generation and multiplexing will take place. The resource management and allocation within the UDWDM-PON are offered by a centralized control plane defined by the network function virtualization (NFV) and software-defined networking (SDN). This leads to a more flexible routing allowing the reduction of the network's complexity increasing control and minimizing the communications latency. Moreover, combining analogue radio over fiber (ARoF) fronthaul with flexible carrier aggregation allows optimum resource utilization. In the optical domain, 5G New Radio (NR) modulated signals, and signals for FTTH (e.g., NRZ, PAM4) propagate within the UDWDM-PON toward their designated terminal (5G base stations or homes, respectively) through the support provided by a UDWDM de-multiplexer that contributes to the separation of these optical waves [2]. Since optical heterodyning will be employed for the mm-wave generation within the future 5G base stations, unmodulated local oscillators are also multiplexed with the rest of the optical signals at the central office [3]. The remainder of this chapter aims to provide solutions toward the integration of optoelectronic integrated components within the 5G base stations, as well as to present digital signal processing techniques to mitigate distortions caused to FTTH signals by the passive optical network.

Within Sect. 7.2, the physical concepts describing optical heterodyning and the operation of uni-travelling carrier photodiodes (UTC-PDs) are explained. Moreover, the processes on the acquisition of the full-equivalent circuit of these devices are analyzed by using both simulation tools and mathematical equations. Furthermore, the properties of transimpedance amplifiers (TIAs) are explained in terms of gain and noise. Finally, the results of a co-simulation between a UTC-PD and a multistage TIA operating in V-band are discussed underlining all the important aspects of high-speed, high-gain optoelectronic co-integration.

For FTTH applications, the advanced 25G avalanche photodiodes with high sensitivity are costly in the passive optical networks. Thus, the combination of semiconductor optical amplifiers (SOA) used as pre-amplifier with photodiodes (PD) is proposed as a low-cost solution to improve system power budget. However, the SOA nonlinearities due to the gain saturation, e.g., the pattern effect, degrade the system performance significantly that increasingly worsen with increasing system speed. Thus, digital signal processing (DSP) is often used as a powerful tool and combined with the use of a neural network to compensate the linear and nonlinear distortions. The performance of NN is investigated in Sect. 7.3 within an intensity modulation with direct detection (IM/DD) system, with 50G PAM4 signals showcasing the nonlinearity compensation and receiver dynamic range improvement.

7.2 Hybrid Photonic-Wireless Interface Unit for 5G Base Stations

The interface between optics and RF high-speed electronics will take place at the 5G base stations of Fig. 7.2. Channel interleaving controlled by the SDN/NFV plane and de-multiplexing (DeMUX) will guarantee that each modulated optical signal is matched to an unmodulated tone (LO) such that their frequency distance is equivalent to the desired RF carrier frequency of the mm-wave.

Each output of the DeMUX will provide input to integrated photonic-wireless transmitting units. Such units are novel photonic-wireless chip interfaces comprising high-speed photodiodes and RF electronic circuits. Uni-travelling carrier photodiodes (UTC-PDs) are promising candidates for the conversion of optical signals to mm-waves through optical heterodyning [4, 5]. The generated mm-waves will be amplified by broadband and high-gain integrated transimpedance amplifiers (TIAs) and transmitted in the wireless domain by mm-wave antennas [6].

The wireless operation in the mm-wave bands supports the transmission using wide wireless channels that are capable of providing peak user data rates in the Gbit/s range. The use of traditional methods for digitized fronthaul and electrical up-conversion of baseband signals to mm-wave frequencies is limited to narrow bandwidth signals not appropriate for high data throughput. Therefore, the signal up-conversion through optical heterodyning and analogue radio over fiber has become a promising solution.

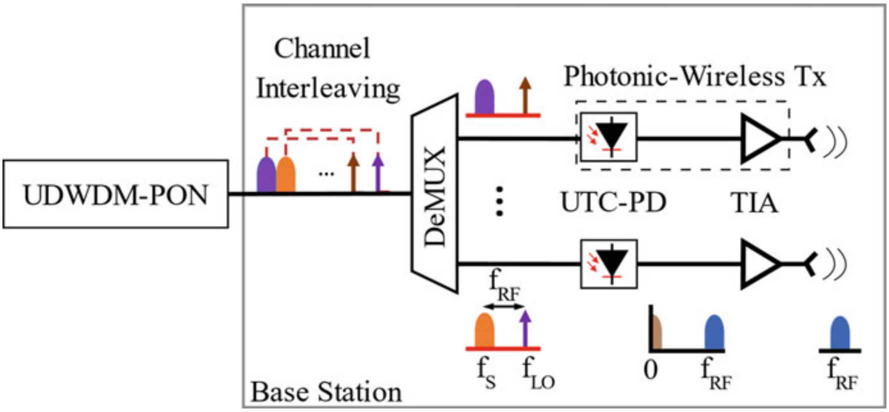


Fig. 7.2 A base station fed by UDWDM-PON where the mm-wave generation and transmission are conducted by photonic-wireless integrated chips that include high-speed UTC-PDs, TIAs, and antenna elements

7.2.1 Optical Heterodyning for mm-wave Generation

The millimeter-wave generation in the optical domain has been widely studied and is based on optical heterodyning, i.e., the beating of two optical signals spaced at a desired RF frequency ($f_{RF} = f_{LO} - f_S$) on a photodiode.

The derivations describing this physical concept are analyzed based on mathematical calculations. As shown at the bottom of Fig. 7.2, the electric fields for both the signal ($\boldsymbol{E_s}$) carrying modulated data and a local oscillator ($\boldsymbol{E_{LO}}$) can be expressed in (7.1) [7] as

$$\boldsymbol{E_s}(t) = \sqrt{P_s}s(t)e^{-i(\omega_s t+\phi_s)}$$

$$\boldsymbol{E_{LO}}(t) = \sqrt{P_{LO}}e^{-i(\omega_{LO} t+\phi_{LO})} \tag{7.1}$$

where $s(t)$ is the modulated signal in baseband and P_s, P_{LO} are the optical signal powers. Both electric fields oscillate at radial frequencies ω_s, ω_{LO} and carry phase components ϕ_s, ϕ_{LO}, which depend mostly on the properties of the sources used (e.g., laser linewidth). Moreover, the local oscillator (LO) is a co-propagating optical tone that carries no modulation.

Typically, in electronic mixers, LO signals are special as these are driving the nonlinear device (i.e., mixer) into large-signal excitation and the RF signal that is up-/down-converted assumed being small, giving rise to a linearized periodic circuit. This is not the case in optical heterodyning within photodiodes.

Thus, assuming that both fields are co-polarized, the electric field at the photodiode ($\boldsymbol{E_{PD}}$) is calculated in (7.2):

$$
\begin{aligned}
\boldsymbol{E}_{PD} &\propto |\boldsymbol{E}_s(t) + \boldsymbol{E}_{LO}(t)|^2 \\
&= [\boldsymbol{E}_s(t) + \boldsymbol{E}_{LO}(t)]\,\overline{[\,\boldsymbol{E}_s(t) + \boldsymbol{E}_{LO}(t)]} \\
&= |\boldsymbol{E}_s(t)|^2 + |\boldsymbol{E}_{LO}(t)|^2 + \boldsymbol{E}_s(t)\overline{\boldsymbol{E}_{LO}(t)} + \overline{\boldsymbol{E}_s(t)}\boldsymbol{E}_{LO}(t) \\
&= P_s s^2(t) + P_{LO} + 2\sqrt{P_s P_{LO}}\, s(t)\ \cos\left(\omega_{RF} t + \phi_{RF}\right)
\end{aligned}
\tag{7.2}
$$

with the desired RF signal given as $2\sqrt{P_s P_{LO}}\, s(t)\ \cos\left(\omega_{RF} t + \phi_{RF}\right)$. The components $P_s s^2(t) + P_{LO}$ are filtered out by the waveguide connected to the mm-wave antenna.

Therefore, the optical heterodyning of two optical tones is an efficient and simple way to generate mm-wave signals. This process is commonly referred to as photonic up-conversion since any modulation present on the optical signals is transferred in the RF domain. Furthermore, this process allows the simplification of the wireless transmitters since no high-frequency sources are needed at the base stations to electrically up-convert the signals. Finally, since the LOs are generated at the central office, the total network topology is even more centralized, and the costs of the base stations are reduced even further.

7.2.2 Physical Properties of Photodiodes

In order to thoroughly comprehend the mm-wave generation based on optical heterodyning, it is essential for the reader to grasp the physical properties and functionalities of photodiodes. In principle, once photons are absorbed by a PD, electron-hole pairs are generated and travel through its junction. There are different methods for coupling light to the absorption layer of a PD such as the evanescent coupling [8] that is illustrated in Fig. 7.3a. In this case, light propagating through the optical waveguide is gradually absorbed by the carrier generation layer.

One example of such optoelectronic structure is the PIN photodiode of Fig. 7.3b that consists of three different types of semiconductor materials: an intrinsic region that acts as the carrier generation layer once the absorbed photons have energy

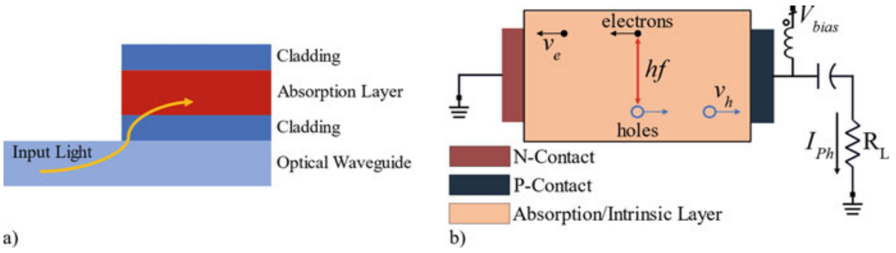


Fig. 7.3 (**a**) A block diagram depicting the evanescent coupling of light to the absorption region of PDs; (**b**) A PIN-PD capable of generating current at its load RL due to the creation of electron-hole pairs by the absorption of photons with energy hf

($E = hf$) higher than its bandgap, a P-type semiconductor collecting the generated holes, and an N-type material attracting the electrons [9]. If an adequate reverse bias (V_{bias}) is applied, the intrinsic region becomes fully depleted, and a strong electric field is established across the junction accumulating the photogenerated carriers that drift toward the contacts of the device inducing a photocurrent (I_{ph}) at its load (R_L).

A fundamental parameter of a PD is its responsivity (R_{opt}), calculating the ratio of the flux of generated electrical carrier over the one of incident photons. Therefore, the responsivity is given as the ratio of I_{ph}/P_{opt} and is measured in A/W. Moreover, except for the generated photocurrent, there is an additional noise parameter that is due to the random generation and annihilation of charges within the depletion region of the PD. This noise source is defined as dark current (I_{dark}), and its amplitude resides within the range of nA.

Apart from the abovementioned static parameters of a photodiode, the response of the device over frequency is also critical. The 3 dB bandwidth (f_{3dB}) of a PD consists of two elements. The first is determined by the RC time constant (τ_{RC}) of the device that depends on the total resistance (R) and capacitance (C) of the PD including its junction capacitance (C_j), its series resistance (R_s), and all the parasitics added from the transmission line (TML). The second element of (7.3) depends on the transit time (τ_{tr}) of the carriers generated in the absorption layer. This parameter varies based on the type of the PD and the speed of the generated electrons (v_e) and holes (v_h). Since $v_e > v_h$ [10], the transit time of the holes within the device defines its speed.

$$f_{3dB} = \sqrt{\frac{1}{\frac{1}{f_{RC}^2} + \frac{1}{f_{tr}^2}}},\ f_{RC} = \frac{1}{2\pi\tau_{RC}} = \frac{1}{2\pi RC},\ f_{tr} = \frac{1}{2\pi\tau_{tr}} \tag{7.3}$$

Except for the PIN-PDs, avalanche photodiodes (APDs) are used in telecommunication systems. APDs contain an avalanche multiplication layer neighboring the absorption region where a single photon produces hundreds of electron-hole pairs leading to an amplification factor [11]. However, these devices are limited in terms of noise and bandwidth comparing to uni-travelling carrier photodiodes (UTC-PDs) that show great potential in terms of power and speed [12].

7.2.3 The Uni-travelling Carrier Photodiode (UTC-PD)

A uni-travelling carrier photodiode (UTC-PD) is an example of a high-speed photodetecting device providing increased sensitivity, broad bandwidth, and high saturation powers.

A block diagram of the overall structure of a UTC-PD is presented in Fig. 7.4. The carrier generation takes place at a thin absorption layer [13]. In addition, a diffusion block layer supports the unidirectional motion of electrons toward the

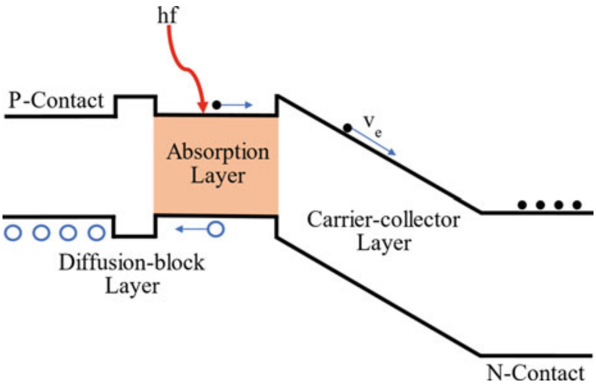


Fig. 7.4 A block diagram showing the different semiconductor material layers within a UTC-PD and the carrier generation while photons are absorbed

N-Contact. The holes are directly swept at the P-Contact within the dielectric relaxation time of the P-doped layer (>THz bandwidth). Thus, the transit time of the device (τ_{tr}) is mainly limited by the speed of the electrons within the total structure from the absorption region to the carrier-collector layer to the N-Contact. Since the velocity of the electrons in the carrier-collector layer is much higher than the velocity of the holes in the absorption layer, the bandwidth of these devices is higher than the traditional PIN-PDs.

Concerning the experimental characterization of UTC-PDs, there are various parameters that can be investigated such as responsivity and dark current measurement in the DC domain or RF saturation powers and 3 dB bandwidth. Moreover, an important characterization process is the measurement of the reflection coefficients of a UTC-PD with a vector network analyzer (VNA) and its correlation with lumped electronic elements.

7.2.3.1 Synthesis of the Equivalent Circuit of UTC-PDs

The measurement of the equivalent circuit of the diodes is of utmost importance since they provide crucial information about their physical properties and limitations. Therefore, it is imperative that the S-parameters of the device are measured.

Hence, as shown in Fig. 7.5a, a VNA is used to extract the reflection coefficients (S_{11}) of the waveguide pad structures (open and short) as well as the UTC-PDs of Fig. 7.5b. An essential component for this measurement is the RF probe touching the pads of the diode. Between the RF probe and the VNA, a power supply is connected providing a reverse bias to the diodes under test. Within the VNA, an internal bias tee is responsible for the isolation of the DC photocurrent generated by the photodiode allowing only the RF signal to be received.

A sample of the S_{11} parameters that are obtained for the measurements of open (OC) and short (SC) on-wafer structures as well as for a UTC-PD at an applied bias of -3 V is depicted in Fig. 7.6a, b. It can be observed that the curvature of the short circuit pad has an inductive behavior, and therefore the measured waveguide

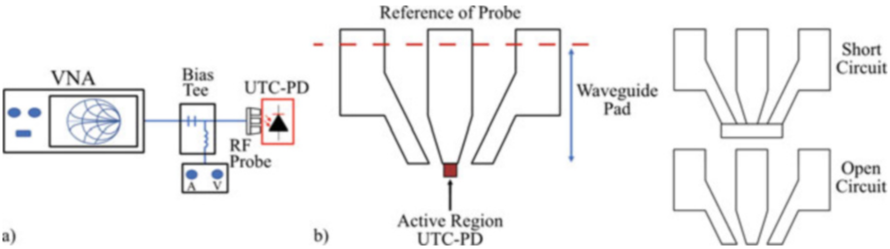


Fig. 7.5 (**a**) The measurement setup extracting the S_{11} parameters of UTC-PDs; (**b**) The UTC-PD, open and short on-wafer structures used for measurements of reflection coefficient parameters

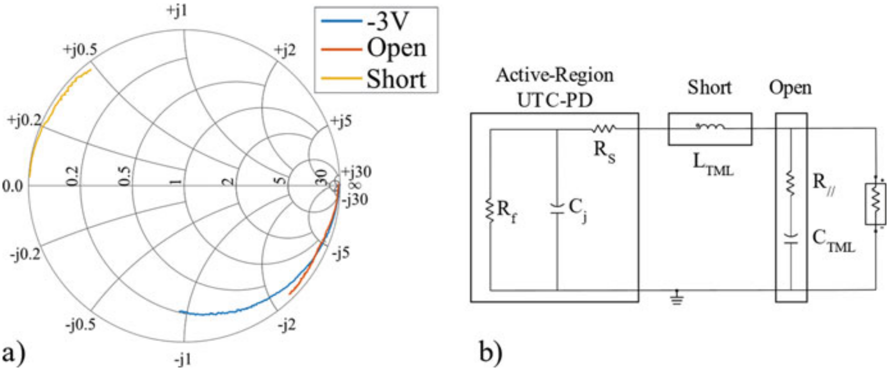


Fig. 7.6 (**a**) The Smith charts of the S_{11} parameters for the open and short waveguide as well as for a 4 × 20 μm^2; (**b**) The lumped equivalent circuit in ADS of a photodiode based on S_{11} parameters

parameters can be matched with an inductor (L_{TML}). The similar process models the open circuit that has a capacitance (C_{TML}). Moreover, a resistor ($R_{//}$) is added in series to C_{TML} to correct for the non-ideal performance of the open structure (e.g., potential leakage to substrate) [14]. In some cases, $R_{//}$ is small and can be omitted. Then, the junction capacitance (C_j) and series resistance (R_S) within the active region of the UTC-PDs are calculated. Finally, a parasitic resistor is shunt (R_f) in parallel to (C_j) achieving a perfect matching between the model and the measurements. In the majority of the models, the (R_f) is calculated to be very high (within $k\Omega$), and this parallel branch can be considered as open. The extraction of the components synthesizing the equivalent circuit can be achieved either by using a circuit design software such as ADS [15] by Keysight or analytically through mathematical derivations.

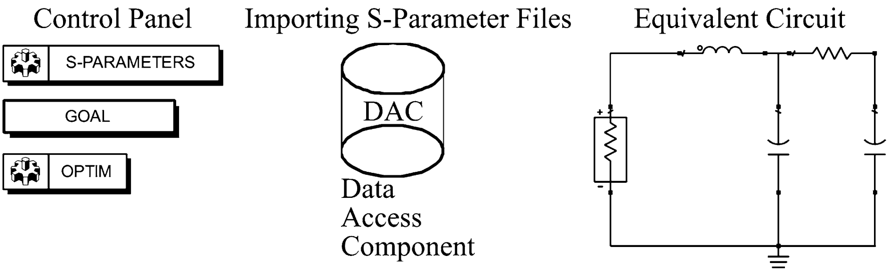


Fig. 7.7 An overview for the simulation in ADS of a lumped circuit based on obtained S_{11} parameters

Table 7.1 The values of the lumped components for the full equivalent simulation of a $4 \times 20\ \mu m^2$ UTC-PD at a reverse bias of bias of 3 V

Symbol	Description	Value
C_{TML}	TML capacitance	21.26 fF
$R_{//}$	TML resistor	1.27 Ω
L_{TML}	TML inductance	55.63 pH
C_j	Junction capacitance	24.16 fF
R_S	Series resistance	24.33 Ω
R_f	Parasitic resistance	3.61 kΩ

Extraction of Equivalent Circuit Parameters with ADS

An example of the processes followed in the software environment of ADS is illustrated in Fig. 7.7. The measured (S_{11}) data are imported to the software in the .s2p file form through an S-parameter reading block. In the next step, a circuit based on lumped components is composed.

Furthermore, the S-parameters of the measured structures (S_{meas}) are matched with the parameters of the model (S_{model}) using the optimization tool of ADS. The optimum point is achieved when the error (ϵ) margin defined by (7.4) is minimized and within the objective range defined. Once the optimization is finalized, the values of the lumped components are automatically updated.

$$\epsilon = \frac{|S_{meas} - S_{model}|}{|S_{meas}|} * 100\% \tag{7.4}$$

The percentage of the error calculation derived from the extraction of the equivalent circuit based on the data for the short, open, and UTC-PD structures is below 5%, confirming that a lumped element equivalent circuit can be synthesized based on reflection coefficient measurements with high accuracy. The values of the lumped elements can be observed in Table 7.1. However, due to the linear nature of the equivalent circuit, the optimized lumped element values in the circuit are not unique.

Analytical Extraction of Equivalent Circuit Parameters

An alternative to the abovementioned method is the calculation of the lumped components through asymptotic equations based on the measured reflection coefficient data. As a first step, all the measured reflection coefficient parameters are translated into impedances, and the parameters of the waveguide are calculated based on (7.5), which are determined from the open circuit measurement as

$$\mathrm{Z}_{meas} = \mathrm{Z}_0 \frac{1 + \mathrm{S}_{meas}}{1 - \mathrm{S}_{meas}}$$

$$C_{TML} = \frac{1}{j\omega Z_{OC}}, L_{TML} = \frac{Z_{SC}}{j\omega} \tag{7.5}$$

It is derived that $C_{TML} = 22.7$ fF and $L_{TML} = 55.2$ pH that differ from the ADS simulation results by 5.7% and 0.7%, respectively.

Then by calculating Z_{Series}, all the parallel components added to the UTC-PD by the transmission lines are removed without taking into account the additional inductors and resistors. Then C_j and R_s can be obtained from (7.6) that is based on the previous calculations.

$$Z_{Series} = \left(Z_{meas}{}^{-1} - Z_{OC}{}^{-1}\right)^{-1}$$

$$R_s = \mathrm{Re}\left[Z_{Series}\right], C_j = \frac{1}{\left(L_{TML}\omega - \mathrm{Im}\left[Z_{Series}\right]\right)\omega} \tag{7.6}$$

The obtained C_j is equal to 23.3 fF and $R_s = 24.5\ \Omega$ with error percentages 3.5% and 0.7%, respectively, in comparison with the results of the simulated ADS lumped components.

Based on the results above, the analytical extraction technique can provide a good estimate of the total S_{11} based equivalent circuit. However, as the percentage error increases, a more complex RC network is needed to be synthesized in order to analyze the UTC-PD once the applied bias voltage is not sufficient to fully deplete the intrinsic region of the device or in the case where the waveguide structures suffer from a leakage on the substrate. In this case, and for verification purposes, one can also employ an analytical extraction method for frequency asymptotes ($f \to \infty, f \to 0$). For this analysis, robust values for all equivalent circuit parameters can be found for different bias values.

7.2.4 The Transimpedance Amplifier

The generated mm-wave signals at the output of the UTC-PD require amplification in order to be transmitted to the 5G end users. A common type of amplifier used in communication systems that require broad bandwidth and high sensitivities is the transimpedance amplifier (TIA).

A simplified TIA architecture is shown in Fig. 7.8. A UTC-PD is connected to a negative feedback amplifier. The TIA circuit is a current-voltage converter, amplifying the mm-wave signal from the photodiode as current, and converts it into a voltage so as to be compatible at the output with equipment that in most cases is designed to be matched to 50 Ω. An important metric for the gain of TIAs is the transimpedance gain (Z_T) that is calculated as the ratio between the output voltage (V_{out}) over the input current (I_{ph}). The term transimpedance stems from the fact that the current and voltage defining Z_T are measured on two different ports and provide broad bandwidth and high gain without compromising the signal to noise ratio – calculated as the ratio between the output voltage (V_{out}) over the input current (I_{ph})) [16]. Z_T can be calculated in Ω or dBΩ based on (7.7):

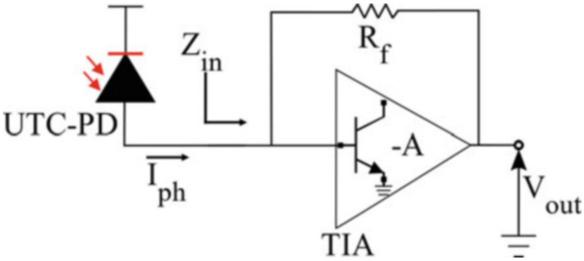
$$Z_T = \frac{V_{out}}{I_{ph}} \ (\Omega)\, , Z_T = 20\log\left(Z_T/\Omega\right) \ (dB\Omega) \tag{7.7}$$

Furthermore, the input impedance (Z_{in}) of the TIA needs to be low in order to match the output of the UTC-PDs. Based on Fig. 7.8, Z_T and Z_{in} are calculated in (7.8) as

$$Z_T = -\frac{A}{A+1}R_f, Z_{in} = -\frac{1}{A}Z_T = \frac{R_f}{A+1} \tag{7.8}$$

where A is the open loop gain and R_f is the feedback resistance. While designing such amplifiers, in order to maximize the transimpedance, the gain of the open loop amplifier needs to be maximized. The higher the Z_T, the higher the match between the TIA and the UTC-PD. In the ideal case where A is infinite, the transimpedance depends only on the feedback resistor and $Z_{in} = 0$. Even though that is not feasible

Fig. 7.8 An overview of the architecture combining a UTC-PD and a TIA supporting the generation of mm-waves



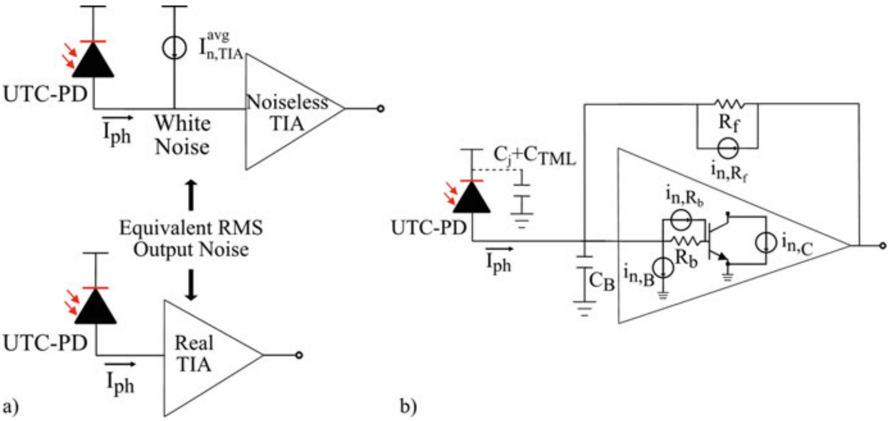


Fig. 7.9 (**a**) Definition of the input-referred noise current; (**b**) the noise sources of a TIA that includes HBT transistors

to be achieved, the abovementioned equations provide a general understanding on the basic function of this device.

7.2.4.1 Noise Analysis of TIAs

Concerning the noise calculations of TIAs, the input-referred noise current is a crucial parameter for this type of amplifier. As illustrated in Fig. 7.9a, this current is defined such that combined with a noiseless TIA, it is capable of reproducing the same noise levels as the actual (noisy TIA). The input-referred noise current can be measured in three different ways.

- Input-referred noise current PSD

  It is the power spectral density (PSD) $I^2_{n,TIA}$ of the input-referred noise current measured in pA^2/Hz and the noise current density ($\sqrt{I^2_{n,TIA}}$) in $pA/\sqrt{Hz}$. As illustrated in Fig. 7.9b, there is a multitude of noise sources contributing to $I^2_{n,TIA}$, including the thermal noise of the feedback resistance and the noise currents at the base and collector of the transistor of the open-loop amplifier [17, 18]. Therefore, this PSD is not white, and it cannot be provided by a single number [19]. A theoretical estimation of $I^2_{in,TIA}$ is given in (7.9).

$$I^2_{n,TIA}(f) \approx \frac{4kT}{R_f} + \frac{2qI_C}{\beta} + 2qI_C\frac{(2\pi C_{Tot})^2}{g_m^2}f^2 + 4kTR_b\big(2\pi C_j\big)^2 f^2 \tag{7.9}$$

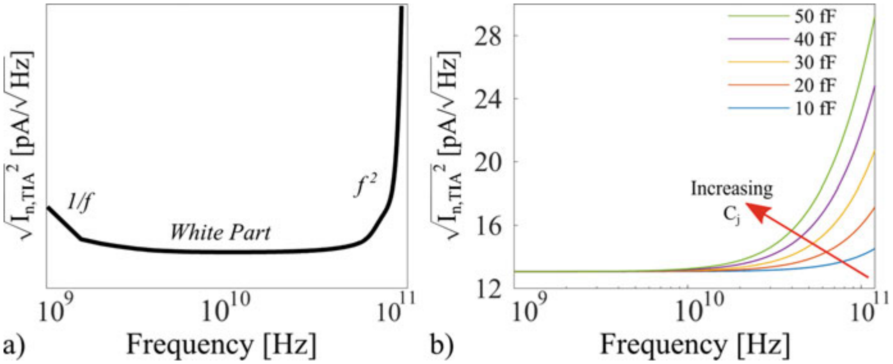


Fig. 7.10 (**a**) The typical curve of the noise current spectral density of a TIA; (**b**) the impact of the increasing junction capacitance of a UTC-PD to the $\sqrt{I_{n,TIA}^2}$

As shown in Fig. 7.10a, the $\sqrt{I_{n,TIA}^2}$ consists of a low-frequency part (1/*f*) due to the shot noise in the base and collector, white noise component, and two high-frequency (f^2) terms that are directly affected by the junction capacitance of the photodiode used since total input capacitance to the TIA is equal to $C_{Tot} = C_j + C_{TML} + C_B$.The dependence of C_j to $I_{n,TIA}^2$ is investigated in Fig. 7.10b for a single-feedback TIA based on (7.9) with R_f at the level of 250 Ω and β =30.2, I_C =9.4 mA, g_m =0.38 S, and R_b =40 Ω. By increasing C_j, the $I_{n,TIA}^2$ increases exponentially at higher frequencies. Thus, in the case where a combined module of a UTC-PD and TIA is designed, the C_j is required to be as low as possible.

- Input-referred RMS noise current

  It is an RMS value ($i_{n,TIA}^{rms}$) given by a single number (in nA or μA) and is extracted from the ratio of the RMS output voltage over the midband transimpedance value of the TIA [20]. The analytical equation is provided in

$$i_{n,TIA}^{RMS} = \frac{1}{Z_T}\sqrt{\int_0^{2\Delta f_1} v_{n,TIA}^{RMS} df} = \frac{1}{\overline{\overline{Z_T}}}\sqrt{\int_0^{2\Delta f_1} \left|Z_T(f)^2\right| I_{n,TIA}^2(f)df} \tag{7.10}$$

where the term $\Delta f_1 = \frac{\pi}{2} f_{3dB}$ is the first-order-equivalent noise bandwidth of the TIA and is dependent on the f_{3dB} of the device.

- Averaged input-referred RMS noise current density

  This noise current definition ($I_{n,TIA}^{Avg}$) is measured in pA/√Hz and obtained by the ratio given by (7.11).

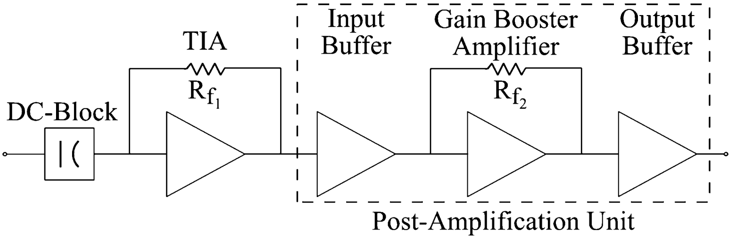


Fig. 7.11 The block diagram of a multistage TIA

$$I_{n,TIA}^{Avg} = \frac{i_{n,TIA}^{rms}}{\Delta f_1} \tag{7.11}$$

As presented in Fig. 7.9a, it is interpreted as the white noise source that needs to be connected to the input of a noise free TIA in order to reproduce the RMS output voltage noise of the real TIA.

The definitions analyzed in these sections provide valuable information on the results obtained in the TIA simulations.

7.2.4.2 Simulations on the Multistage TIA

Figure 7.11 shows a block diagram of the simulated multistage TIA design. At the RF input, a DC-block (capacitor) is added, prohibiting the reverse flow of the DC current toward the UTC-PD. A post-amplification unit after the first TIA further increases the total gain. It consists of an input buffer, a second transimpedance gain booster stage with a second feedback, and an output buffer driving the mm-wave antenna capable of transmitting mm-waves.

The amplifier circuit architecture is based on the common principle of inter-stage impedance mismatch where a transimpedance stage (TIS) is followed by a transadmittance stage (TAS). This method leads to the maximization of the device bandwidth [21]. The TIA is based and simulated on the FBH transferred-substrate InP-DHBT technology combining single- and double-finger transistors with emitter width equal to 500 nm [22]. As derived within the Fig. 7.12a, the simulated TIA achieves a high transimpedance (>75 dBΩ) with a 3 dB bandwidth within the V-band (>75 GHz).

The calculated noise current spectral density of the TIA in Fig. 7.12b follows the typical trend described in the previous section. Finally, the circuit exhibits a gain of 34 dB and a high output saturation power of 10 dBm once it is terminated to a 50 Ω load at both ends at an RF frequency of 60 GHz. The performance and properties of this device is capable of efficiently amplifying the mm-waves generated by the uni-travelling carrier photodiodes.

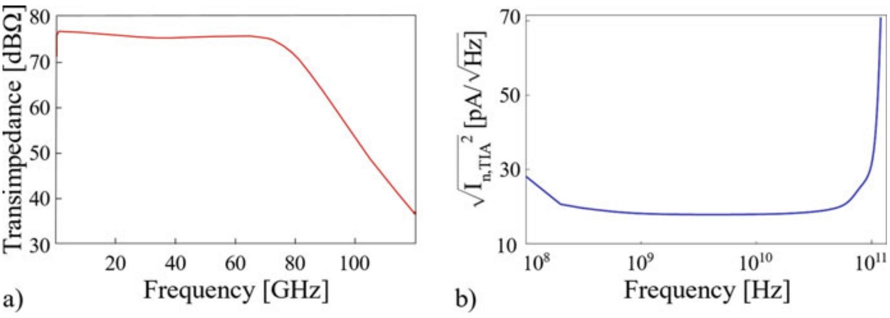


Fig. 7.12 (**a**) The transimpedance of the TIA as a function of frequency; (**b**) input current noise spectral density of the simulated TIA

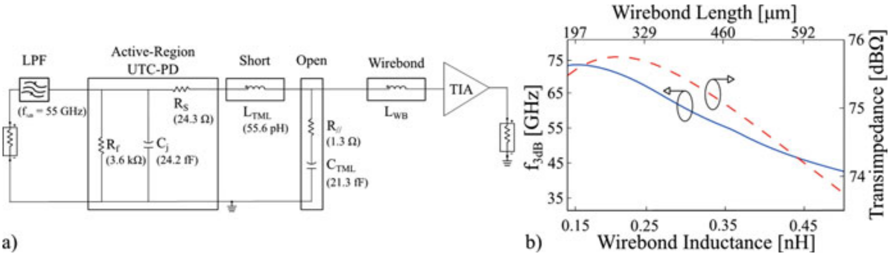


Fig. 7.13 (**a**) A 4 × 20 μm^2 UTC-PD interconnected with a TIA via a wirebond; (**b**) impact of wirebond length to the 3 dB bandwidth (blue) and transimpedance (red) of the total device

7.2.5 Co-simulation of the Interconnection Between UTC-PD and TIA

In Fig. 7.13a, the proposed TIA is co-simulated in ADS with the equivalent circuit of a measured 4 × 20 μm^2 waveguide UTC photodiode from III-V Lab (France) with a responsivity 0.79 A/W [23]. A low-pass filter (LPF) with a 3 dB bandwidth of 55 GHz is used in series to the circuit in order to simulate the transit time of the electrons within the device. This UTC-PD is interconnected with the TIA through a wirebond that is modeled as a lumped inductor.

In the blue curve of Fig. 7.13b, the inductance of the wirebond (L_{WB}) is swept over different values in order to analyze its impact on the f_{3dB} of the device. Also, in the upper x-axis of the diagram, the length of a straight gold wirebond is calculated with an assumed diameter of 17.78 μm [24]. With L_{WB} of 0.15 nH (or 197 μm), the maximum f_{3dB} is obtained at 73.1 GHz where the effect of gain peaking on the response of the component is observed [25]. Therefore, the poles originating from the DC-block and the parasitic capacitances of the UTC-PD that reduce the f_{3dB} are compensated by the L_{WB} [26]. Furthermore, the impact of the wirebond to the transimpedance exhibits a small variation (of approximately 2 dBΩ) over the swept wirebond inductances. Thus, once the UTC-PD and the multistage TIA are simulated as one module, the co-integration process does not limit its frequency response.

These simulation results are very promising providing substantial input on the synthesis of a low-cost and low-energy consumption hybrid optical-electronic component that combined with integrated antenna elements can be replicated and will eventually fulfil one of the requirements that 5G poses which is massive deployment of numerous base stations within urban areas.

7.3 Optical Receiver Design for UDWDM-PON

Passive optical networks (PONs) are most commonly associated with the fiber to the x (FTTx) applications, where x = home, curb, or building. Currently deployed FTTx PONs, based on time division multiplexing (TDM) equipment, use only one wavelength and separate users in time domain where each optical network unit (ONU) can only transmit and receive signal at a specific time slot, and this will influence the system latency and system capacity. To increase the capacity to meet the increasing bandwidth demand for 5G, advanced techniques combined with wavelength division multiplexing are being investigated. Starting from the concept of Dense WDM (DWDM), implementing 40 channels at 100 GHz spacing or 80 channels with 50 GHz spacing, it is possible to further reduce the channel spacing using suitable technologies capable of 12.5GHz or even 6.25 GHz of spacing, introducing the notion of ultradense WDM (UDWDM). Allowing the use of over 256 channels, UDWDM finally enables the "λ-to-the-User" concept. The main challenge for UDWDM-PON is to make the deployment costs affordable to access users. The research emphasis is placed on cost-effective solutions in modulation formats selection, receiver design, and impairments compensation. Receivers based on coherent detection are able to choose the required wavelength by the local oscillator (LO) so as to achieve "colorless" ONU. However, the traditional coherent receiver is too complexed for PON application due to its expensive optical and electronic components. Simplified digital coherent receiver design starts to play an important role in the UDWDM-PON filed. Except for coherent, some proposals also suggest the use of direct detection (see Fig. 7.14), but colored ONU will place pressure on the wavelength management. PON requires enough power budget to support more users so as to reduce system cost, but the signal usually is subject to severe distortions, such as fiber dispersion-induced inter-symbol interference (ISI) and system bandwidth limitation, which become worse with the increase in system capacity. Additionally, when a semiconductor optical amplifier (SOA) is employed as a booster or pre-amplifier to increase the power budget, the nonlinearities due to the gain saturation of SOA will significantly degrade the signal quality. In the following sections, these topics will be discussed in detail.

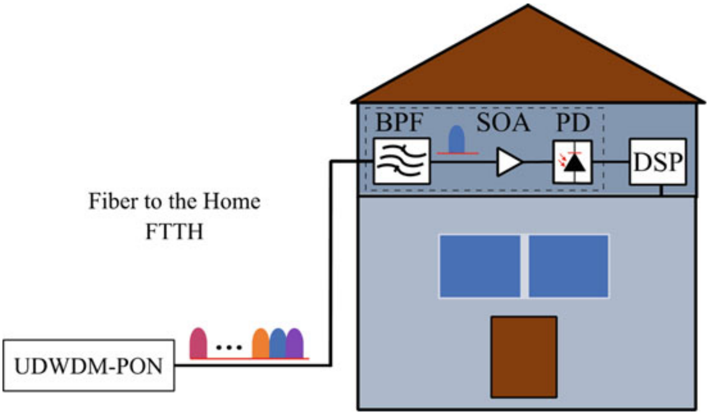


Fig. 7.14 A block diagram of UDWDM-PON for FTTH application

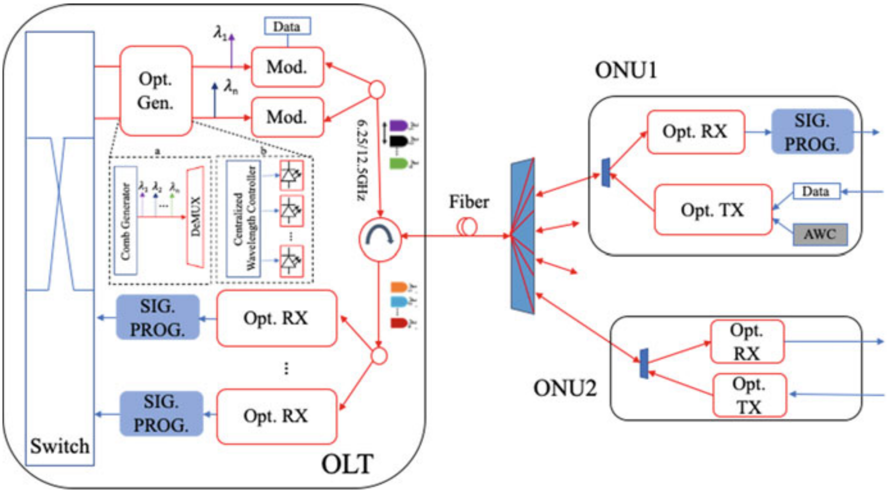


Fig. 7.15 System architecture of UDWDM-PON; inset: (**a**) Comb generator; (**b**) multiple lasers for optical carrier generation (*Opt Gen* Optical generator; *Mod* modulator; *SIG PROG* signal processing; *RX* receiver; *TX* transmitter; *AWC* automatic wavelength controller

7.3.1 Physical Architecture of UDWDM-PON

Figure 7.15 shows the basic physical architecture of UDWDM-PON system. The optical line terminals (OLTs) located in the central office are responsible for the data switching in the whole network as well as for the connection with the upper network layer.

However, the ONUs near to the user side will receive data and send users' request. In the downlink, the sending data is firstly converted form electrical to the

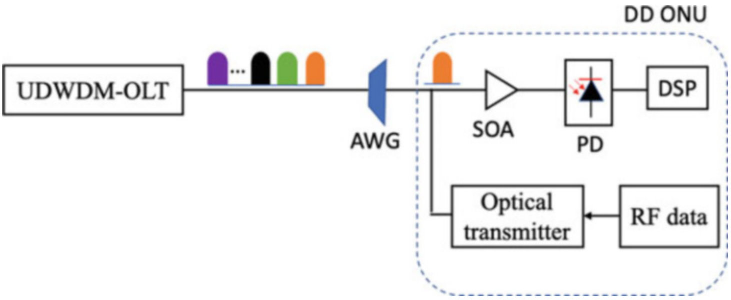


Fig. 7.16 A block diagram of DD-based ONU for UDWDM-PON

optical domain. The optical carriers are densely spaced and are initially generated either by using multiple laser sources or a comb source as show by inset (a) (b). For the multiple laser sources, a centralized wavelength locker combined with a high-performance thermoelectric cooler prohibiting the frequency drift of the sources [27] is employed. In terms of the comb source, it can simplify the tone generation, but an extra demultiplex needs to be added into the OLTs to separate the tones of the comb. The wavelength distance should follow the ITU-T recommendation G.694.1 which is lower to 6.25GHz and 12.5GHz [28]. The generated optical signal will be modulated with electrical signals. After wavelength multiplexing, the signal will travel through an optical circulator enabling bidirectional transmission in the fiber.

In the next step, a UDWDM splitter dictates the separation of signal to each ONU. Such a component can be an arrayed waveguide grating (AWG) separating the overall UDWDM signal to the specific ONU (see the configuration in Fig. 7.16).

In this case, a cost-effective receiver based on direct detection (DD) can be deployed. Since commercial avalanche photodiodes (APD) with high receiver sensitivity is costly and its bandwidth is much lower compared to photodiode (PD), SOA combined with PD is another option for a DD receiver which can keep a similar receiver sensitivity and sufficient bandwidth for high-speed transmission. The problem with the SOA is the nonlinear pattern effect due to the gain saturation; therefore, digital signal processing (DSP) will be employed to compensate this distortion; the detailed discussion will be presented in the following section.

If a coherent receiver is used in the ONU, then a simple passive splitter will first split the optical signal to all ONUs, and the LO in the RX will help detect the target wavelength (see the configuration in Fig. 7.17).

To simplify the structure of the ONU, some of the tones from the OLT can also feed the optical transmitters of the ONUs as the upstream carrier. Note that the traditional coherent receiver is too complicated for PON applications; hence some simplified coherent receiver designs are proposed which will be discussed in the following section. To demodulate the signal and compensate the system impairments, a DSP module will be implemented. For the uplink, the received information from the ONUs travels through the combiner and the circulator. Once more, the received signals will be demultiplexed, and then the data will be extracted

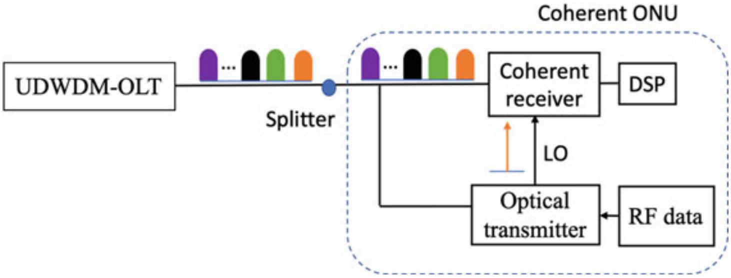


Fig. 7.17 A block diagram of coherent ONU for UDWDM-PON

from an optical receiver. Some signal processing module will help compensate the system impairments.

In order to achieve the power budget required to perform UDWDM, and also reduce the system cost as much as possible, there is a need to further improve the overall performance of some key parameters of the whole system such as the modulation format and receiver design. In the subsequent section, cost-effective receiver design and the modulation formats choices will be presented.

7.3.2 Receiver Design in the ONU: Direct Detection or Coherent

DD is a simple method which uses a single PD to achieve electrical signal recovery from the optical signal. During the past 20 years, direct detection has been employed as a common solution in the optical access network due to its cost-efficient and low-power consumption features. By using simple NRZ modulation formats, a maximum capacity up to 100 Gb/s can be achieved based on direct detection in a combination with time and wavelength-division multiplexed (TWDM) system [29, 30], which is currently considered a highly relevant topic by the IEEE 802.3ca task force. DD means the receiver can use a single PD or APD to detect signal and recover the amplitude of the signal. A demonstration shown in Fig. 7.18 [31] employs DD in the UDWDM system. An AWG filter with a channel spacing of 12.5 GHz at the remote node is used to distribute the wavelengths to the end users, "coloring" the whole network. Colored network means each ONU requires a specific wavelength which will require spectrum management and increase the system cost. Since PD typically has relative lower sensitivity compared to APD, other passive optical components used in the system will increase insertion loss, and thus additional optical amplifiers will be required to increase the system loss budget.

To enable flexible network operation by allowing colorless operation of the ONU without AWG or tunable optical filters in the ODN, receivers based on coherent

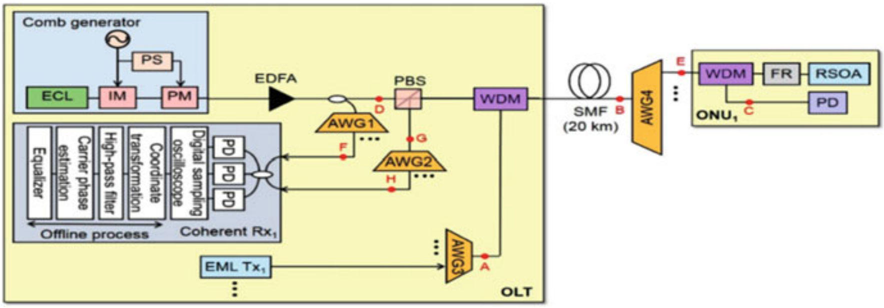


Fig. 7.18 Configuration of the proposed 12.5-Gb/S, 12.5-Ghz spaced UD-WDM PON based on DD [30]

rather than direct detection can be deployed, selecting a wavelength channel simply by tuning the LO laser to the wavelength of the downstream channel of interest. Fine wavelength selectivity in coherent WDM-PONs enables the use of (ultra-) dense wavelength spacing while avoiding the requirement for sophisticated optical wavelength filtering. Recent demonstrations of this technique include 10 Gbps/λ transmission in a 5 GHz grid [32] and 3.75 Gbps/λ and 1 Gbps/λ in a 2.5GHz grid [33, 34]. In addition to this key advantage, coherent receivers offer significantly higher receiver sensitivities in comparison to DD receivers [34, 35]. This will be a major advantage in future PON technologies, operating at multi-Gb/s per subscriber, and offering higher loss budgets, enabling higher split ratios (i.e., increased number of users) and longer reach. The high receiver sensitivity enables high-power budgets that can be shared arbitrarily between reach and split ratio depending on the network requirements. High split ratios reduce per-user costs since more end users can be supported in an access network using a single feeder fiber. Moreover, the optical amplitude, phase, and polarization can all be encoded with information, so it can allow a greater increase in the data rate without putting too much pressure on the electrical devices. Second, an optical coherent detection scheme is a linear detection where transmission impairments can be completely compensated by using DSP, so a higher receiver sensitivity can be obtained. In all, coherent detection is a way that can take advantage of the full potential of fiber transmission in a flexible way. Although coherent technology offers significant advantages, the complexity and high cost of conventional intradyne digital coherent receivers have prevented their use in PON applications. Thus, simplified coherent technology can play an important role in future access and mobile backhaul PONs.

It is acknowledged that if polarization-independent reception can be realized while avoiding the requirement for an optical polarization tracking unit in the coherent receiver, the complexity can be significantly reduced. To date, there are six reported low-complexity polarization-independent coherent receiver architectures employing various techniques; a comparison between the different designs is presented in [36]. The Alamouti receiver [37, 38] as depicted in Fig. 7.19 and Fig. 7.20 is a coherent receiver which detects an Alamouti polarization-time block-

Fig. 7.19 Alamouti intradyne receiver [36]

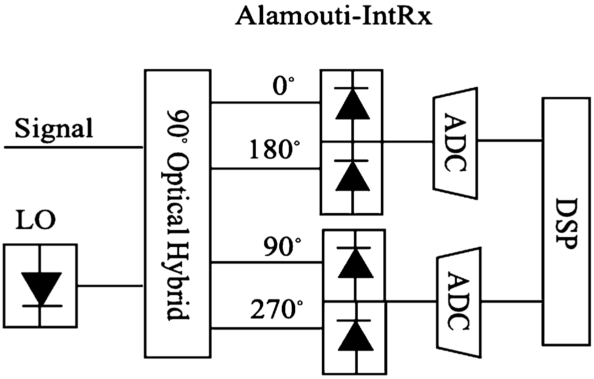
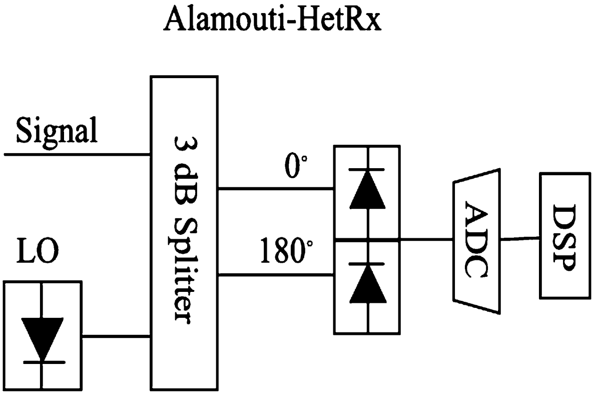


Fig. 7.20 Alamouti heterodyne receiver [37]



coded (PTBC) signal, requiring a dual-polarization modulator to avoid the need for an optical polarization tracking unit in the receiver. Although it introduces 50% redundancy due to the replication of the transmitted symbols, it leads to a significant simplification in the design compared to the conventional polarization- and phase-diverse intradyne (PPDI) coherent receiver.

In the 1980s, Glance has proposed a polarization-independent optical heterodyne receiver [39], referred to as Glance/Cano-HetRx, which has been demonstrated in a 1.25 Gb/s transmission system first [40] and, more recently, in real-time operation using 10 Gbps OOK signal [41]. The receiver consists of a 3 dB coupler and a polarization beam splitter (PBS) followed by two single-ended PDs, each detecting a polarization component, as shown in Fig. 7.21.

Following signal detection, the photocurrents at intermediate frequencies are first filtered, demodulated separately, and, finally, summed to obtain a baseband signal. Due to its capability of detecting two polarization modes, the detection process is independent of the polarization state of the received optical signal.

Moreover, Cano proposed in [42, 43] an alternative low-complexity coherent receiver design for use in ONUs, achieving polarization-independent detection employing a polarization scrambling (PS) method, in which every symbol is

Fig. 7.21 Glance/Cano-HetRx [38]

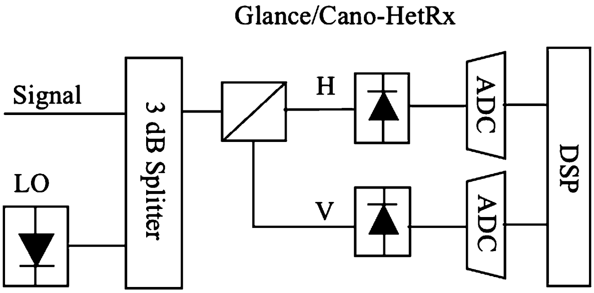


Fig. 7.22 Cano-IntRx [41]

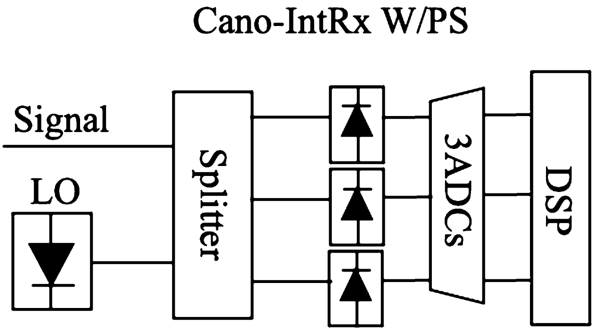
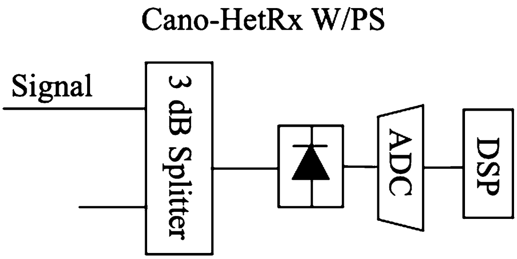


Fig. 7.23 Cano-HetRx [42]



transmitted twice, in orthogonal polarization states during two time slots. The Cano-IntRx consists of a symmetric 3 × 3 (1:1:1) coupler (using only two input ports) followed by three single-ended PDs and three ADCs, as illustrated in Fig. 7.22. In contrast, the Cano-HetRx has a simpler architecture which comprises a 3 dB coupler and a single-ended PD followed by a single ADC, as depicted in Fig. 7.23.

Ciaramella has proposed a simplified coherent receiver achieving polarization-independent reception in [44]. It employs a PBS and a symmetric 3 × 3 coupler (utilizing all three ports) followed by three single-ended PDs, as depicted in Fig. 7.24 [43]. The LO laser is separated into two orthogonal states of polarization (denoted as "H" and "V") using a PBS, and subsequently, they are mixed with the signal component. The output photocurrents are passed through the DC-blocks and then squared and summed to obtain the baseband signal. Finally, the signal is

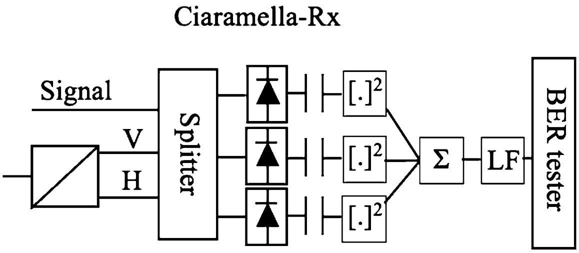


Fig. 7.24 Ciaramella Rx [43]

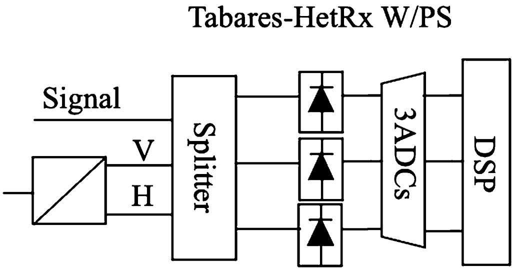


Fig. 7.25 Tabares HetRx [44]

low-pass filtered before being input to a clock and data recovery circuit. The key advantage of the Ciaramella-Rx is that it requires only simple analog processing, i.e., there is no need for an ADC or DSP. However, this receiver design is limited to amplitude-shift keying (ASK) (e.g., OOK or 4-PAM) signaling, and its tolerance to chromatic dispersion is lower than the other proposed receivers since the receiver linearity is lost due to squaring operation after the detection. A further disadvantage of this approach is that the receiver requires a large signal-LO frequency offset (0.9 × symbol rate (fb)) to avoid interference from low-frequency components of the directly detected signal. Therefore, the use of a single laser in the ONU, operating as both the upstream signal source and the downstream signal LO, is not possible.

To regain the phase diversity, Tabares [45] modified the Ciaramella-Rx by replacing the squaring operation with the linear combination of the three output photocurrents to remove the direct detection terms (identical to Cano-IntRx w/PS) while employing the same optical front-end design as the Ciaramella-Rx, as shown in Fig. 7.25 [44]. Although the linear operations to cancel direct detection terms can be performed in the analog domain, it is desirable for it to be performed digitally, requiring three ADCs to achieve high receiver sensitivity.

Simulations results are shown in Fig. 7.26a, b, where BER vs Received Power and bit error ratio (BER) vs photons per bit (PPB) curves are drawn for above proposed architectures and modulation formats. Additionally, Table 7.2 [36] presented a comparison between the proposed receiver, focused on spectrum efficiency, PPB, bit rate, and, of course, sensitivity.

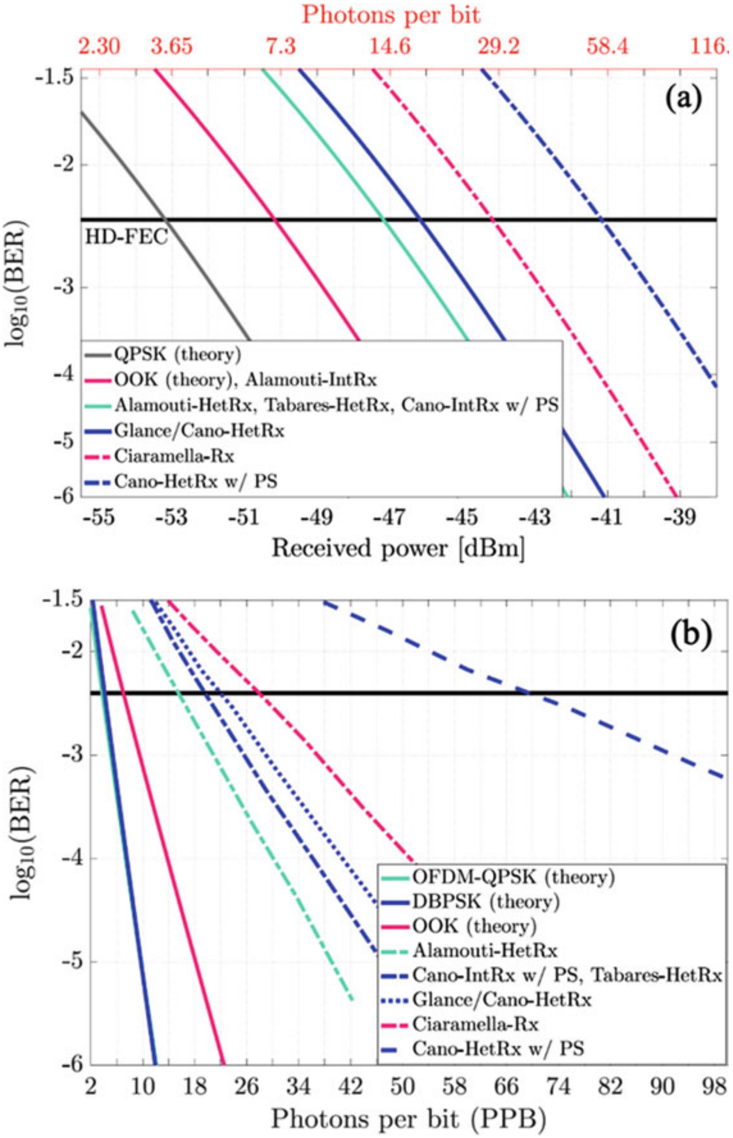


Fig. 7.26 (**a**) BER vs received power, (**b**) BER vs PPB

7.3.3 Modulation Format Choices for UDWDM-PON

It is well acknowledged that the complexity of conventional (dual-polarization digital) coherent receivers has so far prevented their introduction into access networks. Thus, low complexity in coherent receivers is needed. On the other

Table 7.2 Comparison of coherent receivers achieving polarization independent detection

Simplified coherent Rx	Modulation (SE [bis/s/Hz])	B2B req. PPB in sim.	Exp. sensitivity/bit rate [PPB]/[Gbps]	Distance [km]
Ciaramella-Rx	OOK	28	246.8/10	105
Tabares-HetRx	DBPSK	19.5	78.5/1.25	50
Cano-IntRx w/PS	DBPSK	19.5	78.5/1/25	50
Cano-HetRx w/PS	DBPSK	69.3	197.4/1.25	50
Glance/Cano-IntRx	DBPSK	22	123.6/10	25
Alamouti-HetRx	AC-OFDM QPSK	15.5	58/10.7	108
Alamouti-HetRx	AC-OFDM 16QAM	36.5	230/21.4	38

hand, using power-efficient modulation formats is not only very attractive in the challenged power budget of PONs but also one of the most efficient ways to achieve higher data rates, better spectrum efficiency, and thus improved receiver sensitivity.

Discrete multitone (DMT), NRZ, EDB, and M-PAM are widely used as modulation formats in the direct detection PON system. DMT is a multicarrier modulation technique that uses the discrete Fourier transform (DFT) to divide the channel into several sine-shaped orthogonal sub-channels, which might be a promising technology for next-generation PONs while still using low-bandwidth optical components. The principle of the DMT transmitter and receiver is shown in Fig. 7.27. Since DMT has the benefit of loading symbols on subcarriers with different modulation formats depending on the quality of the subcarrier, it is able to realize high spectral efficiency and is robust to spectral deficiencies. Additionally, the DMT time domain signals are real, permitting more cost-effective IMDD (intensity modulation direct detection). Another feature DMT provides is flexibility and increased chromatic dispersion (CD) tolerance. However, it was shown that its high linearity requirements make it hard to outperform EDB and PAM-4 at the same rate for the same receiver sensitivity.

In most optical communication systems, NRZ has been the reference choice for decades due to the inherent simplicity of the electrical components. However, while still considering the IMDD scenario which is desirable for cost-effective applications like PONs, more spectrally efficient modulation formats like EDB and PAM4 are available when multiple amplitude levels are used. A comparison between these multilevel formats can be shown in Table 7.3.

Once a coherent receiver is applied in the UDWDM system, several coherently detected modulation formats can be adopted, namely, dual-polarization quadrature phase-shift keying (DP-QPSK), polarization-switched (PSwitch) QPSK, 4- and 16-pulse position modulation (PPM), and 2- and PAM4 (pulse amplitude modulation).

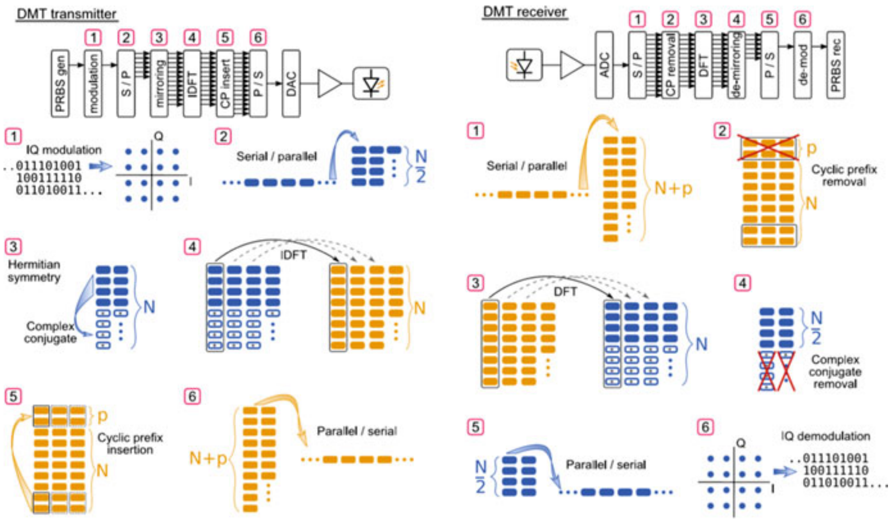


Fig. 7.27 DMT transceiver and receiver

Table 7.3 Comparison between modulation formats [46]

Modulation format	Sensitivity	Complexity	Required bandwidth	Dispersion tolerance
NRZ	√√√	√√√	√	√
NRZ-EDB	√√	√√	√√	√√
EDB	√	√√	√√√	√√
PAM4	√	√	√√√	√√√

A detailed comparison of these formats in terms of power efficiency is shown in [36]:

- PPM: If high sensitivity is the absolute primary requirement (neglecting the spectral efficiency), high-order M-PPM (e.g., $M \geq 16$) is a clear choice; however, it requires an M-fold increase in bandwidth compared to OOK (2-PAM).
- 4-PAM: On the other hand, 4-PAM offers double the information per symbol compared to OOK but requires approximately three times the number of PPBs.
- PSwitch-QPSK: When coherent detection using the conventional PPDI coherent receiver is considered, PSwitch-QPSK stands out as having the lowest required number of PPBs.
- DP-QPSK: PSwitch-QPSK requires 0.6 fewer PPBs than DP-QPSK at the expense of offering 25% lower spectral efficiency. Thus, a trade-off between sensitivity and optical/electrical bandwidth requirements for the targeted capacity needs to be evaluated.
- QPSK: if the polarization diversity is sacrificed to implement a low-complexity coherent receiver, the implementation of PSwitch-QPSK is not possible, whereas single-polarization QPSK might be a reasonable choice.

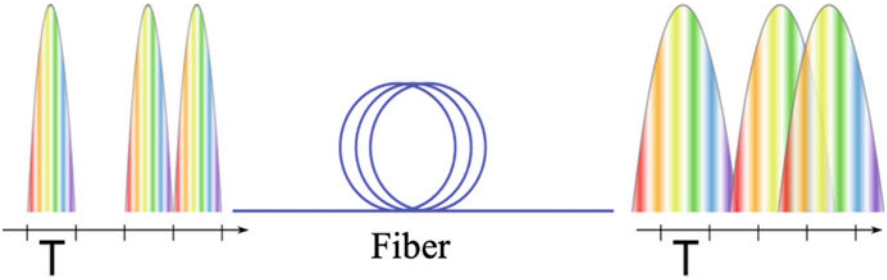


Fig. 7.28 Inter-symbol interference caused by pulse spreading over dispersive fiber

7.3.4 Optical Impairments and Mitigation Strategies in UDWDM-PON

The required increase of bit rate and reach of PONs poses new challenges on the physical layer design in order to meet the desired performance in a cost-effective way. Performance limitations in optical communication systems are due to both optoelectronic components and the optical fiber through which the signal propagates. In this section, the aspects most relevant to PONs are introduced and discussed, while factors that have a significant impact on higher capacity systems are omitted.

7.3.4.1 Chromatic Dispersion

In an optical communication system, dispersion is the dependence of the group velocity on the signal wavelength or on the propagation mode. Dispersion generally causes broadening in time of a light pulse as it propagates along the fiber as shown in Fig. 7.28. The spreading out of the pulse causes ISI (Inter Symbol Interference) which introduces distortions and limits the data rate and maximum length of the optical link.

Chromatic dispersion can be addressed with optical or electronic compensation at the receiver or at the transmitter. Optical compensation, which makes use of dispersion compensating fiber (DCF) or fiber Bragg gratings (FBGs) [47], is effective but requires bulky components and also has to be dimensioned to every specific link to enable optimum performance, making it unsuitable for a point to multipoint networks with an asymmetric physical structure. Besides, optical filtering [48] based on delay interferometer (DI) can compress the chirp of the modulator and then increase the tolerance of dispersion.

Electronic compensation, on the other hand, can provide reduced footprint and high flexibility and can harness the progress in complementary metal-oxide-semiconductor (CMOS)-based DSP technology with potential advantages both in terms of costs and scalability. For a direct detect receiver, feed-forward equalizer

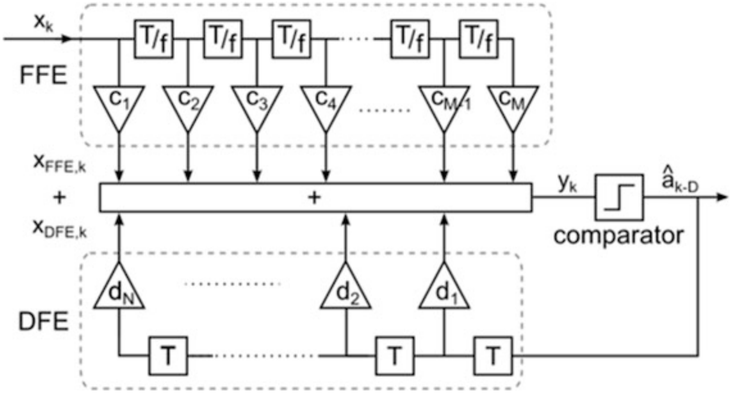


Fig. 7.29 Schematic of linear equalizer with FFE and DFE where delay blocks and tap coefficients are visible

(FFE) and decision feedback equalizer (DFE) are normal equalizers which use a tapped-delay line approach to compensate for ISI whose interference spreads over multiple symbols periods; a simple block diagram of the FFE and DFE is shown in Fig. 7.29.

7.3.4.2 Optical SOA Pattern Effect and Compensation with Neural Network

As mentioned above, the DD receiver can reduce the complexity of ONU. In order to improve the sensitivity of the DD receiver, a pre-amplifier is typically used before the PD. The associated gain saturation will degrade signal quality. Gain saturation is a phenomenon that originates from the dependence of the amplifier gain, G, on the input power of the input signal, $P_{S,in}$. When this value gets close to the amplifier saturation power, P_S, the gain will be reduced. The amplifier gain is defined as the ratio between the output and input power of the signal

$$G = \frac{P_{out}}{P_{S,in}} \tag{7.12}$$

The power of the signal propagating through the amplification medium can be expressed as

$$P(z) = P_{s,in}e^{gz} \tag{7.13}$$

where g is the material gain. After derivation, it can be expressed as

$$\frac{dP(z)}{dz} = gP(z) \tag{7.14}$$

The material gain, g, can be expressed as a function of P_s and the small signal gain g_O, [7]

$$g = \frac{g_0}{1 + \frac{P}{Ps}} \tag{7.15}$$

and hence

$$\frac{dP(z)}{dz} = \frac{g_0 P(z)}{1 + \frac{P}{P_s}} \tag{7.16}$$

By integrating this equation over the amplifier length, the following relation for the large-signal amplifier gain is obtained:

$$G = G_0 e^{\left(-\frac{G-1}{G} * \frac{P_{out}}{Ps}\right)} \tag{7.17}$$

The equation shows that the amplification factor G decreases from its unsaturated value G_0 when P_{out} becomes comparable to P_s. A useful metric commonly used is the output saturation power, which is defined as the output power for which the amplifier gain is reduced by a factor of 2, or 3 dB, from its unsaturated value G_0.

SOAs suffer from carrier depletion-induced saturation. When working in the saturation regime, the SOA becomes nonlinear causing a number of effects to occur including SPM, XPM, FWM, cross-gain modulation, and self-gain modulation. When an SOA amplifies a high-intensity modulated signal, self-gain modulation can lead to a serious waveform distortion commonly referred to as patterning effect (see the eye diagram performance in Fig. 7.30). As a kind of nonlinearity impairment induced by gain saturation, pattern-dependent effect can be mitigated by the nonlinearity compensation DSP at the receiver end.

In recent years, neural network (NN) has become very popular and has shown effective application in the fields of optical performance monitoring and nonlinearity equalization. It can be proved that an NN can fit and express any function if it has at least one hidden layer and enough hidden nodes [50]. Therefore, it is possible to use NN to compensate the nonlinearity from SOA. An experiment demonstration is conducted to verify the effectiveness of NNs – the setup is shown in Fig. 7.31. At the transmitter side, the 25 Gbaud PAM-4 downstream signal is generated by a digital-to-analog converter (DAC) working at 65 GSa/s with the 3 dB analog bandwidth of 20 GHz. A commercial 18Gbps DML (Directly Modulated Laser) with 3 dB bandwidth of 10 GHz and a center wavelength of 1310 nm is employed. The DAC output is first amplified by a 23 dB electrical amplifier (EA) before driving the DML. Then, the 25 Gbaud PAM-4 optical signals are transmitted over up to 20 km SSMF with an average loss of 0.33 dB/km at 1310 nm. In order to support high

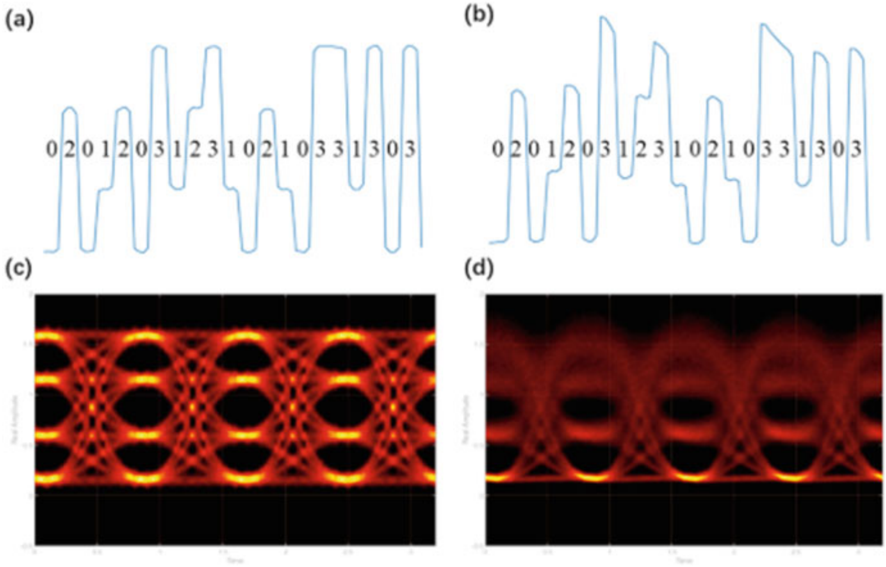


Fig. 7.30 The pattern effect of PAM-4 signal: (**a**) pulse shape and (**b**) eye diagram of the PAM4 signals without pattern effect; pulse shape (**c**) and eye diagram (**d**) of the PAM4 signals suffered from SOA-induced pattern effect [49]

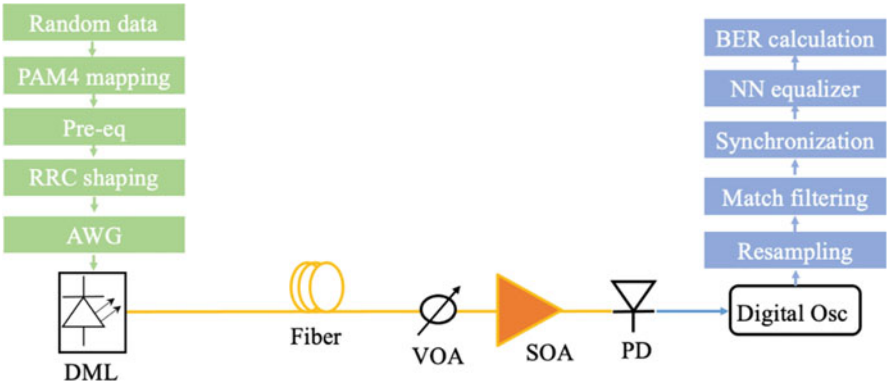


Fig. 7.31 The experimental setup of the NN based IM/DD system at O-band

link loss budget, an O-band SOA preamplifier is used at the receiver side before direct detection. A variable optical attenuator (VOA) is applied to emulate splitter loss and also test the SOA performance. A 50 GHz PD is chosen for this experiment to detect the signal. Then the output signal is captured by an 80 GSa/s oscilloscope with 20 GHz bandwidth and processed by offline DSP.

The DSP block at the transmitter sider (green flow) is shown in Fig. 7.31.

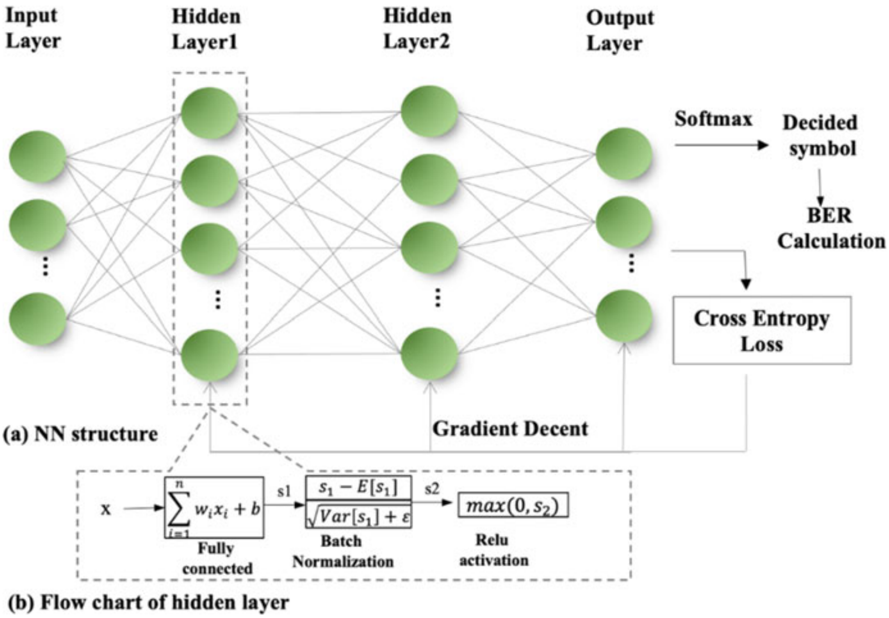


Fig. 7.32 The NN structure (**a**) and connection flow between hidden layer (**b**)

Random data generated by a Mersenne Twister are mapped into PAM-4 symbols with a length of 2^{15} and then upsampled to two times the data rate. The system performance with high baud rate will be seriously limited by the bandwidth limitation of the optoelectronic devices, such as the DAC, DML, and PD. The filtering effect from them can result in inter-symbol interference (ISI) and restrict the baseband signal bandwidth. Time-domain digital pre-equalization (pre-eq) and root-raised-cosine (RRC) filter with the optimal roll-off factor 0.15 are used to reduce the ISI caused by the bandwidth limitation. Next, the symbol sequence is resampled to match the sampling rate of the DAC.

The offline DSP block at the receiver side is shown by the blue flow. The captured offline data is first resampled to 1-sps with the matched filter, and then synchronization is applied to remove the time jitter and extract the sending data. Afterward, a full connected NN equalizer is used to reduce the linear and nonlinear impairments. The structure of the NN-based equalizer we apply in this work is shown in Fig. 7.32.

It is a four-layer network, containing two hidden layers. The circles denote the nodes, also known as neurons. It is a fully connected network, in which the input layer has 57 nodes; hence, the network requires 57 consecutive sampled symbols as input for a judged symbol. Each hidden layer has 128 nodes, while the output layer has 4 nodes, corresponding to the 4 symbols of PAM4 mode. The nonlinear activation function of the hidden layers is rectified linear unit (ReLU) used for nonlinear transformations, which is expressed as follows:

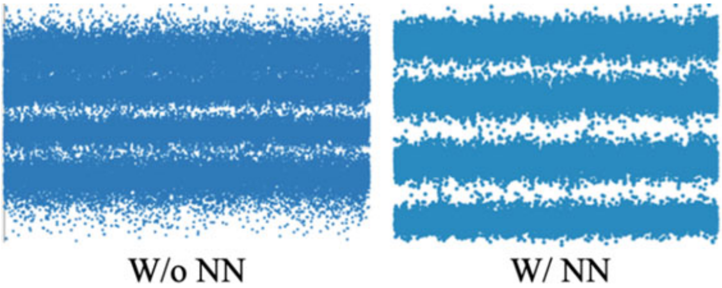


Fig. 7.33 Power-level distribution of received time domain signal with and without NN

$$y = \max(0, x) \tag{7.18}$$

The activation of output layer is softmax which is able to enforce the features to the range of (0, 1). The softmax function is expressed as follows:

$$y_j = \frac{e^{x_i}}{\sum_{j=1}^{N_{class}} e^{x_j}} \tag{7.19}$$

where x and y are input and output feature maps of the softmax function, i and j are the indexes of the neurons, N_{class} is the number of classes which is set as 4 considering the four levels of PAM4 signal, and y_j is the predicted probability of an observation belonging to jth class. The length of the whole data set is 100,000, 50,000 symbols for training and 50,000 symbols for the final test. The final performance is evaluated based on the BER of the test data set.

As mentioned earlier, the higher the input power to the SOA, the pattern effect due to the gain saturation will be more serious. We first set the input power to SOA as −1 dBm and compare the signal quality before and after NN equalization. The power-level distribution of the received time domain PAM4 signal is shown in Fig. 7.33.

The pattern effect will cause jitters for PAM symbols with higher power. Thus, the signals influenced by the pattern effect will be asymmetrical in the time domain. The third and fourth levels of the PAM-4 signals are overlapped. After employing the NN equalizer, the overlap induced distortion will be improved. We also compare the performance between FFE and NNs. As we can see from the result in Fig. 7.34, with the increase in input power, the BER will be improved first due to the increased signal-to-noise ratio (SNR).

However, once the power goes beyond −10 dBm, the FFE is not able to further improve the BER performance due to the nonlinearity introduced from the SOA. On the contrary, the NN-based equalizer can significantly improve the BER; therefore, we conclude that the SOA-induced pattern effect can be well addressed by NN.

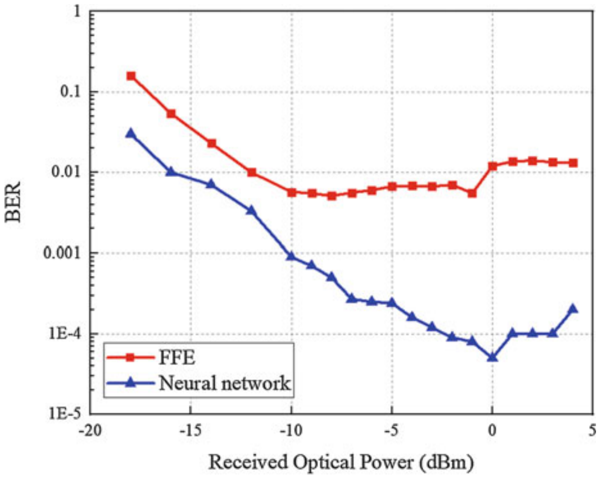


Fig. 7.34 BER versus the input power into the SOA of 50G PAM4 signal transmission over 20 km SMF at O-band

7.4 Conclusion

This chapter provided new insights into the design of future emerging 5G base stations and FTTH terminals that incorporate UDWDM-PONs for enabling high-capacity backhauling.

The physical properties of UTC-PDs are explained, and an equivalent circuit model of the uni-travelling carrier photodiodes is obtained based on reflection coefficient parameter extraction. In addition, a multistage TIA design simulated on InP/InGaAs HBT technology is reported operating within the V-band and exhibiting high gain and high output powers. A co-simulation of these two devices is conducted synthesizing an optoelectronic mm-wave transmitter capable of generating and amplifying 5G NR signals. Finally, the wirebond between a simulated UTC-PD and the TIA does not significantly degrade the bandwidth of the device, making the proposed co-integrated mm-wave transmitter an ideal choice for the generation of mm-wave signals within the future 5G base stations interconnected with UDWDM-PONs.

For FTTH applications, two types of optical receivers based on coherent and direct detection, respectively, are introduced. PON is sensitive to the cost; therefore, the complexity and high cost of conventional intradyne digital coherent receivers have prevented their use in PON applications. Thus, a variety of simplified coherent receiver designs are introduced. On one hand, the combination of SOA used as pre-amplifier with PDs is proposed as a low-cost solution for the receiver in a DD ONU system.

In practical PON systems, different ONUs are dimensioned with different receiving powers to cater for the diverse distances to the OLT. Varied optical power into the SOA will induce nonlinear distortions to the signal. Therefore, NN is proposed at the receiver side to mitigate SOA nonlinearities generated within the optical signals. The performance of NN is experimentally investigated in an IM/DD system with 50G PAM4 signals. After optimizing the convergence factor of the NN, the SOA nonlinearities can potentially be compensated leading to improved receiver dynamic range, making the proposed SOA/PD package an alternative for the future ONU design in the UDWDM-PONs system.

Acknowledgments This work was partially funded by EU through the projects 5G STEP FWD (no. 722429), 5G-PHOS (no. 761989), TERAWAY (no. 871668), and blueSPACE (no. 762055).

References

1. Chin, W. H., Fan, Z., & Haines, R. (2014). Emerging technologies and research challenges for 5G wireless networks. *IEEE Wireless Communications, 21*(2), 106–112.
2. Konstantinou, D., et al. (2020). 5G RAN architecture based on analog radio-over-fiber fronthaul over UDWDM-PON and phased array fed reflector antennas. *Optics Communications, 454*, 124464.
3. Konstantinou, D., Morales, A., Rommel, S., Raddo, T. R., Johannsen, U., & Monroy, I. T. (2019). Analog radio over fiber fronthaul for high bandwidth 5g millimeter-wave carrier aggregated OFDM. In *International conference on transparent optical networks*, 2019, Vol. 2019-July.
4. Rommel, S., Olmos, J. J. V., & Monroy, I. T. (2017). 15Gbit/s duobinary transmission over a W-band radio-over-fiber link. In *Proceedings* – 2016 *advances in wireless and optical communications,* RTUWO *2016*, 2017, pp. 197–200.
5. Rommel, S., Morales, A., Konstantinou, D., Raddo, T. R., & Monroy, I. T. (2018). MM-wave and THz analog radio-over-fiber for 5G, wireless communications and sensing. In *Optics InfoBase conference papers*, 2018, Vol. Part F123-.
6. Johannsen, U., et al. (2019). ARoF-Fed antenna architectures for 5G networks. In 2019 *Optical fiber communications conference and exhibition, OFC 2019 – Proceedings*.
7. Agrawal, G. P. (2011). *Fiber-optic communication systems* (4th ed.).
8. Erman, M., et al. (1991). Optical circuits and integrated detectors. *IEE Proceedings. Part J, Optoelectron, 138*(2), 101–108.
9. Xu, Z., & Gao, J. (2017). Semi-analytical small signal parameter extraction method for PIN photodiode. *IET Optoelectronics, 11*(3), 103–107.
10. Maloney, T. J., & Frey, J. (1977). Transient and steady-state electron transport properties of GaAs and InP. *Journal of Applied Physics, 48*(2), 781–787.
11. Othman, M. A., Taib, S. N., Husain, M. N., & Napiah, Z. A. F. M. (2014). Reviews on avalanche photodiode for optical communication technology. *ARPN Journal of Engineering and Applied Sciences, 9*(1), 35–44.
12. Shimizu, N., Miyamoto, Y., & Ishibashi, T. (1999). Uni-traveling-carrier photodiodes. *Conference Proceedings – Lasers and Electro-Optics Society Annual Meeting-LEOS, 2*, 808–809.
13. Ito, H., Nagatsuma, T., & Ishibashi, T. (2007). Uni-traveling-carrier photodiodes for high-speed detection and broadband sensing. *Quantum Sensing and Nanophotonic Devices IV, 6479*, 64790X.
14. Konstantinou, D., Caillaud, C., Rommel, S., Johannsen, U., & Tafur Monroy, I. (2020). Investigation of de-embedding techniques applied on uni-traveling carrier photodiodes. In *EuMiC*.
15. "Advanced Design System (ADS), ADS 2020.02, Keysight EEsof EDA, Keysight Technologies."
16. Razavi, B. (2008). Fundamentals of microelectronics. *Change, 13*, 1–4.

17. Paasschens, J. C. J., Havens, R. J., & Tiemeijer, L. F. (2003). Modelling the correlation in the high-frequency noise of (hetero-junction) bipolar transistors using charge-partitioning. In *Proceedings of the IEEE Bipolar/BiCMOS circuits and technology meeting*, 2003, pp. 221–224.
18. Meyer, R. G., & Blauschild, R. A. (1986). A wide-band low-noise monolithic transimpedance amplifier. *IEEE Journal of Solid-State Circuits, 21*(4), 530–533.
19. Tagami, H., et al. (2005). A 3-bit soft-decision IC for powerful forward error correction in 10-Gb/s optical communication systems. *IEEE Journal of Solid-State Circuits, 40*(8), 1695–1703.
20. Säckinger, E. (2005). Broadband circuits for optical fiber communication.
21. Rein, H. M., & Möller, M. (1996). Design considerations for very-high-speed Si-bipolar IC's operating up to 50 Gb/s. *IEEE Journal of Solid-State Circuits, 31*(8), 1076–1090.
22. Shivan, T., et al. (2019). A 175 GHz bandwidth high linearity distributed amplifier in 500 nm InP DHBT technology. In *IEEE MTT-S international microwave symposium digest*, 2019, Vol. 2019-June, pp. 1253–1256.
23. Achouche, M., et al. (2004). High performance evanescent edge coupled waveguide unitraveling-carrier photodiodes for >40-Gb/s optical receivers. *IEEE Photonics Technology Letters, 16*(2), 584–586.
24. Qi, X., et al. (1998). Fast 3D modeling approach to parasitics extraction of bonding wires for RF circuits. In *Technical digest – International electron devices meeting*, pp. 299–302.
25. Konstantinou, D. (2020). Simulation of an integrated UTC-photodiode with a high-speed TIA for 5G mm-wave generation. In *NUSOD*.
26. Chen, G., Yu, Y., Deng, S., Liu, L., & Zhang, X. (2015). Bandwidth improvement for germanium photodetector using wire bonding technology. *Optics Express, 23*(20), 25700.
27. Pachnicke, S. et al. (2014). First demonstration of a full C-band tunable WDM-PON system with novel high-temperature DS-DBR lasers. In *Optics InfoBase conference papers*.
28. International Telecommunication Union – ITU-T, "Rec. ITU-T G.694.1 (02/2012)," 2012.
29. Ji, H., et al. (2017, May). Field demonstration of a real-time 100-Gb/s PON based on 10G-class optical devices. *Journal of Lightwave Technology, 35*(10), 1914–1921.
30. Xue, L., Yi, L., Ji, H., Li, P., & Hu, W. (Jan. 2018). Symmetric 100-Gb/s TWDM-PON in O-band based on 10G-class optical devices enabled by dispersion-supported equalization. *Journal of Lightwave Technology, 36*(2), 580–586.
31. Shim, H. K., Kim, H., & Chung, Y. C. (2014). Practical 125-Gb/s, 125-GHz spaced ultra-dense WDM PON. *Optics Express, 22*(23), 29037.
32. Luo, M. et al. (2018). Demonstration of 10-Gb/s, 5-GHz spaced coherent UDWDM-PON with Digital signal processing in real-time. In *Optics InfoBase conference papers*, 2018, vol. Part F84-O.
33. Ferreira, R. M., Shahpari, A., Reis, J. D., & Teixeira, A. L. (2017). Coherent UDWDM-PON with dual-polarization transceivers in real-time. *IEEE Photonics Technology Letters, 29*(11), 909–912.
34. Smolorz, S., Gottwald, E., Rohde, H., Smith, D., & Poustie, A. (2011). Demonstration of a coherent UDWDM-PON with real-time processing. In *Optics InfoBase conference papers*.
35. Shahpari, A., et al. (2017). Coherent access: A review. *Journal of Lightwave Technology, 35*(4), 1050–1058.
36. Erkilinc, M. S., et al. (2018). Comparison of low complexity coherent receivers for UDWDM-PONs (λ-to-the-user). *Journal of Lightwave Technology, 36*(16), 3453–3464.
37. Shieh, W., Yi, X., Ma, Y., & Yang, Q. (2008). Coherent optical OFDM: Has its time come? [Invited]. *Journal of Optical Networking, 7*(3), 234–255.
38. Erkilinç, M. S., et al. (2016). Polarization-insensitive single-balanced photodiode coherent receiver for long-reach WDM-PONs. *Journal of Lightwave Technology, 34*(8), 2034–2041.
39. Glance, B. (1987). Polarization independent coherent optical receiver. *Journal of Lightwave Technology, 5*(2), 274–276.
40. Cano, I. N., Lerín, A., Polo, V., & Prat, J. (2014). Simplified polarization diversity heterodyne receiver for 1.25Gb/s cost-effective udWDM-PON. In *Optics InfoBase conference papers*.

41. Altabas, J. A., et al. (2018). Real-time 10Gbps polarization independent quasicoherent receiver for NG-PON2 access networks. In *Optics InfoBase conference papers*, 2018, Vol. Part F84-O.
42. Cano, I. N., Lerin, A., Polo, V., & Prat, J. (2015). Flexible D(Q)PSK 1.25-5 Gb/s UDWDM-PON with directly modulated DFBs and centralized polarization scrambling. In *European conference on optical communication, ECOC*, 2015, Vol. 2015-Novem.
43. Cano, I. N., Lerín, A., Polo, V., & Prat, J. (2014). Polarization independent single-PD coherent ONU receiver with centralized scrambling in udWDM-PONs. In *European Conference on Optical Communication, ECOC*, 2014.
44. Ciaramella, E. (2014). Polarization-independent receivers for low-cost coherent OOK systems. *IEEE Photonics Technology Letters, 26*(6), 548–551.
45. Tabares, J., Polo, V., & Prat, J. (2017). Polarization-independent heterodyne DPSK receiver based on 3×3 coupler for cost-effective udWDM-PON. In *Optics InfoBase conference papers*, 2017, Vol. Part F40-O.
46. Dalla Santa, M. (2019). *Next generation technologies for 100 Gb/s PON systems*. University College Cork.
47. Yi, L., Wang, X., Li, Z., Huang, J., Han, J., & Hu, W. (2015). Upstream dispersion management supporting 100 km differential reach in TWDM-PON. *Optics Express, 23*(6), 7971.
48. Li, Z., et al. (2014). Symmetric 40-Gb/s, 100-km passive reach TWDM-PON with 53-dB loss budget. *Journal of Lightwave Technology, 32*(21), 3389–3396.
49. Wang, K., Zhang, J., Zhao, L., Li, X., & Yu, J. (2020). Mitigation of pattern-dependent effect in SOA at O-band by using DSP. *Journal of Lightwave Technology, 38*(3), 590–597.
50. Nielsen, M. A. (2015). *Neural networks and deep learning*. Determination Press.

Chapter 8
Modulation and Equalization Techniques for mmWave ARoF

Javier Pérez Santacruz, Umar Farooq, Kebede Tesema Atra, Simon Rommel, Antonio Jurado-Navas, Idelfonso Tafur Monroy, Amalia Miliou, Giancarlo Cerulo, Jean-Guy Provost, and Karim Mekhazni

Abstract Fifth generation (5G) is the emerging mobile communications platform that aims to meet the market requirements in terms of enhanced broadband connectivity based on harnessing small cell and mmWave technology. These two in synergy will provide high capacity gain not only through the hyperdense deployment of small cell but also through accessing large swathes of untapped spectrum at mmWave frequencies. The envisaged architecture entails an integrated optical wireless network architecture, where optical technology will complement radio in order to handle the new demands on capacity over the backhaul and fronthaul network, leading to the notion of analog radio over fiber (ARoF). The goal of this chapter is to provide novel approaches to optimize the performances of mmWave ARoF systems that includes developing enabling technology from a digital to signal processing (DSP) and device perspective.

8.1 Introduction

The fifth generation (5G) of mobile networks is the new solution to cater for emerging market requirements that aims to deliver in terms of flexibility, cost, power

J. P. Santacruz (✉) · S. Rommel · I. T. Monroy
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: j.perez.santacruz@tue.nl

U. Farooq · A. Miliou
Aristotle University of Thessaloniki, Thessaloniki, Greece
e-mail: umfarooq@csd.auth.gr

K. T. Atra · G. Cerulo · J.-G. Provost · K. Mekhazni
III-V Lab (a joint laboratory between Nokia Bell Labs, Thales R&T and CEA Leti), Palaiseau, France
e-mail: kebede.atra@3-5lab.fr

A. Jurado-Navas
University of Málaga, Málaga, Spain
e-mail: navas@ic.uma.es

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_8

consumption, latency, bit rate, reliability, and coverage. A centralized radio access network (C-RAN) is defined for 5G that can deliver enhanced received signal quality through centralized processing, with the potential to deliver reduced service latency, and greater network flexibility to the mobile operators.

This had led to a new disruptive 5G architecture that entails an integrated optical wireless network architecture, where optical technology will complement radio in order to handle the new demands on capacity over the backhaul and fronthaul network, leading to the notion of analog radio over fiber (ARoF). ARoF implies attractive benefits such as wide area coverage, high spectral efficiency, simple receivers, reduced power consumption, and low latency. Nevertheless, the combination of these technologies implies new challenges, such as high free-space path loss (FSPL), chromatic dispersion, phase noise, and combined mmWave radio and optical channel impairments, that will affect radio detection performance.

This chapter aims to study and analyze techniques to reduce the degradation introduced by the mmWave ARoF channel by revisiting the signal processing in the radio-optical transceiver link. The right modulation format selection is key to achieve higher performance in any communication system. The analysis of modulation formats in optical and wireless channels has been well studied independently. However, modulation formats have not been analyzed comprehensively in mmWave ARoF scenarios, where optical and wireless channels are joined. In this context, Sect. 8.2 compares experimentally the main radio modulation candidate formats in a mmWave ARoF setup and provides new insights into the best modulation options that suggests that legacy OFDM modulation might not tick all the boxes as it once did. Moreover, in such as a converged system, the analog radio signal will be subject to chromatic dispersion in the standard single-mode fiber (SSMF), spurring the need for equalization to compensate dispersion in the fiber. Section 8.3 studies channel equalization at the radio receiver based on a simulated mmWave ARoF scenario and provides new insights in performance. Optical amplifiers and modulators are crucial devices in mmWave ARoF systems. The REAM (reflective electroabsorption modulator)-SOA (semiconductor optical amplifier) integrated into a single chip is investigated as an alternative to the directly modulated lasers (DMLs) in the optical link, where EAM-based transmitters have the potential to provide better transmission performances because of the absence of adiabatic chirp. On the one hand, the SOA is sought to increase signal propagation distances to envisage the 5G coverage requirements. At the same time, the transceiver module of a 5G fronthaul must be able to operate at any wavelength of the WDM (wavelength-division multiplexing) system by being either wavelength-tunable or wavelength-independent (colorless). In this context, Sect. 8.4 investigates the device and optical link performance in terms of key parameters such as extinction ratio, insertion losses, and gain. In particular, an experimental digital transmission is demonstrated by utilizing this device, achieving a bit rate of 50 Gb/s. Finally, Sect. 8.5 provides the chapter conclusions.

8.2 Feasibility Study on New Modulation Formats for mmWave ARoF

The right modulation format selection is key to achieve higher performance in any communication system. The analysis of modulation formats in optical and wireless channels has been well studied independently. However, modulation formats have not been analyzed thoroughly in mmWave ARoF scenarios, where optical and wireless channels are joined.

Since mmWave ARoF combine wireless and optical channels, the selection of the modulation format becomes more crucial [1]. Orthogonal frequency-division multiplexing (OFDM) is the selected modulation format by 3rd Generation Partnership Project (3GPP) in the first 5G standard. However, it is not clear that OFDM is the best choice for mmWave ARoF systems. Therefore, in this section, a comprehensive comparison and analysis of the main modulation formats for 5G is realized in a mmWave ARoF system.

This section is structured as follows: first, a 5G architecture based on mmWave cells over an ARoF layer is explained in Sect. 8.2.1; next, the examined modulation formats are described in Sect. 8.2.2, highlighting its advantages and disadvantages; finally, the experimental setup and results of the modulation format comparisons are shown in Sect. 8.2.3.

8.2.1 ARoF Architecture for 5G Fronthauling

5G aims to support many types of services with different requirements and needs. These services are classified by 5G into three scenarios [2]: enhanced mobile broadband (eMBB), where the main goal is to achieve high bit rate; ultra-reliable low-latency communications (URLC), where latency and reliability are the main requirements; and massive machine-type communications (mMTC) supporting a huge quantity of connected devices. Video streaming is the most common user caser for eMBB, while autonomous driving and smart cities are clear examples of URLC and mMTC scenarios, respectively.

Network slicing is one of the most suitable strategies to manage and adapt the resources of the network according to the service requirements. Network slicing is based on software-defined networking (SDN) and network function virtualization (NFV), allowing to slice the physical network into several logical networks and, then, provide resources for a distinct application scenario [3].

As mentioned before, exploiting the mmWave domain is necessary to achieve the 5G requirements in terms of bit rate. However, the use of mmWave frequencies implies high free-space path loss (FSPL), and, thus, the cell radius is reduced into ranges of 10–200 m [1]. Hence, mmWave cells lead to a huge increase in the number of cells in contrast to the sub-6 GHz cells, to cover the same surface. Moreover, 5G

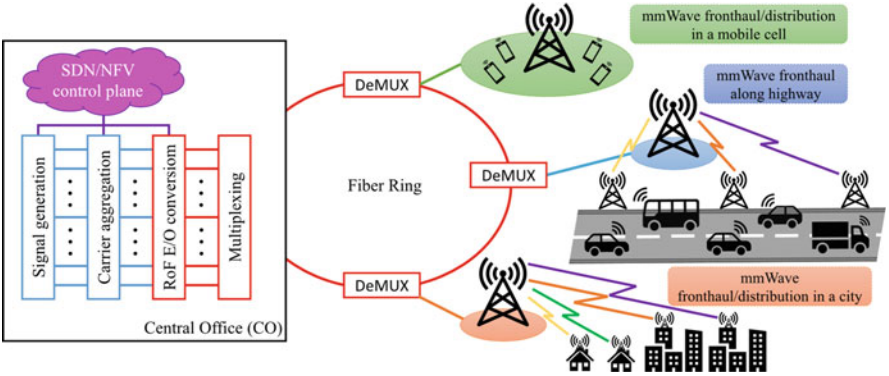


Fig. 8.1 ARoF fronthauling architecture employing mmWave cells for different applications

scenario exhibits a heterogeneous network where the mmWave microcells and the sub-6 GHz macrocells coexist and cooperate together.

The complexity of the remote unit (RU) is much less in ARoF system than in the current common public radio interface (CPRI) [4]. Therefore, an ARoF solution is a suited technology to give support to the enormous quantity of mmWave cells. Furthermore, C-RAN is a preferred option in terms of flexibility, latency, and energy consumption since most of the processes can be performed from a central office (CO) [5].

Therefore, due to the inherent benefits of the individual components, mmWave cells over C-RAN ARoF transport layer in synergy play a pivotal role in the 5G architecture. However, with several challenges still to solve, the ARoF will continue to be at the forefront of 5G research as we head toward the beyond 5G era.

Figure 8.1 shows the scheme of a C-RAN ARoF architecture with mmWave cells for the 5G fronthauling. In this system, most of the processes are arranged in CO. First, the electrical signals are generated in multiple streams. Then, the radiofrequency (RF) carrier is aggregated to each signal. Next, the resulted signals are converted to the optical domain. These three processes are managed and monotonized by a SDN/NFV control plane in order to perform the network slicing. Last, the multiple optical signals are multiplexed into an optical fiber ring. This multiplexing process can be arranged by different technologies: wavelength-division multiplexing (WDM), spatial-division multiplexing (SDM), or mode-division multiplexing (MDM). For instance, using WDM technology, each mmWave cell is located in a single wavelength.

The optical fiber ring achieves the different mmWave cells through demultiplexing access points. After the demultiplexing process, the optical signal reaches the RU of its corresponded mmWave cell. In the RU, the optical signal is converted to the electrical domain with the desired mmWave carrier frequency. Then, the mmWave wireless link is established between the RU and receiver point.

There are depicted several examples of mmWave cell use-cases in Fig. 8.1. The first use-case consists of a typical mobile cell where the end users correspond to mobile phones. The mobile phones can request resources for many types of services such as video streaming, augmented reality, video calls, etc. Another example of a mmWave cell user is the fronthauling along a highway. In this example, the mmWave cell supports a certain number of microcells that are distributed along the highway. Applications, such as autonomous driving, can be achieved through this deployment. The last example is based on a point-to-point communication. In this case, the RU is connected to several buildings in a city. In this context, smart houses with Internet of things (IoT) solutions can be supported in this communication system. Moreover, sub-6 GHz cells can also be supported through the C-RAN ARoF solutions. In this way, the 5G heterogeneous network can be managed by a single technology. Thus, the remote processes can be simplified, and the cell cooperation management can be optimized.

8.2.2 Proposed Modulation Formats Under Test

Modulation formats study is essential to enhance the performance of any communication system. Each communication system requires determined specification for the employed modulation format. For example, the wireless communication requires high robustness to frequency-selective channels due to the multipath effect in this type of communications. Since mmWave ARoF combines the optical and wireless channels, the design requirements for the modulation formats are more extensive and complex. The mmWave ARoF requirements for modulation formats can be expressed in terms of key performance indicators (KPIs). The main KPIs for modulation formats in a mmWave ARoF system are the following [6]:

- ***Peak-to-average power ratio (PAPR)***: this parameter is obtained by dividing the maximum peak by the power average of the signal. High PAPR leads to important degradations of the signal in devices such as digital-to-analog converters (DAC), MZM, and RF amplifiers. Since all these devices are involved in a mmWave ARoF system, low PAPR is a relevant KPI for this type of systems. Moreover, there are multiple techniques that take a step toward reducing the PAPR by DSP. However, any PAPR reduction technique implies at least one of the following: power increase, bandwidth expansion, or bit error rate (BER) degradation.
- ***Robustness to phase noise***: phase noise represents a random fluctuation in the phase of the waveform. This impairment is produced in the mmWave tone generation to upconverting or downconverting the information signal. In ARoF systems, this generation can be performed in the optical or electrical part. The phase noise level depends on the employed technique to generate the mmWave tone. The phase noise represents one of the major limiting factors in ARoF systems [7]. For this reason, high robustness to phase noise is a crucial KPI to reach high performances in ARoF systems.

- ***Robustness to multipath channels***: as the signal of the transmitter antenna follows different paths in a wireless environment, the receiver antenna receives copies of the transmitted signal with delays and attenuations corresponding to each of these multipaths. This multipath effect produces a degradation in the received signal. In this context, high robustness to multipath channels is an important KPI for any wireless link. However, since the secondary paths are more attenuated in high carrier frequencies due to reflections and refraction processes, the mmWave channel displays less multipath effect than in lower-frequency bands. Therefore, robustness to multipath channels is not a determinative KPI for mmWave ARoF systems.
- ***Spectral efficiency***: this parameter is related to the achieved throughput of the used bandwidth. When the bandwidth of the system is fixed, the spectral efficiency determines the maximum bit rate that can be reached. As most of the communication channels are limited in spectrum, high spectral efficiency is a relevant KPI to be considered. This KPI is more critical in the mmWave wireless channel in contrast to its optical counterpart since optical spectrum is much wider than mmWave spectrum. Furthermore, since the frequency spectrum was commercialized, high spectral efficiency implies lower costs.
- ***Complexity***: a commercial communication system requires a real-time DSP process. Field-programmable gate array (FPGA) is a suitable solution to achieve real-time DSP in prototypes with reduced manufacturing time process. However, FPGA devices are limited in process blocks. Thus, low complexity is a KPI that allows to integrate modulation formats in FPGA systems. In addition, low complexity KPI reduces latency and cost because the DSP is simplified.

Modulation formats can be classified into two big groups: single-carrier (SC) waveforms and multi-carrier (MC) waveforms. The "under-test" MC waveforms are the following: OFDM, universal-filtered multi-carrier (UFMC), and generalized frequency-division multiplexing (GFDM). On the other hand, single-carrier frequency-division multiplexing (SC-FDM) and multiband carrierless amplitude phase modulation (multi-CAP) are the examined SC waveforms. However, SC-FDM and multi-CAP are not pure SC waveforms. Therefore, these modulation formats present hybrid SC and MC waveform properties.

SC waveforms present lower PAPR level than MC waveforms. Moreover, SC waveforms are more robust in terms of phase noise. However, SC waveforms are less tolerant to multipath channels, and MC waveforms can imply higher spectral efficiency due to its low out-of-band (OOB) emissions. Hence, the characteristics of a modulation format can be briefly resumed depending on the waveform group that it belongs.

Table 8.1 shows a qualitative comparison of the under-test modulation formats according to the KPIs previously explained. As it can be observed, it does not rationalize to assess the best modulation format since each one of these presents advantages and disadvantages. For example, GFDM is a good choice in terms of spectral efficiency and robustness to multipath channels. However, GFDM presents high PAPR and complexity. On the other hand, multi-CAP performs better than

Table 8.1 Comparison of the evaluated modulation formats in terms of mmWave ARoF KPIs

	OFDM	SC-FDM	UFMC	GFDM	Multi-CAP
PAPR	High	Low	High	High	Low
Robust. to phase noise	Medium	Medium/high	Medium	Medium/high	Medium/high
Spectral efficiency	High	High	Very high	Very high	Medium/high
Robust. to multi. chan.	High	Medium/high	High	High	Medium/high
Complexity	Medium	Medium/high	Medium/high	High	Medium/low

GFDM in terms of complexity and PAPR. However, the spectral efficiency and robustness to multipath channels of multi-CAP is lower than in the GFDM case.

As mmWave ARoF channel is very complex to analyze, it is very difficult to determine the most relevant KPI. Therefore, a qualitative comparison of the mmWave ARoF KPIs is not sufficient to decide the best modulation format choice for this type of system. For this reason, a comparison of the examined modulation formats in an experimental mmWave ARoF system is realized, which is elaborated in the next subsection.

8.2.3 Practical Experiment

Since it is very complex to select the best modulation format in mmWave ARoF systems through a qualitative comparison, an experimental comparison is needed to determine the best modulation format candidate. In this subsection, this experimental comparison is presented and explained as two parts: in the first part, the experimental mmWave ARoF setup is explained thoroughly, while the second part provides the analysis and interpretation of the results.

8.2.3.1 ARoF Testbed Description

Figure 8.2 shows all the components used in the experimental comparison. In this experimental comparison, a scenario of Fig. 8.1 is simulated.

Figure 8.2a represents the mmWave ARoF scheme. First, an external cavity laser (ECL) generates an optical carrier at 1550 nm. Then, this optical signal is modulated by an MZM, biased in the null point, and driven by a sinusoid of 12.5 GHz. This sinusoid is produced by a vector signal generator (VSG). At the MZM output, two optical tones with a separation of 25 GHz are generated. These two optical tones correspond to the second harmonic of the MZM. The optical spectrum of these two tones can be observed in Fig. 8.2b. In this graph, it can be noticed that the undesired central tone is not totally removed. In addition, as the MZM is sensitive to optical polarization, a polarization controller (PC) is set in the input of this device. Therefore, the optical output power of the MZM can be optimized by tuning this

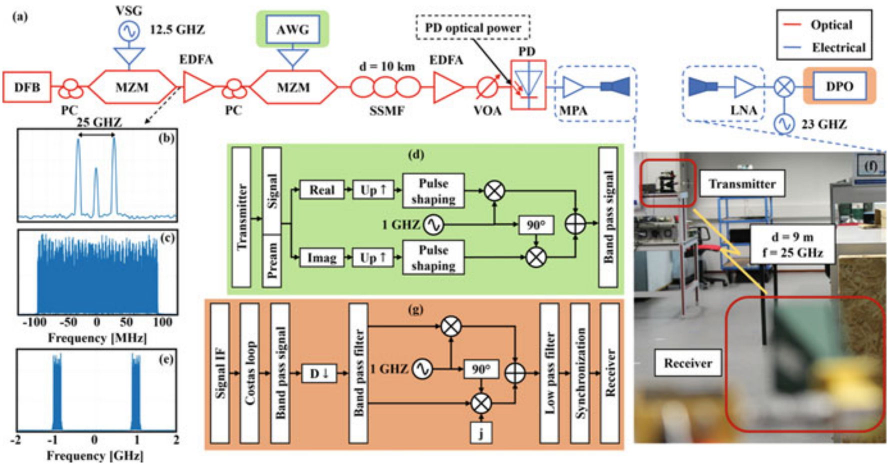


Fig. 8.2 Experimental testbed description: (**a**) experimental setup; (**b**) optical spectrum after the first MZM; (**c**) baseband spectrum of the transmitted signal; (**d**) block diagram of the DSP in the transmitter; (**e**) IF signal spectrum of the transmitted signal; (**f**) photo of the mmWave wireless link; (**g**) block diagram of the DSP in the receiver

PC. After the MZM, the two optical tones are boosted with an erbium-doped fiber amplifier (EDFA) because the MZM introduces high attenuation.

The signal of the under-test modulation formats is produced electrically by an arbitrary waveform generator (AWVG) with a sampling rate of 12 GSa/s. In the AWVG, the DSP process of the transmitter is realized. This process is performed offline. Each evaluated modulation format has its specific baseband transmitter scheme. The spectrum of the OFDM baseband signal is represented in Fig. 8.2c as a reference. The bandwidth of this baseband spectrum is 245.76 MHz. After the baseband process of each modulation format, an intermediate-frequency (IF) process is realized. This IF process is common to all the evaluated modulation formats, and its block diagram is represented in Fig. 8.2d. In this diagram, the first block corresponds to the specific process of a determined modulation format in the baseband domain. Then, a preamble is included in the baseband signal for synchronization at the receiver. Next, the real and imaginary parts are separated and upconverted independently. The real and imaginary parts are filtered by a pulse shaping. Subsequently, the real and imaginary parts are multiplied by a sine and cosine of 1 GHz, respectively. Finally, the signals of both branches are combined. In this way, the signal of each modulation format is upconverted to an IF of 1 GHz. The spectrum of this IF signal is shown by Fig. 8.2e. The two optical tones are modulated by a second MZM and driven by the IF signal. As in the previous MZM, the second MZM needs a PC to maximize its output power. The two modulated tones are transmitted through a standard single-mode fiber (SSMF) of 10 Km. All the processes behind the SSMF are realized in the CO of Fig. 8.1. The SSMF corresponds to the optical fiber ring of Fig. 8.1.

The output of the SSMF is boosted by a second EDFA. A photodiode (PD) is employed to convert the optical tones into the electrical domain. In the PD, the two tones beat generating an RF signal at 25 GHz corresponding to the separation of both tones. Then, the electrical signal is boosted by a 30 dB medium power amplifier (MPA), and the boosted signal is transmitted over a wireless mmWave link through two 18.5 dBi horn antennas. The RU of Fig. 8.1 are compounded by the PD, MPA, and transmitter horn antenna in this experiment. Figure 8.2f shows a photo of the experimental wireless link at 25 GHz. The distance of this wireless link is 9 m.

In the second horn antenna, the received signal is amplified by a 40 dB low-noise amplifier (LNA). Then, the amplified signal is multiplied with a sinusoid of 23 GHz by a RF mixer. Thereby, the electrical signal is downconverted to a second IF at 2 GHz. Finally, the downconverted IF signal is sampled by a digital phosphor oscilloscope (DPO) with a sampling rate of 12.5 GSa/s.

An offline DSP process is realized with the sampled IF signal. The block diagram of this DSP process is depicted in Fig. 8.2g. First, the IF signal is downconverted to the first IF (1 GHz) by a Costas loop process. The Costas loop process enables phase tracking in this downconversion. Then, the resulted signal is downsampled and filtered by a band-pass filter (BPF). After the BPF process, the real and imaginary parts are obtained by multiplying the filtered signal with a sine and cosine of 1 GHz, respectively. Hence, a complex baseband signal is achieved. Next, the baseband signal is filtered by a low-pass filter (LPF) and synchronized by using the preamble inserted in the transmitter. Finally, the specific baseband receiver process of each modulation format is performed.

According to the configuration parameters employed in each under-test modulation format, the OFDM configuration is based on the first 5G standard [2]. The employed OFDM parameters for this experimental comparison are the following: subcarrier spacing of 60 kHz, cyclic prefix (CP) of 1.2 μs, 4096 total subcarriers, 3168 active subcarriers, 928 null subcarriers to reduce the OOB emissions, and 1 pilot tone inserted on every 12^{th}active subcarrier. In respect of the parameters of the remaining modulation formats, the used UFMC configuration employs 128 sub-bands; GFDM uses 3 sub-symbols; and multi-CAP uses 9 sub-bands. The rest of the parameters are adapted to the OFDM configuration described previously.

In order to have a fair comparison, all the evaluated modulation formats employ the same bandwidth (245.76 MHz) and modulation order. The employed modulations are quadrature phase-shift keying (QPSK) and 16-quadrature amplitude modulation (16-QAM). Thereby, the spectral efficiency is the same for all the modulation formats. By using QPSK modulation, the throughput is 325 Mbps, and by using 16-QAM, 650 Mbps. Therefore, the spectral efficiencies for QPSK and 16-QAM cases are 1.32 bit/s/Hz and 2.64 bit/s/Hz, respectively.

8.2.3.2 Analysis and Interpretation

This subsection presents a thorough analysis and interpretation of the experimental results obtained in the setup explained in the previous subsection. The results are

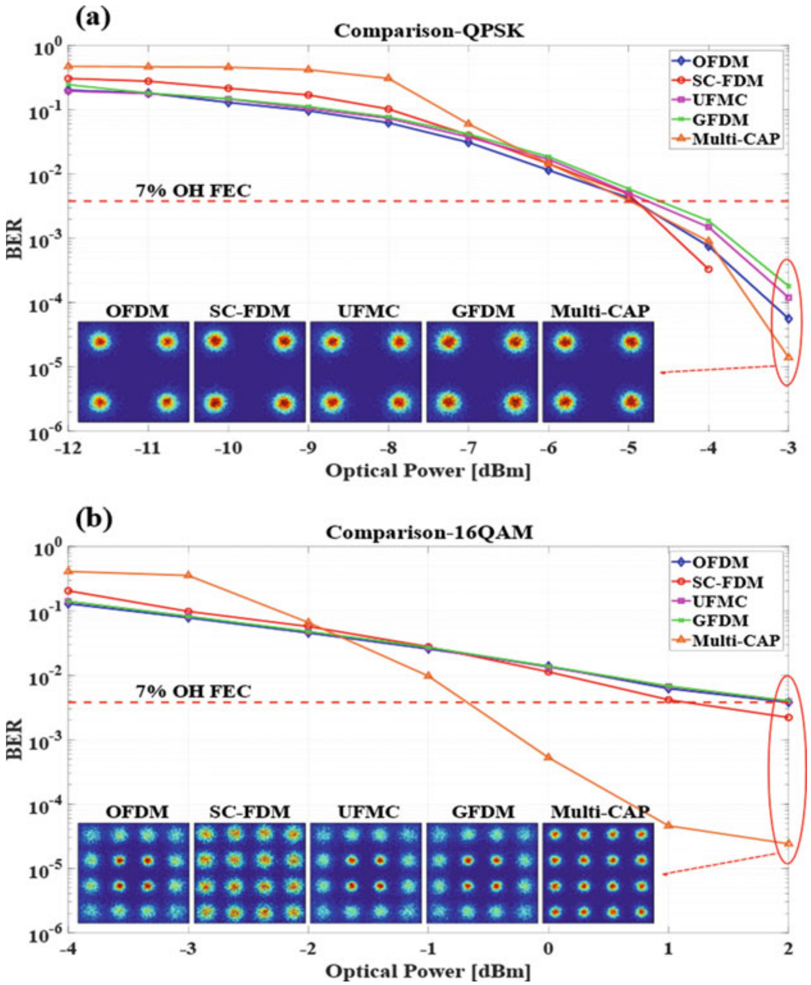


Fig. 8.3 BER as a function of the optical power in the PD for the evaluated waveforms using different modulation orders: (**a**) QPSK modulation (upper graph); (**b**) 16-QAM modulation (lower graph)

shown in Fig. 8.3. These graphs represent the BER results as a function of the optical power received in the PD. As mentioned before, QSK and 16-QAM modulations are used to compare the under-test modulation formats, Fig. 8.3a–b, respectively. Furthermore, the 7% overhead forward error correction (OH FEC) is represented by the red dotted line. In addition, the constellation of the received symbols of all the modulation formats in the maximum power points is also illustrated in both graphs.

Observing the maximum power point (−3 dBm) in the graph of Fig. 8.3a, it can be noticed that the performances of the modulation formats in terms of BER follow this order: SC-FDM, multi-CAP, OFDM, UFMC, and GFDM; this also corresponds to the PAPR level exhibited by each modulation, being SC-FDM the modulation format with minimum PAPR and GFDM the modulation format with maximum PAPR (see Table 8.1). Therefore, a low PAPR of the transmitted signal is an important KPI to achieve better performance in the proposed experimental mmWave ARoF setup.

On the other hand, examining the maximum optical power point (2 dBm) in the graph of Fig. 8.3b, the best modulation formats in terms of BER follow this order: multi-CAP, SC-FDM, OFDM, UFMC, and GFDM. In the 16-QAM case, multi-CAP outperforms SC-FDM despite presenting higher PAPR. Multi-CAP presents 2.5 dB of optical power gain respecting the SC-FDM solution for the 7% OH FEC threshold. The reason for this being that the employed channel equalizer is different in the multi-CAP case in contrast to the remaining modulation formats. Multi-CAP uses decision feedback equalizer (DFE) with the least mean square (LMS) algorithm in the time domain, while the remaining modulation formats use least-squares (LS) equalizer in the frequency domain. Multi-CAP utilizes a different equalization process because each sub-band is independent as an SC waveform, and thus the LS equalizer is not suitable for compensating the channel on an SC waveform. Multi-CAP performs much better than the rest in the 16-QAM comparison because its equalizer amplifies less the noise in contrast to the LS equalizer. Thereby, the equalizer strategy is also crucial in mmWave ARoF scenarios to obtain the best yield.

The conclusions that can be raised through the analysis and interpretations of the presented experimental results are the following: low PAPR and optimum equalizer selection are keys to achieving high performances in this type of system. Moreover, the results show that the standardized OFDM is not clearly the best modulation format candidate for mmWave ARoF systems and other modulation formats, such as SC-FDM and multi-CAP, should be considered.

According to the future lines, this work can be followed by varying the mmWave wireless channel conditions: longer distance of the link, outdoor experiment, non-line-of-sight (NLOS) propagation, etc. Moreover, several optical schemes to generate mmWave tones can be compared: external modulation, two free-running lasers, phase-locked lasers, etc. One of the key challenges of this work consists of increasing the bit rate by using higher modulation orders such as 64-QAM and 256-QAM. To achieve this, both DSP and hardware systems must be optimized. Another key challenge is related to the adaptation of the bit rate according to the channel conditions in real time. This adaptable bit rate system can be realized by utilizing machine learning.

8.3 Channel Equalization for OFDM-Based mmWave ARoF Systems

8.3.1 Introduction

In OFDM (orthogonal frequency-division multiplexing) systems, channel estimation, and channel equalization play a key role in overcoming distortions caused by phenomena like fading, delay spread, and multipath effect. Channel equalization is needed in optical fiber communication systems due to the effect of chromatic dispersion (CD) in the optical link making it very difficult to decode the received OFDM symbols as the bit symbols get broadened and distorted.

Currently, the Common Public Radio Interface (CPRI) is mainly utilized in the fronthaul links between central units (CUs) and distributed units (DUs), which is a digital transmission scheme of quantized waveforms of baseband signals, called "digital radio over fiber (DRoF)." However, CPRI requires extreme high transmission capacity compared to the original user throughput due to the digitization process. Considering the approaching 5G system, in which the peak throughput will be around 20 Gb/s, it is obvious that CPRI is not scalable [8]. To improve the bandwidth utilization efficiency, analog radio over fiber (ARoF) has been used instead of digital radio over fiber (DRoF).

8.3.1.1 Noise and Distortion Effects on OFDMmmWave RoF Systems

When OFDM signal is carried by mmWave RoF systems, its performance is affected by various physical layer impairments such as noise, dispersion, and nonlinear distortion. The nonlinear distortion effect is a critical problem that affects the performance of OFDM signal causing the loss of orthogonality among the different subcarriers and resulting in performance degradation.

The dispersion leads to pulse broadening due to the different spectral components of the signal having different group delays. The chromatic dispersion of a fiber mode is given by

$$D = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{1}{\mathrm{c}}\frac{\mathrm{d}\beta}{\mathrm{d}k}\right] \tag{8.1}$$

where c is the speed of light in vacuum. The bracketed quantity in (8.1) is the group-delay time per unit length. The chromatic dispersion (CD) is the change in group-delay time per unit fiber length per unit wavelength interval and is typically expressed in units of ps/nm· Km. For conventional telecommunications fiber operating near 1550 nm, this dispersion is typically +16 ps/nm. Km [9].

For bit rates up to 2.5 Gbit/s, problems related to dispersion can be solved using narrowband transmitters. For high bit rate systems ($\geq$ 10 Gb/s) as in the case of orthogonal frequency-division multiplexing (OFDM), the dispersion limits the transmission distance. This raises the need for some sort of dispersion compensation. Also, for high bit rate systems ($\geq$ 10 Gb/s), the dispersion slope becomes an important factor since strength due to CD is reduced by the square of the bit rate [10].

8.3.1.2 Equalization Requirements for Converged mmWave and RoF System

Cloud-based radio access networks (C-RANs) provide a cost-effective, energy-efficient, and high spectral-efficient solution for future access network. A backhaul network connecting a growing number of small base stations and supporting a C-RAN network is extremely important. Such a backhaul network should have high energy efficiency, flexibility, low transmission delay, low cost, and high capacity. The most obvious solution to realize such a backhaul network would involve the use of a converged mmWave network. At the mmWave bands, 60 GHz has 7–9 GHz of license-free bandwidth worldwide, while the 70–90 GHz spectrum (71–76, 81–86, and 92–95 GHz) has 13 GHz of licensed bandwidth available [11]. The convergence of mmWave with optical fiber networks provides high capacity, flexibility, wide area coverage, cost-effectiveness, and energy efficiency for multiple-gigabit signal delivery in high-speed mobile and broadband wireless access networks [12]. However, such a converged system would suffer from the chromatic dispersion in the standard single-mode fiber (SSMF). So, there is a need of equalization to compensate dispersion in the fiber as it deteriorates the output signal.

8.3.2 Enabling Technologies

8.3.2.1 MmWave ARoF Systems

RoF jointly takes advantage of the huge bandwidth offered by the optical fibers with the mobility and flexibility provided by wireless systems [13, 14]. The scarcity of the spectrum in the lower region of the microwave frequency band, where many mobile and wireless communication services operate, has led to the interest in mmWave communication systems. MmWave RoF technology plays a key role in the next-generation optical wireless networks, having great potential to deliver multi-gigabit wireless services through centralized control unit and simplified remote base stations with low loss and bandwidth abundant fiber-optic connectivity. RoF systems operating at 60 GHz have gained much attention due to the large unlicensed bandwidth availability.

The mmWaves are the electromagnetic waves with wavelength ranging from 1 to 10 mm. Therefore, mmWave band ranges from 30 to 300 GHz. Different mmWave frequency bands have been proposed for high capacity wireless systems employing

RoF: in the 24–30 GHz band and 75–110 GHz band [15], at 120 GHz, at 250 GHz, and more recently at 220 GHz. However, the frequency band that has attracted major importance in recent research activities is around 60 GHz, mainly due to two reasons: 1) reduction of the cell size due to high atmospheric attenuation [16] leads to frequency reuse, expanding the wireless system capacity [17] the spectrum license at 60 GHz is free and such a high frequency can provide huge bandwidth.

8.3.2.2 OFDM Techniques for RoF Passive Optical Networks

PONs use multicarrier modulation like OFDM, which provides an opportunity for increased bandwidth at affordable cost. OFDM provides better spectrum utilization and high transmission by harnessing M-ary modulation on its subcarriers, such as phase-shift keying (PSK) or quadrature amplitude modulation (QAM). In the OFDM-PON, a passive optical splitter is used to connect two or more optical network units (ONUs) [18].

In [19], the authors design a WDM RoF PON based on OFDM and optical heterodyne, which can achieve 40 Gbit/s per wavelength channel and wired/wireless access synchronously.

In [20], the authors experimentally demonstrated a WDM-PON system to provide the triple-play services using centralized light wave sources with symmetric data at 10 Gbit/s per channel for both downstream and upstream data.

8.3.2.3 Generation of mmWave Signal

A CW laser and an external modulator such as Mach-Zehnder modulator (MZM) or electroabsorption modulator (EAM) are used to modulate the intensity of light. In this scheme, an electrical signal from “FuncSinEl” is input into a Mach-Zehnder modulator (MZM). At the output of the “FuncSinEl” module in VPI, the signal consists of many sidebands as shown in Fig. 8.4, with frequency separation between two successive components equal to the frequency of the input electrical signal. In our simulation setup, a signal with a frequency separation of four times the input electrical signal frequency is selected (Fig. 8.4). The input electrical signal frequency is 15 GHz.

The generated optical signal is then passed through an arrayed waveguide grating (Filter_AWVG) module, and frequency spacing between adjacent channels is adjusted. Subsequently, a “bus selector” module is used to select the appropriate sidebands where two sidebands are selected as shown in Fig. 8.5 separated at 60 GHz.

Next a “WDM_MUX” module is used to merge the two sidebands to a single sideband. An optical band-pass filter (OBPF) is used to reduce the amplified spontaneous emission noise. With this scheme, a flexible-frequency mmWave signal could be generated, presenting a simple and flexible method for mmWave signal

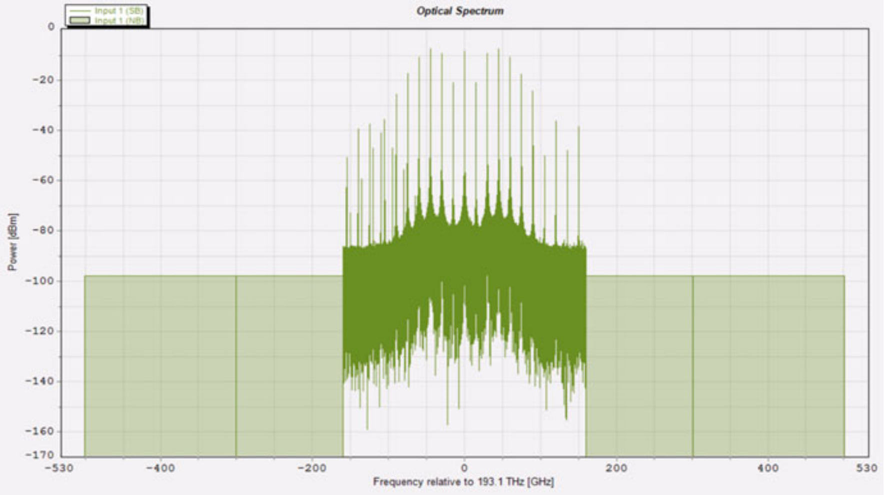
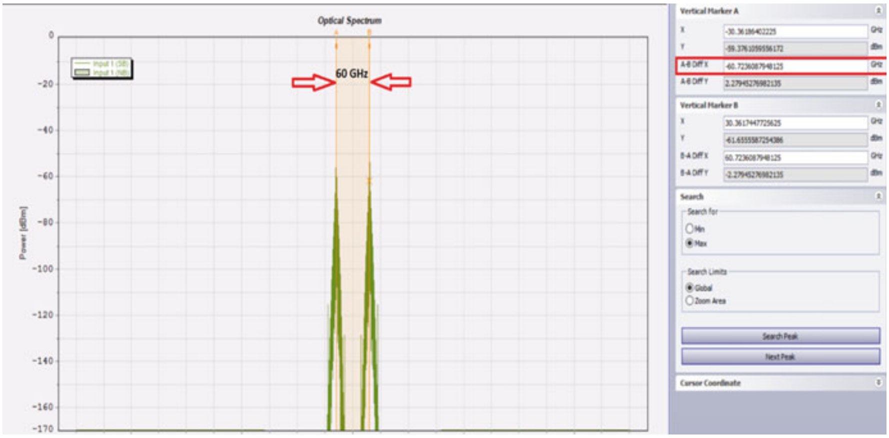


Fig. 8.4 Optical spectrum after Mach-Zehnder modulator



Vertical Marker A

X -30.36186402225 GHz
Y -59.3761059556172 dBm
A-B Diff X -60.7236087948125 GHz
A-B Diff Y 2.27945276982135 dBm

Vertical Marker B

X 30.3617447725625 GHz
Y -61.6555587254386 dBm
B-A Diff X 60.7236087948125 GHz
B-A Diff Y -2.27945276982135 dBm

Search

Search for: ○ Min ● Max

Search Limits: ● Global ○ Zoom Area

Fig. 8.5 Selection of mmWave sidebands (60 GHz)

generation. Figure 8.6 shows the generation of mmWave signal in a commercially available VPI simulation software.

8.3.2.4 Channel Equalization and LMS Approach

In OFDM systems, channel equalization plays a key role in overcoming distortions caused by phenomena like fading, delay spread, and multipath effect. The basic operation of channel equalization is to inverse the transmission channel impairments

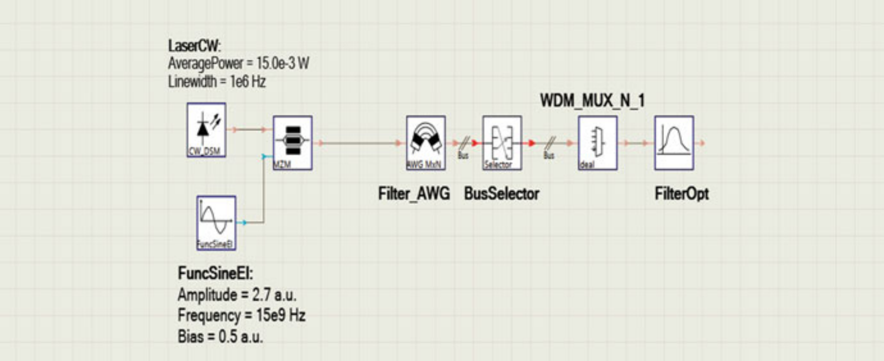


Fig. 8.6 Generation of mmWave

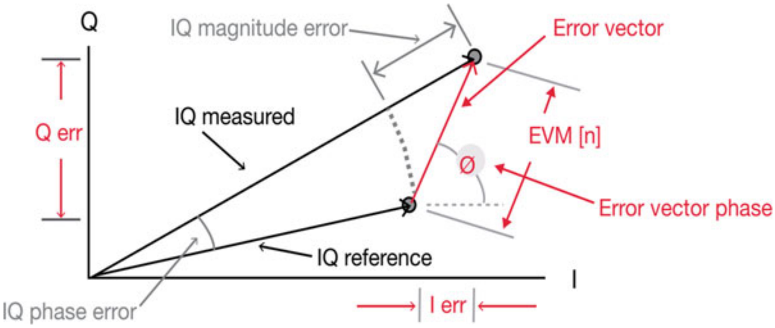


Fig. 8.7 Calculation of error vector and error vector magnitude [J. Cacazos, D.McGrath, and N. Faubert ‘Engineering the 5G world- Design and Test Insights’ Keysight Technologies, 2020]

such as frequency-dependent phase and amplitude distortion. Adaptive equalization is a technique that automatically adapts to the time-varying properties of the communication channel. The least mean square (LMS) algorithm is one such popular technique that can be used for adaptive channel equalization. The criterion used in this algorithm is to minimize the mean square error (MSE) between the desired output and the actual output [21]. The approach used here in this work is to provide the distorted signal and ideal signal extracted from the setup built in VPI software and import these to LMS equalizer implemented in MATLAB. First, both signals are normalized in the equalizer, and the equalizer taps are then found by an iterative method that is used to find the equalized signal. After equalization of the received distorted symbols, error vector magnitude (EVM) of the equalized constellation is calculated. Figure 8.7 shows the calculation of the EVM metric.

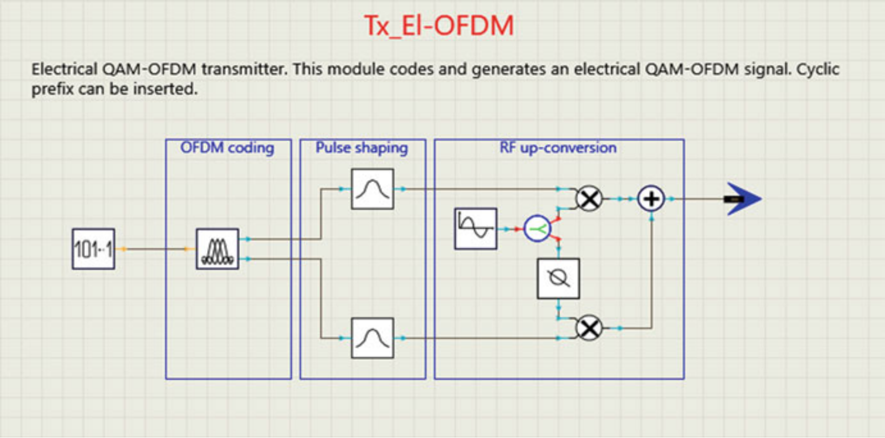

Tx_El-OFDM
Electrical QAM-OFDM transmitter. This module codes and generates an electrical QAM-OFDM signal. Cyclic prefix can be inserted.
OFDM coding
Pulse shaping
RF up-conversion
101-1

Fig. 8.8 Implementation of OFDM transmitter in VPI

8.3.3 *Simulation Model and Performance Evaluation*

8.3.3.1 Direct Detection Converged OFDM RoF mmWave System: 60GHz

To design a complete system consisting of a converged OFDM RoF and mmWave, first we design an OFDM transmitter in a commercially available simulation software VPI transmission maker. The blocks used inside the OFDM transmitter module in VPI (Tx_El_OFDM_vtmg1) are shown in Fig. 8.8. This module generates electrical signals corresponding to the real and imaginary parts of an OFDM signal.

The pseudorandom binary sequence (PRBS) block is used to generate data, at a rate determined by the modulation level and the bit rate. The raw digital binary bits are distributed into data streams, and each stream is then encoded according to the settings in the “subcarrier Modulation” parameters group of the coder block. The number of bits encoded in each symbol is given by the parameter “Bits Per Symbol QAM.” Individual modulation formats can also be specified for different subcarriers. The OFDM coding stage is then followed by pulse shaping. In this stage, the rectangular input pulses are pulse-shaped by a filter with a raised cosine characteristic. After the pulse shaping stage is the RF upconversion stage, in which the in-phase channel data modulates a cosine wave carrier, while the quadrature channel data modulates the sine wave carrier via an RF phase shifter, using a sine wave generator and mixers. Finally, an OFDM signal is upconverted to the chosen RF frequency, with multilevel in-phase and quadrature phase M-QAM coded symbols. Subsequently, a 60GHz mmWave signal is generated as described in the previous section. The generated converged mmWave and OFDM signal is then sent to a standard single-mode fiber (SSMF). The signal is then detected on a single photodiode and then sent to OFDM receiver to recover the subcarriers. The blocks used inside the OFDM receiver module in VPI are shown in Fig. 8.9.

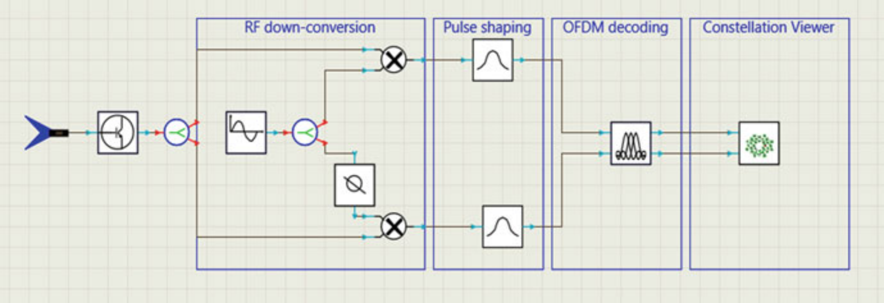


Fig. 8.9 Implementation of OFDM receiver in VPI

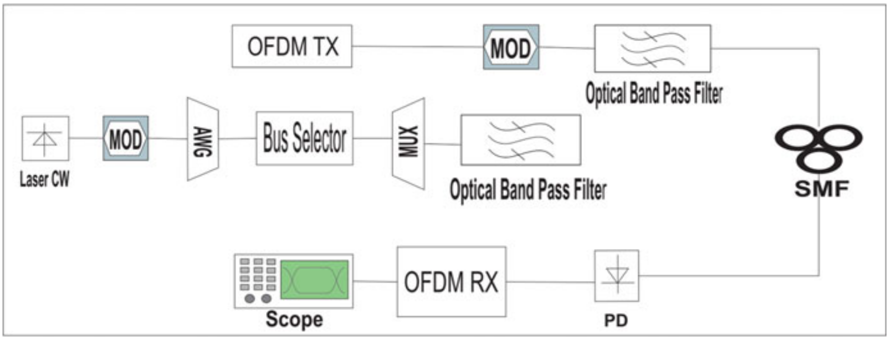


Fig. 8.10 Direct detection converged OFDM RoF mmWave system at 60GHz

OFDM receiver module decodes an electrical QAM-OFDM signal by reversing the process of the transmitter. The demodulation process begins with the electrical signal first being downconverted to baseband. Following this, pulse shaping is applied, and the signal is decoded in the decoder module. After decoding the signal, constellation viewer is used to view the constellation. A complete simulated setup is shown by Fig. 8.10.

8.3.3.2 Performance Evaluation

The system described in Sect. 8.3 is simulated in a commercially available design suite “VPI transmission maker,” and the data is extracted from the simulation and processed offline in MATLAB with the proposed algorithm identified in Sect. 8.3.2 (“Channel Equalization and LMS Approach”).

Offline processing resulted in constellation diagrams and error vector magnitude (EVM) measurements, as shown by Figs. 8.11 and 8.12, where 4-QAM and 16-QAM are the modulation formats used in each OFDM subcarrier. The performance

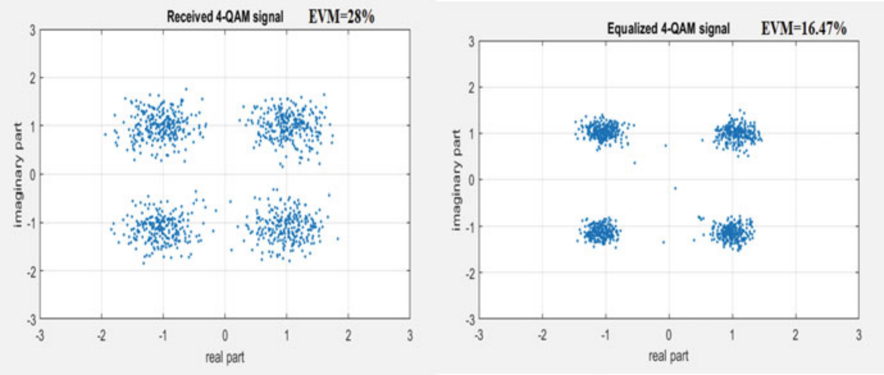


Fig. 8.11 Constellation diagram of 4-QAM as a modulation format in OFDM subcarriers before and after equalization (30 Km SMF link)

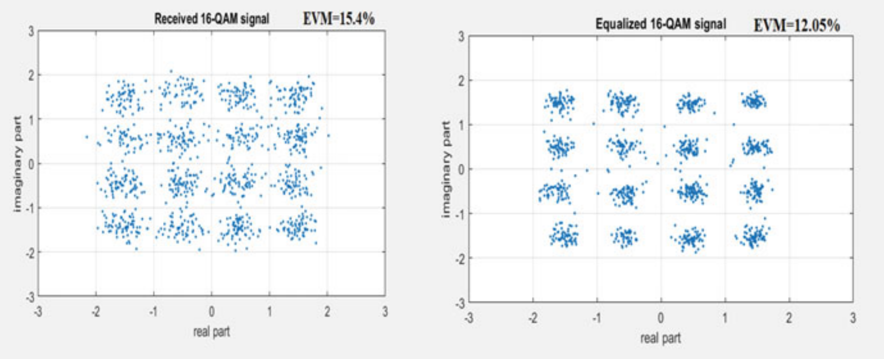


Fig. 8.12 Constellation diagram of 16-QAM as a modulation format in OFDM subcarriers before and after equalization (15 Km SMF link)

of the algorithm is tested in the form of maximum transmission distance achieved in the presence of fiber dispersion and other nonlinearities. It can be seen that for the 4-QAM case, as a modulation format in each subcarrier, a maximum transmission distance of 30 Km has been achieved with an error vector magnitude (EVM) of 16.47% after equalization and in the case of 16-QAM as a modulation format, a maximum transmission distance of 15 Km has been achieved with an error vector magnitude of 12.05% after equalization. Both EVM values are below the suggested threshold of 3GPP for the corresponding modulation formats. In particular, Figs. 8.11 and 8.12 present the constellation diagrams obtained before and after equalization in a converged mmWave radio over fiber setup as described by Sect. 8.3; EVM vs fiber length for 16-QAM is shown by Fig. 8.13.

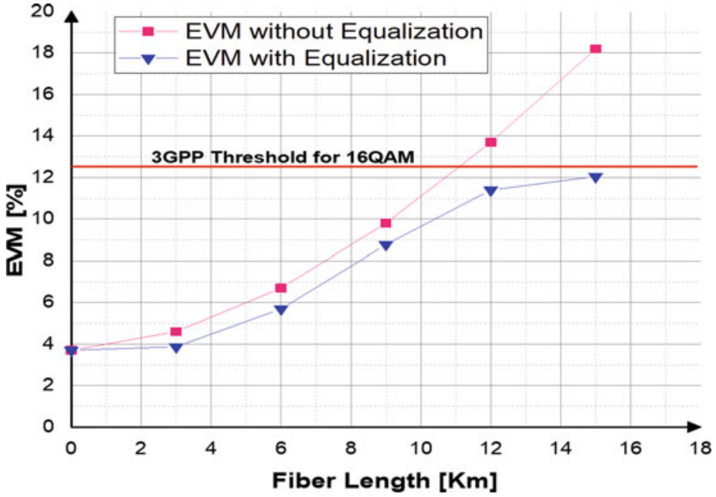


Fig. 8.13 EVM vs fiber length for 16-QAM

8.3.4 *The Key Challenges Ahead*

The main contributions presented in this work spur several new research lines for future investigation on coherent OFDM mmWave systems. These include comprehensive analysis on the range of different parameters involved in the system such as the number of channels used in the OFDM framework; modulation formats (e.g., 64-QAM, 128-QAM, 256-QAM) on different OFDM subcarriers and equalization approaches, among others, are currently missing; and the direct detection analysis of the converged OFDM RoF mmWave at 60 GHz investigated in Sect. 8.3.3 can be further extended to a coherent setup with a similar analysis for equalization.

The importance of this future work would be to ensure better equalization with coherent detection and long reach in single-mode fiber.

8.4 Reflective Electroabsorption Modulator for 50 Gb/s Colorless Transmission

8.4.1 *Optical Components for 5G Fronthauling*

The fronthaul is an optical link between the digital unit (DU) or the baseband unit (BBU), depending on the modulation scheme, and the remote radio head (RRH). Some of the key requirements of a 5G fronthaul are ultralow latency, high reliability, large number of connected users, and high data rate [22]. The number of connected users can be increased via network densification. To minimize the end-to-end (E2E) delay, wavelength-division multiplexing passive optical network (WDM-PON) is a promising solution. The actual WDM-PON implementation can follow a logical point-to-point (PtP) or a point-to-multipoint topology. The latter introduces additional delay though as multiple end users share a single wavelength channel. The International Telecommunication Union (ITU) in its next-generation PON2 (NG-PON2) standard recommends implementing a time-and-wavelength-division multiplexing (TWDM) scheme to share up to eight WDM channels [23]. As a result, the standard is commonly referred to as TWDM-PON.

On the other hand, the high-speed requirement of a 5G fronthaul network can be satisfied in a cost-effective manner by employing transmitters that are based on electroabsorption modulators (EAMs). Compared to directly modulated lasers (DMLs), EAM-based transmitters provide better transmission performances because of the absence of adiabatic chirp which is not the case in DMLs. In TWDM-PON, the transceiver components are also required to be wavelength-tunable on both the optical line terminal (OLT) and the optical network unit (ONU). Although IEEE's 50 Gb/s Ethernet PON (50G-EPON) standard recommends using fixed-wavelength sources, it is based on multiplexing only two 25 Gb/s wavelength channels [24]. In any case, it is necessary to increase the number of WDM channels for a large-scale deployment of 5G networks satisfying the transport latency requirement. At the same time, the transceiver module of a 5G fronthaul must be able to operate at any wavelength of the WDM system by being either wavelength-tunable or wavelength-independent (colorless).

However, by the nature of tunable devices, they require tight wavelength control. If such devices are used in a burst mode transmission, which causes frequency drift due to thermal variation between burst on and off states [25], complex control circuitry will be required to stabilize the emission wavelength. Colorless transmitters, on the other hand, do not require wavelength tunability and thus can be utilized to realize low-cost 5G fronthaul transmitters. For example, a reflective EAM monolithically integrated with a semiconductor optical amplifier (SOA) can be used as a standalone colorless transmitter at the ONU or as an array at the OLT. The main drawback of such a configuration is the need for an external optical source. But a single multi-wavelength fixed source such as a comb laser can be used to support multiple sites. Figure 8.14a shows a schematic diagram of an uplink WDM-PON transmission network topology using reflective EAM-SOAs (REAM-SOAs) as ONU transmitters. In this section, we present a complete characterization of REAM-SOAs that can operate up to 50 Gb/s with simple digital modulation formats such as non-return-to-zero (NRZ) without equalization. Although the results presented here are for digital signals, the components can also be effectively applied to analog transmissions.

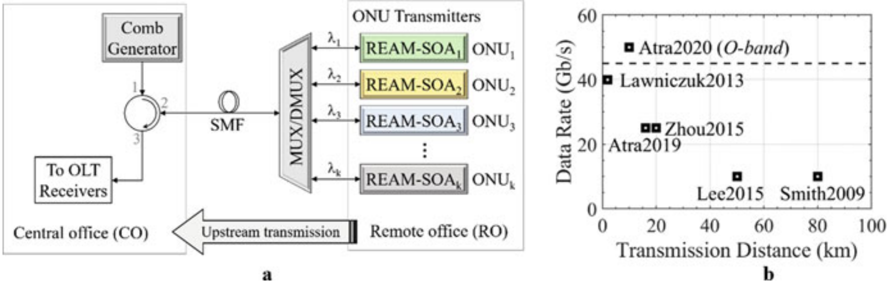


Fig. 8.14 (**a**) Schematic diagram of a remote-seeded WDM-PON uplink transmission using REAM-SOAs as ONU transmitters. (**b**) State-of-the-art reflective EAM-SOAs

8.4.2 *Review on Reflective EAM-SOAs*

Reflective EAM-SOAs were widely studied until 2015, the year NG-PON2 was standardized by ITU, mainly for using the devices as colorless transmitters. As a result, for compatibility reasons, most device demonstrations available in the literature are in the C-band, operating at 10 Gb/s for a minimum transmission distance of 20 km [26, 27]. Figure 8.14b summarizes state of the art of REAM-SOAs in terms of data rate and transmission distance.

For C-band transmissions, there is always a tradeoff between data rate and the transmission distance because of high chromatic dispersion in optical fibers. A few C-band devices were demonstrated at higher bit rates than 10 Gb/s. For example, [28] demonstrated a REAM-SOA based on AlGaInAs/InP technology that can operate up to 40 Gb/s NRZ, but the transmission distance was limited to 2 km. When operated at 10 Gb/s NRZ, the same device is capable of transmitting up to 20 km. Similarly, [29] demonstrated a 25 Gb/s REAM-SOA transmitting up to 20 km, but the experiment involved offline digital signal processing (DSP), which is not generally desirable for access network components as it increases the transceiver cost. We demonstrated a dispersion-limited 16 km C-band transmission at 25 Gb/s NRZ without equalization using an InGaAsP-based REAM-SOA [30].

For the O-band wavelengths, on the other hand, chromatic dispersion is very low, and it is possible to achieve long-distance transmission at very high bit rates. However, without external amplification, the transmission distance will be limited by fiber attenuation which is higher in the O-band than the C-band. We recently demonstrated a 50 Gb/s NRZ transmission using a 100 μm O-band EAM in REAM-SOA configuration [31].

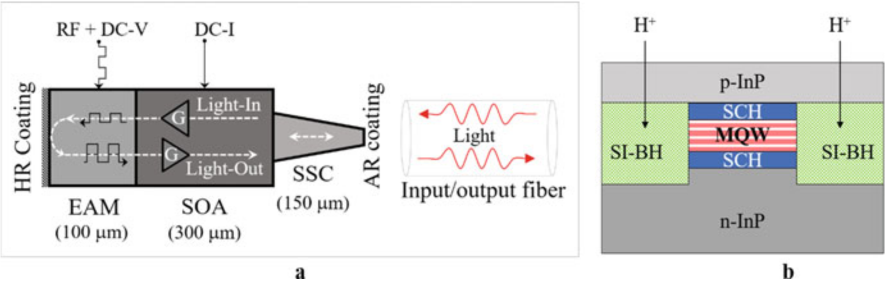


Fig. 8.15 (**a**) Reflective EAM-SOA and (**b**) SI-BH technology

8.4.3 Device Design and Technologies Applied

Figure 8.15a shows schematic diagram of a reflective EAM-SOA. Our photonic integrated circuit (PIC) comprises an EAM, an SOA, and a spot-size converter (SSC). To realize a reflective device configuration, the back facet is treated with a high-reflection (HR) coating, and the front facet is treated with an anti-reflection (ARF) coating. The devices are based on InGaAsP multiple quantum well (MQW) structures on an n-doped InP substrate, and they are realized with industrially compatible technologies. More specifically, the waveguide structures are defined by using a semi-insulating buried heterostructure (SI-BH) technology whose schematic diagram is shown in Fig. 8.15b. The semi-insulating behavior is achieved by hydrogen ion (H^+) implantation applied to the undoped InP cladding surrounding both sides of the waveguide region. Compared to other technologies used for waveguide definition, SI-BH technology provides efficient thermal dissipation, more circular (symmetric) optical mode, and improved modulator bandwidth [32].

On the other hand, the epitaxial growth process involves independent optimization of the EAM, the SOA, and the SSC sections and combining them by using a butt-joint integration technology (a multi-step growth process) which is schematically illustrated in Fig. 8.16. The design layout of a REAM-SOA with integrated SSC is also shown in the figure. One important aspect of this approach is the flexibility to separately design and engineer the bandgaps of the various active and passive components integrated into a single chip (e.g., laser, modulator, amplifier, waveguide, etc.). Detailed information about the components and operating principles of a REAM-SOA can be found in [30] and [31].

To study the effect of EAM length on both modulation bandwidth and achievable extinction ratios, we fabricated and characterized devices with different EAM lengths. In particular, we studied reflective devices with 150-μm- and 80-μm-long EAMs in the C-band and a 100-μm-long EAM in the O-band. The C-band devices are presented for comparison, especially to demonstrate the effect of chromatic dispersion on the transmission distance.

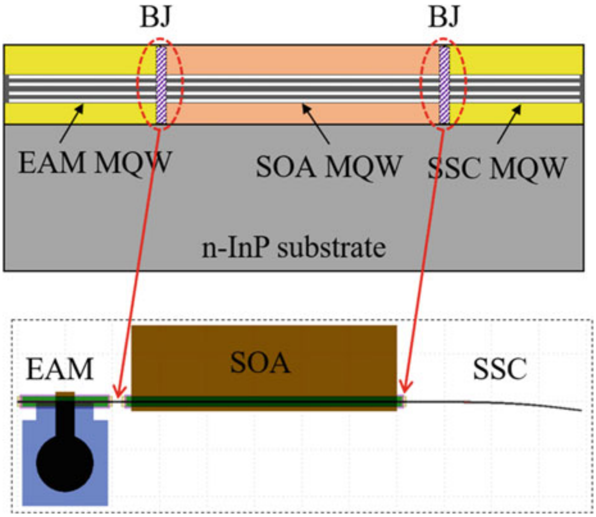


Fig. 8.16 Schematic diagram of butt-joint integration technology and design layout of a REAM-SOA with integrated SSC

8.4.4 Device Performance Evaluation

8.4.4.1 Spot-Size Converter

The SSC is a tapered waveguide with a 7° tilt. The taper increases the size of the optical mode and thus improves fiber coupling efficiency. The tilt, on the other hand, is intended to minimize optical feedback to the gain (SOA) section. A linear taper running from 1.3 μm (for O-band devices or 1.5 μm for the C-band) to 0.7 μm is integrated with the REAM-SOAs. To avoid an abrupt change of curvature in the optical path, the tapered waveguide starts with a straight section and slowly bends toward its end until the final tangential angle becomes 7°.

Figure 8.17a–b shows contour plots of simulated optical mode profiles at the taper input and output, respectively. As expected, the mode size increases at the output of the taper. For example, for the C-band devices that integrate a taper running from 1.5 μm to 0.7 μm, the simulated optical mode has a horizontal width of 2.2 μm and a vertical width of 2 μm. Similarly, the O-band taper having a narrower input waveguide has a mode diameter of 1.4 μm × 1.2 μm. After fabrication, we estimated the size of the optical beam at the output of each taper by using far-field measurement data and taking the beam diameter at $1/e^2$ point as shown in Fig. 8.17c.

Table 8.2 summarizes the estimated mode diameters of the tapers and a lensed fiber used in our experiments. The difference between simulated and measured mode diameters is attributed to fabrication tolerances. In the next subsection, we will see contributions of input/output coupling efficiencies to the total insertion loss.

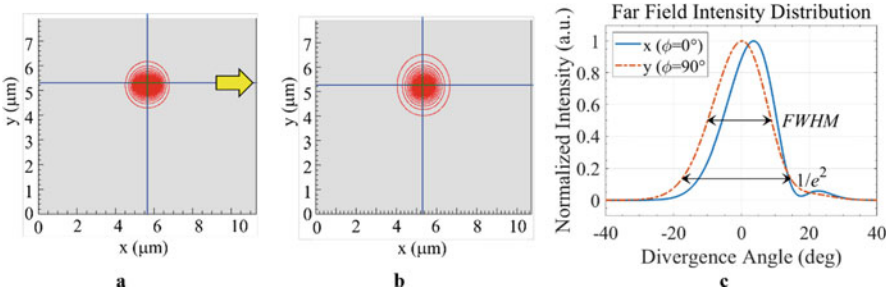


Fig. 8.17 Contour plot of simulated optical mode profile at (**a**) the input, (**b**) the output of a linear taper, and (**c**) measured far-field spectrum of an optical beam at the output of a 0.7 μm taper

Table 8.2 Simulated and measured mode diameters of a 0.7 μm taper and a lensed fiber

Component	Simulated mode diameter (μm)		Measured mode diameter (μm)	
	Horizontal	Vertical	Horizontal	Vertical
C-band taper (1.5 μm–0.7 μm)	2.2	2	3	2.5
O-band taper (1.3 μm–0.7 μm)	1.4	1.2	2.7	3.25
Lensed fiber	–	–	3	2.5

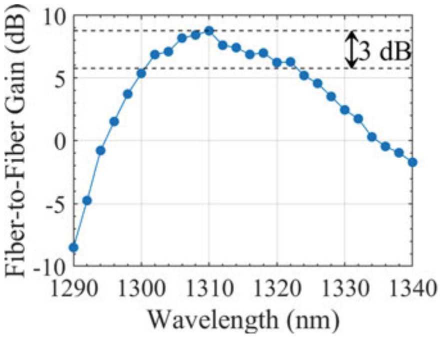
8.4.4.2 Insertion Losses and SOA Gain

The dominant insertion loss contributions in a REAM-SOA (or its opposite RSOA-EAM) come from input/output facet coupling losses and intrinsic (material) absorption losses of the EAM and the SSC sections. By using a lensed fiber having a mode diameter of 3 μm to couple light to and from the chip, we estimated a total insertion loss of 12 dB for our reflective devices (EAM biased at 0 V). The input/output coupling efficiency contributes a 2 dB loss in each direction (4 dB total), and the two-way EAM and taper absorption losses are 6 dB and 2 dB, respectively.

Insertion loss generally limits the system power budget which is a key parameter in WDM-PON systems. However, the presence of an SOA on chip with the EAM allows to fully compensate insertion losses and thus improve the system power budget. For wavelengths close to the peak of the SOA gain, we even obtain a net gain in the optical power. Figure 8.18 shows the fiber-to-fiber gain spectrum of a 300-μm-long O-band SOA in a reflective device configuration.

The results are obtained after all component-related losses are compensated by the SOA (only external losses are calibrated). The SOA gain peak is designed to be close to the zero-dispersion region. At 1310 nm, for example, the net device gain is >8.5 dB for an SOA current of 60 mA. The PIC also shows a net gain of >5 dB over a 25 nm range (1300 nm–1325 nm), which is another important device characteristic for colorless operation.

Fig. 8.18 Fiber-to-fiber gain spectrum of a 300-µm-long O-band SOA in RSOA-EAM



8.4.4.3 EAM Absorption and Extinction Ratio

The absorption strength of an EAM depends on both its cavity length and thickness of the active region (expressed in terms of optical confinement factor, Γ). In a reflective EAM-SOA configuration, the EAM absorbs the input light twice—before and after reflection. As a result, for a given EAM length and vertical structure, a higher extinction ratio can be obtained from a reflective EAM compared to its single-pass counterpart. The modulator's length also affects its bandwidth, but in a negative way. Hence, the EAM's length can be shortened when it is used in a REAM-SOA configuration to obtain high modulation bandwidth while providing acceptable extinction ratios.

The physical process that enables electroabsorption is a change in the absorption coefficient of the EAM (red shift) due to an externally applied electric field [33]. In a quasi-two-dimensional system such as MQWs, this process is known as quantum-confined Stark effect (QCSE). Since electroabsorption is directly related to the bandgap energy of the active region, the EAM's absorption increases with the increasing photon energy (decreasing wavelength). As a result, higher extinction ratio is obtained at shorter wavelengths.

Figure 8.19 shows static extinction ratio of the 100-µm-long EAM in a REAM-SOA configuration for different input wavelengths (1300 nm and 1320 nm). At 1300 nm, the device exhibits a very high extinction ratio of 16.5 dB between 0 V and −3 V. However, a stronger absorption at a shorter wavelength also means a reduced modulated output power. Hence, there is a tradeoff between extinction ratio and output power. The extinction ratios obtained from our devices are high enough even at longer wavelengths. For example, the extinction ratios at 1310 nm and 1320 nm are 12.5 dB and 10 dB, respectively. Since the SOA has no effect on the static performance of the EAM, there is no difference in static extinction ratios between REAM-SOA and RSOA-EAM configurations as shown in Fig. 8.19b. However, the position of the EAM along the optical path has a strong effect on its dynamic performance (bandwidth) as discussed in the following subsection.

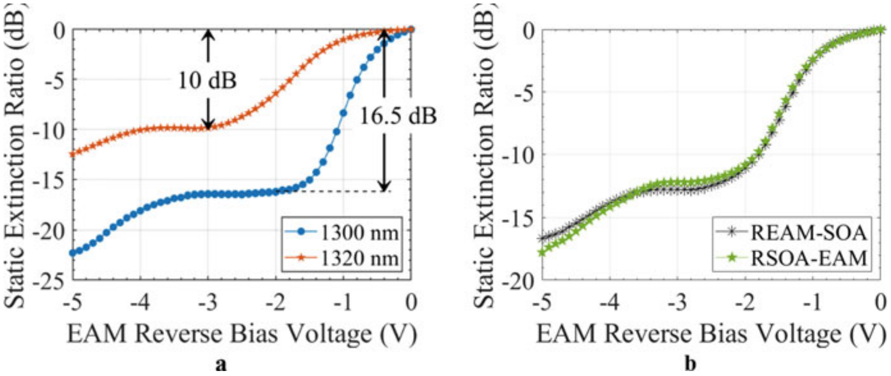


Fig. 8.19 Static extinction ratios of 100 μm O-band EAM: (**a**) in REAM-SOA configuration at 1300 nm and 1320 nm and (**b**) in both REAM-SOA and RSOA-EAM configurations at 1310 nm

8.4.4.4 Small-Signal Frequency Response

To estimate the modulation bandwidths of the EAMs in our PICs, we measured their frequency responses (S_{21} parameters) by applying a small modulation signal, typically <300 mV peak-to-peak voltage (V_{PP}). The electro-optic (E/O) bandwidth of an EAM is inversely proportional to its length because its bandwidth becomes RC-limited—parasitic capacitance increases with increasing EAM length. Similarly, bandwidth decreases as thickness of the active region increases because it takes a longer time to sweep out photogenerated charge carriers from the active region [33]. Although the RC limit is the dominant limiting factor to bandwidth, thickness of the active region must also be kept at minimum in order to reduce the required EAM bias voltage (reduce its power consumption). In a reflective device configuration, the effective EAM length is doubled. As a result, a reduction in the E/O bandwidth is generally expected compared to a single-pass configuration.

Figure 8.20 shows measured frequency responses of different EAM lengths. The 3 dB cutoff bandwidth of the 100 μm O-band EAM in REAM-SOA configuration is ~34 GHz as shown in Fig. 8.20a [$\lambda = 1310$ nm]. To demonstrate the effect of EAM length on its bandwidth, the frequency responses of a 150 μm and an 80 μm C-band EAMs are shown in Fig. 8.20b. The E/O bandwidth of the 150 μm EAM in REAM-SOA configuration is ~23.5 GHz, whereas that of the 80 μm EAM is still flat at 26.5 GHz (setup limit; estimated value is >40 GHz).

On the other hand, an EAM's bandwidth is significantly reduced, regardless of its length, when it is used in an RSOA-EAM configuration. This reduction is primarily due to the round-trip delay inside the SOA section. For example, the 3 dB bandwidth of the 100 μm O-band EAM in a RSOA-EAM is only 12 GHz (see dashed line in Fig. 8.20a). Therefore, for next-generation high-speed applications (e.g., ≥50 Gb/s NRZ), the REAM-SOA configuration with a reasonably short EAM is preferable.

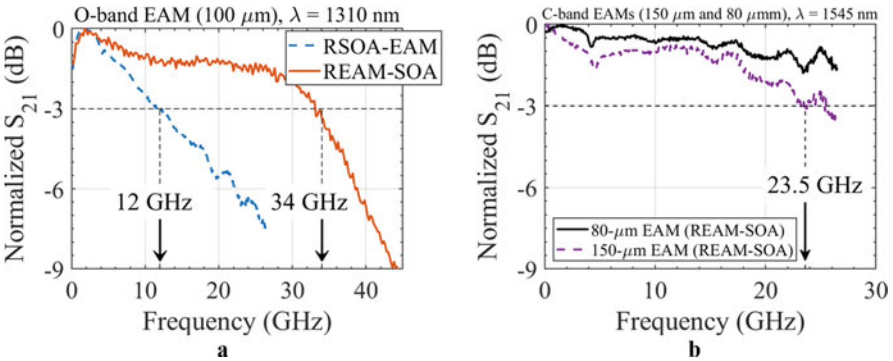


Fig. 8.20 Small-signal frequency responses of EAMs: (**a**) 100 μm O-band EAM in both REAM-SOA and RSOA-EAM configurations and (**b**) 150 μm and 80 μm C-band EAMs in REAM-SOA

8.4.5 System Demonstration

8.4.5.1 Colorless Transmission at 25 Gb/s NRZ

One important aspect of a reflective device configuration is the prospect of using the devices as colorless transmitters over a wide range of operating spectrum. As such, a single transmitter module can cover the entire width of a given WDM system without any additional tuning complexities. When mass produced, such devices can be used to realize low-cost transceivers, which is particularly important for large-scale 5G fronthaul deployment.

To demonstrate the colorless capabilities of our components at very high bit rates, we performed a preliminary test at 25 Gb/s NRZ, simply because it can be done with a standard bit error rate (BER) test setup.

For our experiments, the 25 Gb/s NRZ digital signal is generated by a signal quality analyzer with a bit pattern of 2^{31}–1 pseudorandom binary sequence (PRBS31). Then it is mixed with a DC biasing voltage via a Bias-T and applied to the EAM section to modulate the optical carrier. We obtained clearly open eye diagrams in both back-to-back (BtB) and 10 km configurations. At 1310 nm, the BtB dynamic extinction ratio (DER) obtained from the 100 μm O-band EAM in REAM-SOA configuration is ~10.2 dB when it is biased at −1.3 V with a 2.6 V_{PP} voltage swing. Similarly, the DERs at 1300 nm and 1320 nm are 13.5 dB and 8.1 dB, respectively. As expected, for the longer 150 μm C-band EAM, the DER is higher than that of the 100 μm EAM. For example, at 1530 nm, the DER is ~14.5 dB at 2.2 V_{PP}. The DER at 1545 nm is ~11.5 dB (−1.3 V_B/2.6 V_{PP}). All results are obtained at 25 °C.

Figure 8.21a shows BER performances of the 100 μm O-band EAM in REAM-SOA configuration at 1300 nm and 1320 nm ($\Delta\lambda = 20$ nm) when modulated with a 25 Gb/s NRZ signal.

Since chromatic dispersion is very low in this region, there is no dispersion penalty after 10 km transmission over a standard single-mode fiber (SSMF)—the

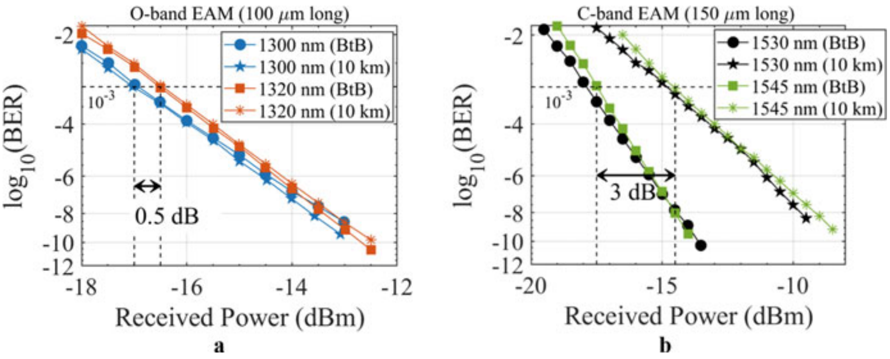


Fig. 8.21 BER performances (25 Gb/s NRZ): (**a**) 100 μm O-band EAM and (**b**) 150 μm C-band EAM. Both devices are in a REAM-SOA configuration

BtB and the 10 km BER curves overlap for each wavelength. Here, the transmission distance is rather limited by external losses (fiber attenuation, switches, filter, etc.). External amplification is not used during the experiment so as to demonstrate a *passive* optical network. The slight difference in the BER curves between 1300 nm and 1320 nm is attributed to the lower extinction ratio obtained at a higher detuning as shown earlier in Fig. 8.19 (>5 dB difference in extinction ratio). As a result, a slightly higher input power (<0.5 dB penalty) is required at longer wavelengths to obtain the same BER as the one obtained at shorter wavelengths.

On the contrary, chromatic dispersion in optical fibers is very high for the C-band wavelengths, and its effect on the transmission distance is clearly visible in the BER curves shown in Fig. 8.21b. For a 10 km transmission at 25 Gb/s NRZ, the dispersion penalty is ~3 dB between 1530 nm and 1545 nm ($\Delta\lambda = 15$ nm) for a BER of 10^{-3}. Therefore, to obtain the same BER as the one obtained at 1530 nm, we need 3 dB more input power at 1545 nm. Since the DERs at both wavelengths are very high (>11 dB), the penalty observed here is primarily caused by chromatic dispersion.

Nevertheless, such wide spectral ranges ($\Delta\lambda \geq 15$ nm) may not be required in practice for the devices to operate in colorless modes. For example, the spectral window defined for TWDM-PON is only ~2.5 nm wide, which is far lower than the range our C-band devices are tested. That means our components can be easily optimized to support the entire TWDM-PON spectral window without any significant performance degradation. The O-band devices, on the other hand, can support a 20-nm-wide WDM system with only a 0.5 dB tolerance requirement. Therefore, we can conclude that our reflective devices can support colorless operation at 25 Gb/s NRZ in their respective frequency bands.

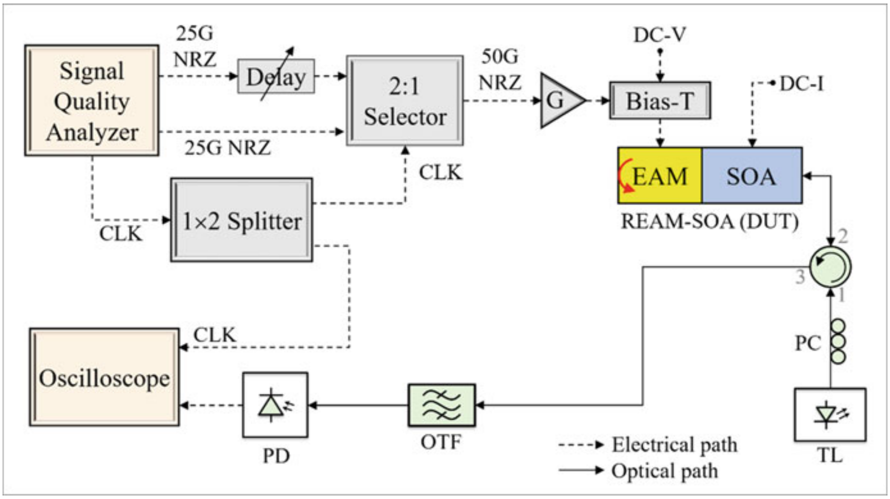


Fig. 8.22 Schematic of the experimental setup for 50 Gb/s transmission using a REAM-SOA

8.4.5.2 50 Gb/s Digital Transmission

The downlink transmission capacities of both TWDM-PON and 50G-EPON standards are based on multiplexing four (up to eight) or two wavelength channels, respectively [23, 24]. The latter reduced the number of wavelength channels from four to two to minimize cost associated with channel multiplexing. However, as technology matures, satisfying the required data rate with a single device will have the advantage of either saving cost or doubling the total bit rate for the same multiplexing cost. Similar to the 50G-EPON, ITU-T also selected 50G PON technology with NRZ modulation format as the focus for its next-generation high-speed PON (HSP) standard [34]. Thus, it is imperative to realize high-speed transmitters that can operate beyond the nominal 50 Gb/s line rates. High-speed EAMs operating up to 64 Gb/s NRZ are already demonstrated in a single-pass (transmitter) configuration.

To demonstrate the high-speed capabilities of our reflective devices, we modulated the 100 μm O-band EAM in REAM-SOA configuration with a 50 Gb/s NRZ signal. Figure 8.22 shows a schematic diagram of the experimental setup used to generate and transmit the 50G signal. The setup is customized to capture the 50 Gb/s eye diagrams only (without BER data). Further details about the device's dynamic characteristics including its 50G PAM-4 performance can be found in [31].

Two 25 Gb/s NRZ PRBS31 signals are generated by a signal quality analyzer and injected into a 2:1 selector module from III-V Lab. A delay line is connected to one arm of the input to interleave the two signals in time and generate a 50 Gb/s NRZ signal. Both the selector (transmitter side) and the oscilloscope (receiver side) are provided with a common clock source for synchronization by splitting the clock output of the signal quality analyzer. The 50G electrical signal is then amplified by a

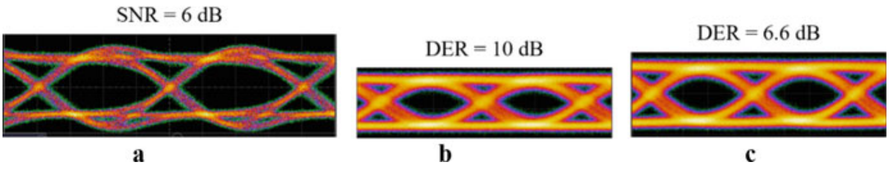


Fig. 8.23 (**a**) 50 Gb/s NRZ signal and (**b**) 50 Gb/s BtB eye diagrams at 1300 nm and (**c**) at 1320 nm

linear driver to obtain a peak-to-peak voltage swing of 2.4 V (eye amplitude). Figure 8.23a shows eye diagram of the 50 Gb/s NRZ electrical signal generated in this way.

The REAM-SOA chip is mounted on a high-frequency (HF) carrier with a ground-signal-ground (GSG) electrical connection provided to access the EAM. An external tunable laser (TL) is used to generate a continuous wave light. Since our components are sensitive to the polarization state of the input light, we used a polarization controller (PC) in order to inject transverse electric (TE)-polarized light into the chip. An optical circulator is inserted for bidirectional transmission over a single fiber. Finally, the RF signal and the EAM biasing voltage (-1.1 V) are combined inside a Bias-T and applied to the EAM. A DC current of 40 mA is applied to the SOA section to amplify the optical signal. At the receiver side, a tunable optical filter (OTF) of 2 nm bandwidth is inserted to suppress the amplified spontaneous emission (ASE) noise generated by the SOA. A wideband photoreceiver (PD) converts the optical signal to electrical signal, and its output is connected to a sampling oscilloscope to capture the eye diagrams.

Figure 8.23b–c shows the 50 Gb/s NRZ BtB eye diagrams obtained from the REAM-SOA at 1300 nm and 1320 nm, respectively. At 1300 nm, we obtained a clearly open eye diagram with a high DER of ~10 dB. The DER at 1320 nm is ~6.6 dB, which is slightly lower than expected. The maximum peak-to-peak voltage swing is limited to 2.4 V (setup limited). As a result, a fixed V_{PP} is used over the 20 nm range. Hence, the relatively lower DER observed at 1320 nm is attributed to the lower V_{PP} used during the 50G modulation. Dynamic performance of the EAM is not significantly degraded at 50 Gb/s compared with the one at 25 Gb/s NRZ. For example, at 1310 nm, the 50 Gb/s NRZ DER (which is ~9 dB) is only 1.2 dB lower than the DER at 25 Gb/s. Based on these observations, we also expect no significant degradation in BER performance at 50 Gb/s for the same 10 km transmission.

At this point, it is worth to mention that we recently demonstrated a 10 Gb/s multi-channel analog radio over fiber (ARoF) transmission using an RSOA-EAM satisfying 3GPP's requirement for 16-QAM signals [35]. Therefore, our devices can find application in several next-generation high-speed optical networks such as HSPs, data centers, and 5G fronthauling (analog and digital). Finally, further improvements in the future can be achieved by employing on-chip impedance matching techniques. Moreover, one can also focus on designing devices for uncooled operations, low polarization dependence, low bias voltage, and integrating array of REAM-SOAs to realize low-cost WDM transmitters.

8.5 Conclusion

In this chapter, a comprehensive study on mmWave ARoF systems is presented that focuses on the modulation and equalization approaches for enhancing the integrated fiber-radio link performance. In this context, the relevance of using mmWave ARoF technologies is highlighted for 5G and beyond is addressed in Sect. 8.1. However, mmWave ARoF still presents several challenges (such as nonlinearities, phase noise, or high-power losses) to be resolved. In particular, the choice of modulation format is pivotal toward ensuring spectral efficiency and reliability in the communication link. Therefore, Sect. 8.2 conducted a study on the link level modulation formats that are able to take a step toward attaining the stringent 5G KPI targets, where it was concluded that legacy OFDM was clearly not the modulation format of choice. Alternative schemes, such as SC-FDM and multi-CAP, should be considered.

Signal degradation introduced by optical fiber dispersion in mmWave ARoF systems is shown in Sect. 8.3. Moreover, the enabling technologies for mmWave ARoF are also explained. A mmWave ARoF simulation is performed over different fiber lengths. In these simulations, the LMS equalization is used to reduce the nonlinearity effects of the fiber and thus achieve reduced lower EVM values.

Optical amplifiers and modulators are crucial devices in mmWave ARoF systems. The characterization of a REAM-SOA integrated into a single chip is realized in Sect. 8.4. In this characterization, key parameters (such as extinction ratio, insertion losses, and gain) of this device are thoroughly investigated. Furthermore, an experimental digital transmission is demonstrated by utilizing this device, achieving bit rate of 50 Gb/s.

Therefore, this chapter serves to highlight the viability of mmWave ARoF links for 5G and beyond by analyzing techniques and methods to achieve high performance. Regarding the future lines of this work, Sect. 8.2 can be extended by changing the conditions of the experimental setup: different distances of the wireless link, NLOS propagation, or wireless channel with greater multipath effect. For Sect. 8.3, varying the parameters involved in the simulations is a straightforward way to continue with this part. Achieving high modulation orders such as 64-QAM and 256-QAM is one of the most relevant key challenges in both Sects. 8.2 and 8.3. Finally, for Sect. 8.4, future work can be achieved by employing on-chip impedance matching techniques.

Acknowledgment This research was funded by the European Commission through H2020 ITN 5G STEP FWD (grant agreement 722429) project.

References

1. Pérez Santacruz, J., et al. (2020). Candidate waveforms for ARoF in beyond 5G. *Applied Sciences, 10*, 3891.
2. 3GPP, FG IMT-2020, User Equipment (UE) radio transmission and reception; Part 2: Range 2 Standalone. 3GPP TS 38.101-2, version16.3.1, Release 16, Geneva, Switzerland, 2020.

3. Shunliang, Z. (2019). An overview of network slicing for 5G. *IEEE Wireless Communications, 26*(3), 111–117.
4. Rommel, S., et al. (2020). Towards a scaleable 5G Fronthaul: Analog radio-over-fiber and space division multiplexing. *Journal of Lightwave Technology, 38*(19), 5412–5422.
5. Fiorani, M., et al. (2016). Modeling energy performance of C-RAN with optical transport in 5G network scenarios. *Journal of Optical Communications and Networking, 8*(11), B21–B34.
6. Santacruz, J. P., et al. (2020). Experimental assessment of modulation formats for beyond 5G mm-wave ARoF systems. In *2020 European Conference on Networks and Communications (EuCNC), Dubrovnik, Croatia*, IEEE (pp. 300–304).
7. Delmade, A., et al. (2019). OFDM baud rate limitations in an optical heterodyne analog Fronthaul link using unlocked fibre lasers. In *2019 International Topical Meeting on Microwave Photonics (MWP), Ottawa, ON, Canada*, OSA (pp. 1–4).
8. Ishimura, S., et al. (2018). 1.032-Tb/s CPRI-equivalent rate IF-over-fiber transmission using a parallel IM/PM transmitter for high-capacity mobile Fronthaul links. *Journal of Lightwave Technology, 36*(8), 1478–1484.
9. Poole, C. D., et al. (1994). Optical fiber-based dispersion compensation using higher order modes near cutoff. *Journal of Lightwave Technology, 12*(10), 1746–1758.
10. Zulkifli, N., et al. (2006). Dispersion optimized impairment constraint based routing and wavelength assignment algorithms for all-optical networks. In *2006 international conference on transparent optical networks*, IEEE (Vol. 3, pp. 177–180).
11. Verma, L., et al. (2015). Backhaul need for speed: 60 GHz is the solution. *IEEE Wireless Communications, 22*(6), 114–121.
12. Dat, P. T., et al. (2014). High-capacity wireless backhaul network using seamless convergence of radio-over-fiber and 90-GHz millimeter-wave. *Journal of Lightwave Technology, 32*(20), 3910–3923.
13. Agrawal, G. P. (2002). *Fiber-optic communication systems* (3rd ed.). Wiley.
14. Aragon-Zavala, A., et al. (2011). Radio-over-fiber systems for wireless communications. *Recent Patents on Electrical Engineering, 4*(2), 114–124.
15. Sambaraju, R., et al. (2010). Up to 40 Gb/s wireless signal generation and demodulation in 75–110 GHz band using photonic techniques. In *2010 IEEE International Topical Meeting on Microwave Photonics, Montreal, QC, Canada*, IEEE (pp. 1–4).
16. Giannetti, F., et al. (1999). Mobile and personal communications in the 60 GHz band: A survey. *Wireless Personal Communications, 10*(2), 207–243.
17. Velez, F. J., et al. (2001). Frequency reuse and system capacity in mobile broadband systems: Comparison between the 40 and 60 GHz bands. *Wireless Personal Communications, 19*(1), 1–24.
18. Almasoudi, F., et al. (2013). Study of OFDM technique on RoF passive optical network. *Optics and Photonics Journal, 3*(2), 217–224.
19. Chen, L., et al. (2009). A novel scheme for seamless integration of ROF with centralized lightwave OFDM-WDM-PON system. *Journal of Lightwave Technology, 27*(14), 2786–2791.
20. Yu, J., et al. (2005). Seamless integration of an 8/spl times/2.5 Gb/s WDM-PON and radio-over-fiber using all-optical up-conversion based on Raman-assisted FWM. *IEEE Photonics Technology Letters, 17*(9), 1986–1988.
21. Elangovan, K. (2012). Comparative study on the channel estimation for OFDM system using LMS, NLMS and RLS algorithms. In *International conference on pattern recognition, informatics and medical engineering (PRIME-2012)*, IEEE (pp. 359–363).
22. International Telecommunication Union, ITU-R M.2083–0, IMT Vision–Framework and overall objectives of the future development of IMT for 2020 and beyond, 2015.
23. 40-Gigabit-capable passive optical networks 2 (NG-PON2): Physical media dependent (PMD) layer specification. ITU-T Recommendations G.989.2 (Amendment 2, 2017), 2017.
24. Knittle, C. (2020). IEEE 50 Gb/s EPON (50G-EPON). In *2020 optical Fiber communications conference and exhibition (OFC)*, IEEE (pp. 1–3).
25. Bonk, R., et al. (2015). The underestimated challenges of burst-mode WDM transmission in TWDM-PON. *Optical Fiber Technology, 26*, 59–70.

26. Lee, D.-H., et al. (2015). Design and performance of 10-Gb/s L-band REAM-SOA for OLT transmitter in next generation access networks. *Optics Express, 23*(3), 2339–2346.
27. Smith, D., et al. (2009). Colourless 10Gb/s reflective SOA-EAM with low polarization sensitivity for long-reach DWDM-PON networks. In *35th European Conference on Optical Communication, Vienna, Austria*, IEEE (pp. 1–2).
28. Lawniczuk, K., et al. (2013). 40-Gb/s colorless reflective amplified modulator. *IEEE Photonics Technology Letters, 25*(4), 341–343.
29. Zhou, X., et al. (2015). A 25-Gb/s 20-km wavelength reused WDM system for mobile fronthaul applications. In *2015 European Conference on Optical Communication (ECOC), Valencia, Spain*, IEEE (pp. 1–3).
30. Atra, K., et al. (2020). 25 Gb/s colorless transmitter based on reflective Electroabsorption modulator for ultra-dense WDM-PON application. In *Proceedings of 13th IMCL conference*, Springer (Vol. 1192, pp. 1089–1100).
31. Atra, K., et al. (2020). O-band reflective electroabsorption modulator for 50 Gb/s NRZ and PAM-4 colorless transmission. In *2020 Optical Fiber Communications Conference and Exhibition (OFC), San Diego, CA, USA*. IEEE.
32. Debrégeas, H., et al. (2018). Components for high speed 5G access. In *2018 Optical Fiber Communications Conference and Exposition (OFC), San Diego, CA, USA*, IEEE (pp. 1–3).
33. Fox, A. M., et al. (1991). Quantum well carrier sweep out: Relation to electroabsorption and exciton saturation. *IEEE Journal of Quantum Electronics, 27*(10), 2281–2295.
34. Zhang, D., et al. (2020). Progress of ITU-T higher speed passive optical network (50G-PON) standardization. *IEEE/OSA Journal of Optical Communications and Networking, 12*(10), D99–D108.
35. Atra, K., and al. *O-band Reflective SOA-EAM for 10 Gb/s Multichannel Analog Fiber-Wireless Uplink Transmission [submitted to ECOC 2020].*

Chapter 9
Optical Wireless System Performance, Deployment, and Optimization

Eugenio Ruggeri, Apostolos Tsakyridis, Christos Vagionas, Amalia Miliou, Shafiullah Malekzai, George Agapiou, George Datseris, and George Stavroulakis

Abstract As the 5G milestone approaches, there needs to be concerted effort towards practical deployment strategies and optimization in order to fully capitalize on the 5G benefits and KPIs (key performance indicators). This chapter aims to provide just that; the authors provide new insights on the performance of the optical wireless link and how this can be deployed as an integrated mobile architecture to provide small cell coverage to highly dense hotspot areas. Not only we investigate the system performance and practical deployment approaches, but we also address their optimization based on ML approaches. Virtual resource management and big data represent prominent 5G paradigms that offer new opportunities in terms of network management and optimization. In this chapter, we consider how ML can play a major role in network optimization by bringing intelligence to the limelight and by investigating how learned patterns or relationships between network states and QoE can be used towards intelligent prediction and optimization.

9.1 Introduction

We are on the verge of 5G deployment, where the increasing cellular traffic and the push towards broadband and delay-sensitive traffic are placing stringent requirement on the network KPIs as defined by expert alliances [1]. In order to face such a surge in data traffic head on, new mmWave technology is being deployed to unlock huge swathes of available spectrum [2] while at the same time enabling

E. Ruggeri (✉) · A. Tsakyridis · C. Vagionas · A. Miliou
Aristotle University of Thessaloniki, Thessaloniki, Greece
e-mail: eugenior@csd.auth.gr

S. Malekzai · G. Agapiou
OTE Academy, Athens, Greece

G. Datseris · G. Stavroulakis
Nessos Information Technologies S.A., Athens, Greece
e-mail: gdatseris@nessos.gr

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_9

massive multiple-input-multiple-output (MIMO) radio access achieved through beamsteering and spatial multiplexing that will provide the potential to achieve the 10 Gbps connectivity targets. However, to convert potential into a practical reality requires system optimization and validation through practical experimentation. In this context, Sect. 9.2 aims to take a step towards this direction by providing new insights on the experimental performance of a Fiber-Wireless (FiWi) link for converged FiWi/mmWave 5G networks, characterized by wide steering angle of 90° and multi-user support. The intermediate-frequency over fiber (IFoF)/V-band link under study constitutes a microwave photonics optical transmission stage and a wireless link employing a beam-steerable phased antenna array that is investigated in terms of frequency and linear response, beamsteering capabilities, and multi-user transmission targeting hotspot enhanced mobile broadband (eMBB) and dense fixed wireless access (FWA) 5G scenarios.

The small cell market is increasingly growing [3], while the development of software-based small cells running on open-source code is allowing the research community to perform quick implementation and testing by widely simplifying the building process of common radio technologies and maintaining compatibility with the legacy networks for fast deployment. Along with already developed software-defined networking (SDN) tools [4], major players such as Facebook are recently developing RAN hardware/software wireless access platforms towards further opening the current proprietary small cell interfaces. In Sect 9.3.1, we provide a comparative study between vendor-specific and open-source small cells that is referred to as the OAI (OpenAirInterface) that is commonly used for in-house testing.

As the radio access network (RAN) scales, fronthaul networks based on central offices, responsible for signal and data processing, are being deployed, either in a centralized or as a distributed architecture, that is referred to as C-RAN or D-RAN network approaches, respectively. Depending on the chosen approach, different requirements have been defined [5] in order to support fronthauling for all 5G network scenarios [6], where the enhanced mobile broadband (eMBB) use-case is the most challenging in terms of data rate, in particular for the conventional Common Public Radio Interface (CPRI) transport of digitized time domain IQ data. In this context, both enhanced CPRI (eCPRI) [7] and analog radio over fiber (ARoF) [8] schemes have been proposed as the fronthaul transport technology solution, with the first allowing for enhanced CPRI digital transport, while the latter technology unlocks high spectral efficiency and low computational loads inherent to the optical analog transmission, although demonstrating backward compatibility challenges with the already commercially deployed digital schemes. In Sect. 9.3.2, we target our analysis to the design of small cells for deploying real small cell networks in hotspot scenarios, taking in consideration the ARoF fronthaul and providing general design/deployment guidelines for such network portion, while in Sect. 9.3.3, we provide an overview on current backhaul solutions as part of integrated mobile network solution.

Migrating from integrated fronthaul-backhaul technologies towards a more global perspective of network performance, Section 9.3.4 provides a holistic view on network performance using the IxChariot application. The objective is not only to assess the viability of IxChariot as an experimental tool but to compare the OAI with vendor-specific equipment. IxChariot platform instantly assesses network performance, including wireless performance by using a simple server-client topology.

Given the expected huge increase in traffic, big data will be exploited in order to gain insight into network operation, as well as playing a pivotal role towards machine learning-based network management. According to [9], the data generated in case of cellular networks can provide the infrastructure provider (InP) with insight on their operation and maintenance needs, while machine learning (ML) techniques [10], defined as a combination of algorithms and computational statistics that enable a computer to perform actions based on sample data, will aid in further enhancing network security, optimization, and design. In Sect. 9.4, we aim to exploit 5G network data such as traffic patterns and QoS-related data (e.g., bandwidth) along with QoE parameters attained by users, in order to develop a new mapping based on ML between the input and output data, where the output state would be the equivalent network optimization parameters.

9.2 Physical Layer Performance for Fiber-Wireless Links

We present an experimental evaluation on a Fiber-Wireless (FiWi) link for converged FiWi/mmWave 5G networks, characterized by wide steering angle of 90° and multi-user support. The IFoF/V-band link under study, comprising a microwave photonics optical transmission stage and a wireless link employing a beam-steerable phased antenna array, is investigated in terms of frequency and linear response, beamsteering capabilities, and multi-user transmission performance towards hotspot enhanced mobile broadband (eMBB) and dense fixed wireless access (FWA) 5G scenarios.

9.2.1 Fiber-Wireless Link Featuring Beam-Steerable Phased Antenna Array Approach

The used experimental setup of the investigated system is shown in Fig. 9.1a, including one portable V-band horn transmitter (Horn Tx) with 22.5 dBi gain, featuring an I/Q modulator, a 10 GHz local oscillator (L.O.), and an upconversion stage to generate a 60 GHz wireless signals of 10° beamwidth. The portable transmitter is used hereby to emulate the traffic of the endpoint user terminal towards the 32-element phased antenna array receiver (PAA Rx) by Siklu followed by the IFoF link. The data traffic is generated by a Keysight arbitrary waveform generator (AWG) M8190A and fed to the Horn Tx in a differential data input configuration

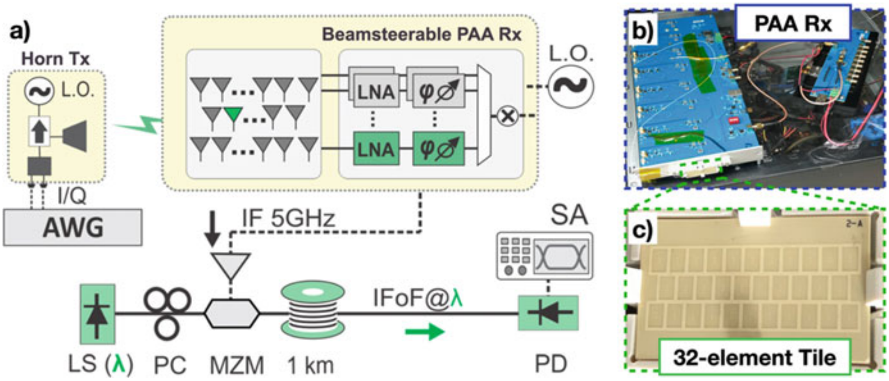


Fig. 9.1 (**a**) Experimental setup. (**b**) Phased antenna array (PAA) receiver along with (**c**) zoom in of the 32-element Tile PCB

(I/ $\bar{I}$ and Q/ $\bar{Q}$). At first, while performing the static characterization, the Tx is fed with either one or two tones and, then, is fed by two different modulation formats of QPSK and QAM16 signals when performing data transmission over the end-to-end FiWi channel. At last, for the multi-user experimental demonstration, two additional V-band transmitters are employed within the same setup, as described in Sect. 9.2.3.

The 60 GHz signals transmitted by the Horn Tx are received after 1 m V-band link by the phased array Rx antenna, composed by a Tile PCB and Feed Board PCB, both supplied with 5.3 V and drawing 1.2A and 300 mA, respectively. The Tile PCB is integrated on a low-temperature ceramic and hosts the 32 radiating elements, each of which comprises a dipole with a 6dBi gain featuring almost isotropic radiation across a 120° sector, and it is placed at the front panel of the antenna system. The 32 elements are then followed by an RF IC with 32 channels each featuring a series of a low-noise amplifier (LNA), a phase-shifting element (φ), and an integrated downconversion stage, before combining all 32 channel outputs using a 32:1 combiner. The Tile Feed Board PCB also takes input from an external 10 GHz L.O. with an RF input power of −10 dBm, in order to generate a single 5GHz IF electrical output that carries the initially transmitted data. A photo of the boards of the phased antenna Rx system during the experiment is shown in Fig. 9.1b, while the inset Fig. 9.1c shows the 32-element Tile placed at the front panel of the antenna. When all elements are on, they can be configured to constructively interfere over the air (OTA), concentrating the radiation in a tightly focused beam across a certain direction, that can be steered across a 120° direction, while the signal distortion across 90° range was experimentally found to be negligible, as discussed in Sect. 9.2.2.

The wireless link is then extended with a fiber-optical transmission over 1 km of standard single-mode fiber (S-SMF). Specifically, the -2 dBm (0.5Vpp) 5GHz data output of the Rx antenna is amplified to 5 V by a driver amplifier before driving a zero chirp LiNbO3 Mach-Zehnder modulator (MZM), biased at the quadrature point. The MZM is then modulating an optical carrier at $\lambda = 1550$ nm generated by a DFB laser source (LS), imprinting the wirelessly received RF signals onto the wavelength (λ) for IFoF transmission. At the end of the 1 km fiber link, the signal is o-e converted by a 10G InGaAs avalanche photo-receiver (APR) with 0.7A/W responsivity, and its output is captured by a Keysight signal analyzer (SA) for monitoring purposes.

9.2.2 Fiber-Wireless Link Characterization

Initially, the frequency response of the full FiWi link is presented, showing a 3 dB bandwidth of more than 1 GHz, along with two-tone measurements separately for either the optical or the wireless channels, revealing IIP3s of +8.1 and − 2.4 dBm, respectively, for each section. Following the static characterization, single-user data transmission of 300 MBaud QPSK and 250 MBaud 16-QAM signals is evaluated in an uplink scenario, reaching up to 1 Gbps user data rate, and the beamsteering capabilities are characterized across a 90-degree sector.

9.2.2.1 Static Characterization: End-to-End Frequency Response and Two-Tone Measurement

First, the end-to-end FiWi link was statically characterized in terms of supported bandwidth and available frequency spectrum, by transmitting a single tone from the Horn Tx to the PAA Rx and sweeping its carrier frequency while measuring the received power after the APR in the SA corresponding to the frequency of the received IF at the output of the APR. The overall tested bandwidth was 3 GHz, with the tone frequency at the output of the Rx starting from 3.2 GHz and swept up to 6.2 GHz, which is translated in sweeping of the actual wirelessly transmitted mmWave carrier over the air (OTA) from 58.3 to 61.5 GHz. Figure 9.2 illustrates the received power versus frequency plot, revealing a gain peak of −16 dB around 4.5 GHz central frequency, while the 3 dB bandwidth of the FiWi link was measured to be above 1GHz.

Following the frequency response, in order to evaluate the linearity of the channel and identify the operational conditions, we performed two-tone measurements selectively for the V-band wireless link and the fiber-optical one [11]. The two-tone measurement of the wireless link aimed to evaluating mainly the impairments induced by the whole signal processing chain, including the upconversion, modulation, and amplification stages of the portable Horn Tx as well as the LNA, phase shifter, and downconversion stage of the phased array antenna. Thus, in order to simplify the measurements of the phased array antenna comprising 32 channels, we activated only 1 radiating element and 1 channel of the RF-IC. However, we expect the response to be the similar for all channels as they are featuring identical components and signal chain.

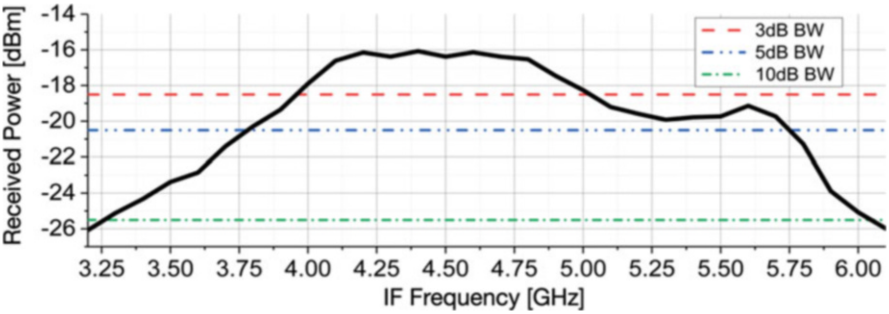


Fig. 9.2 Frequency response of the end-to-end FiWi link, featuring 3 dB bandwidth of more than 1 GHz

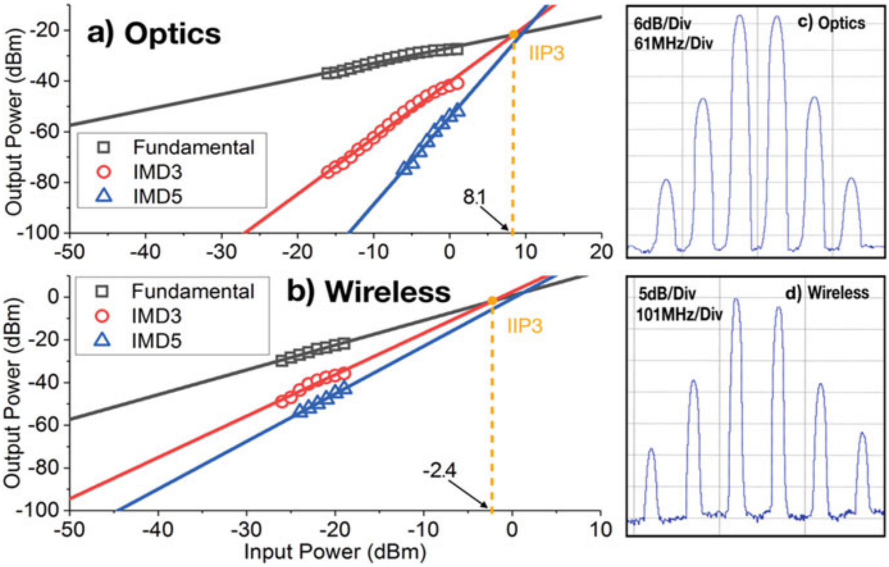


Fig. 9.3 Two-tone measurements of (**a**) optical and (**b**) wireless sections. (**c**) Optical and (**d**) wireless RF spectra captured in high-saturation regime

Specifically, the results for the optical link are shown in Fig. 9.3a, which depicts the two-tone measurements of the link with square, circle, and triangle dots corresponding to the fundamental curve and the third- and fifth-order intermodulation distortions, respectively, while the RF spectrum is illustrated in Fig. 9.3c. Equivalently, Fig. 9.3b shows the two-tone measurements for the wireless link, while the RF spectrum of the two-tone measurement for the wireless link is depicted in Fig. 9.3d. After plotting the dots of the measurements for the fundamental, IMD3 and IMD5, and extrapolating the curves, we aimed to evaluate the intercept point of the extrapolation of the fundamental and the IMD3 curve, widely referred to as the third-order intercept point (IP3) [11]. The input power on the horizontal

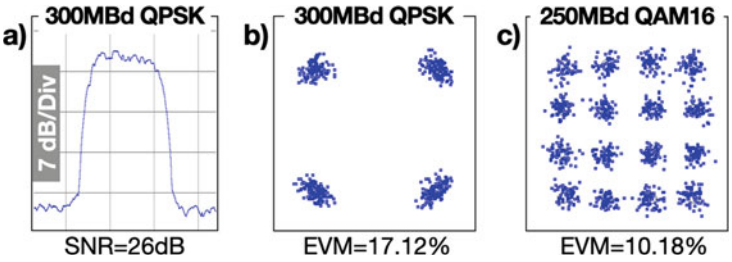


Fig. 9.4 Dynamic characterization featuring (**a**) spectrum and (**b**) constellation of the 300 MBaud QPSK received signal and (**c**) constellation diagram of the 250 MBaud 16-QAM received signal

axis corresponding to our measurements (third-order input intercept point, IIP3) is −2.4 dBm for the wireless link and 8.1 dBm for the optical link, while the inclination of the red IMD3 curves was also found to be higher for the optical link rather than the wireless link. Consequently, the undesired saturation regime is reached by the wireless section several dBs before affecting the optical channel. This evaluation reveals how a higher distortion stems from the radio link rather than the optical, attributing the higher impact on the transmission impairment to the wireless section of the V-band FiWi link.

9.2.2.2 Dynamic Characterization: Link Data Rate Performance with Beamsteering

Aiming to evaluate the FiWi link at various user data rates envisioned for 5G use-case scenarios and KPIs [1], varying from 50 Mbps for hotspot environments up to 1 Gbps foreseen for broadband access in indoor areas, FiWi transmissions of different modulation formats and beamsteering angles were performed. In this case, all 32 antenna elements were activated and configured to steer initially at 0°, while the modulation formats were varied from 300 MBaud QPSK to 250 MBaud 16Q-AM, resulting, respectively, in data rates of 0.6 Gbps and 1 Gbps, thus meeting the 5G KPI of 1 Gbps. The recorded RF spectrum for the 300 MBaud QPSK is shown in Fig. 9.4a, occupying a spectral bandwidth of around 400 MHz and revealing a high SNR of 26 dB, while similar spectrum was recorded for the 16-QAM transmission. The constellation diagrams of the signals and the respective EVM values obtained for the two single-user uplink scenarios are shown in Fig. 9.4b and c, indicating EVM values of 17.12% and 10.18%, respectively, as recorded at the SA and without applying any additional digital signal processing, achieving a maximum spectral efficiency of 3.85 b/s/Hz. It is worth noting that all measured EVM values are below the respective requirements of 17.5% and 12.5% set by 3GPP for new radio transmission and reception [12].

Following the dynamic characterization, the FiWi link was also benchmarked against high data rate coverage at various angles within a complete 90-degree sector

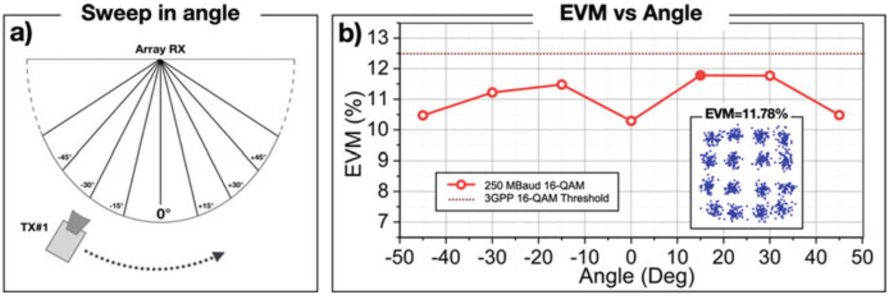


Fig. 9.5 (**a**) Schematic describing how the captured measurements were acquired; (**b**) experimental results for single-user uplink of a 250 MBaud 16-QAM signal at beamsteering angles from −45° to +45° with a step of 15°. The constellation diagram corresponding to the worse case of 11.78% EVM at +15° is shown in the inset

while employing the 1 Gbps 250 MBaud 16-QAM signal. The PAA system was configured to steer its beam across a 90-degree range, from −45° to +45°, using a step of 15° and resulting in seven different angular positions of the portable terminal transmitter, as illustrated by the schematic of Fig. 9.5a. The 15°-step value was chosen as it is comparable to the beamwidth of 10° and experimentally the minimum value to appreciate consistent performance variation. The same signal was transmitted by the portable Horn Tx for all angles while keeping a constant radius distance from the PAA of 1 meter, in order to make the EVM performances comparable. The plot depicting the captured EVM values versus angles is shown in Fig. 9.5b, exhibiting EVM values below the 12.5% requirement [12] within the angle range of −45° to +45°, with an average EVM of 11.07%.

Moreover, it is worth noting that the reported 1 Gbps data transmission verifies that high data rate can be achieved for a single-user data stream that can be delivered anywhere within a 90° sector when beamsteering the antenna, meeting the most demanding 5G KPI for user rates envisioned for indoor broadband connectivity of FWA services [1].

9.2.3 Fiber-Wireless Multi-user Experimental Demonstration

Finally, we present a true multi-user experimental demonstration, employing three end user portable transmitters (namely, TX#1, TX#2, and TX#3) scattered within a 90-degree angular sector, taking advantage of linearity measurements, 1 GHz available bandwidth, and enhanced beamsteering performance. Specifically, the three end user terminals are employed to simultaneously transmit either on the same carrier frequency, demonstrating frequency-reuse enabling spatial-division multiplexing by means of beamsteered reception, or on three separate spectral frequencies by employing frequency-division multiplexing along with isotropic reception, comparing the two approaches.

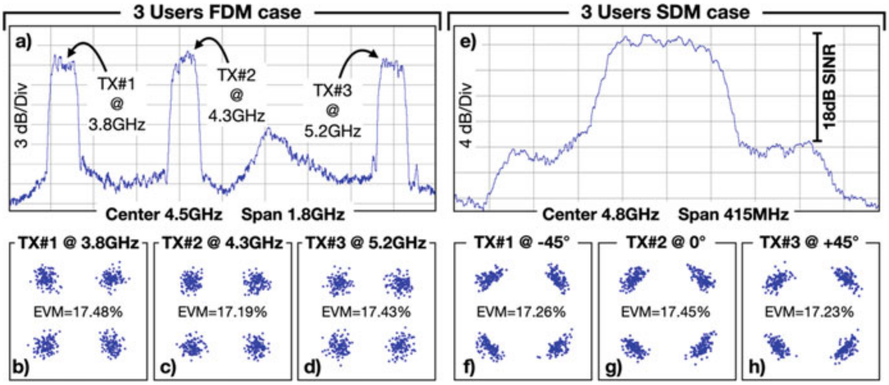


Fig. 9.6 Experimental results for the multi-user transmissions including, for the FDM scenario, (**a**) spectrum of the three distinct data signals around the received IF frequency of 5 GHz (corresponding to 60.5 GHz in mmWave) and (**b c d**) constellation diagrams of each simultaneous transmission and, for the SDM scenario, (**e**) spectrum of the three superimposed data signals at the received IF frequency of 5 GHz (corresponding to 60.5 GHz in mmWave) and (**f g h**) constellation diagrams of each simultaneous transmission

Multi-user uplink communication was evaluated for two different multiplexing schemes, i.e., FDM and SDM. In both cases, the FiWi link was tested having the three V-band transmitters simultaneously emitting from the −45°, 0°, and + 45° angular positions. Due to different internal amplification stages of the three portable transmitters, the I/Q data input signal powers for TX#1, TX#2, and TX#3 have been adjusted in order to transmit at equal power levels.

In the FDM case, the three transmitters are simultaneously radiating at the three mmWave frequency carriers of 59.3 GHz, 59.8 GHz, and 60.7 GHz (i.e., IF frequencies of 3.8 GHz, 4.3 GHz, and 5.2 GHz), as shown in the received RF spectrum in Fig 9.6a. The PAA is configured in isotropic mode by activating just a single element of the array, accommodating for all users scattered within a 120° sector. After the fiber propagation and the opto-electric conversion at the APR, the three constellation diagrams are shown in Fig. 9.6b–d, obtained by adjusting the monitoring bandwidth of the SA selectively at each of the three frequency carriers, while all transmissions were taking place simultaneously. The EVM values recorded for each of the three 100 MBaud QPSK transmitted signals were 17.48%, 17.19%, and 17.43%, respectively, as obtained by the SA and without applying any additional DSP (digital signal processing), all still satisfying the 3GPPP EVM threshold of 17.5% for QPSK [12].

In the SDM case, the three transmitters are still operating simultaneously, but this time over the same frequency carrier, i.e., 60.3 GHz (IF frequency of 4.8 GHz). The reception is thus performed with the PAA configured in beamsteering mode, achieving spatial division by activating all 32 elements of the array and then tuning the 32 respective Tile PCB phase shifters so as to steer the main beam in the desired direction and point at TX#1, TX#2, and TX#3 placed at −45°, 0°, and + 45°, respectively. Figure 9.6e shows the spectrum of the three superimposed signals, as received at the SA, after the end-to-end FiWi transmissions. The obtained corresponding constellation diagrams are shown in Fig. 9.6f-h, along with the respective EVM values of 17.26%, 17.45%, and 17.23%, demonstrating that the multi-user SDM operation can be still performed within the required 3GPP threshold values for each transmitter [12], regardless of the simultaneous reuse of the exact same wireless and optical intermediate-frequency channel. The presented SDM scenario, simultaneously carrying three 100 MBaud QPSK signals, showed an average EVM penalty of less than 0.2% with respect to the 300 MBaud single-user transmissions of Fig. 9.4c, achieving the same aggregate data rate but allowing three different users with a reduction in spectrum allocation, all with a negligible penalty. The main source of this penalty comes from the additional phase noise visible in the constellation diagrams of the beamsteered transmissions, attributed to the use of multiple phase-shifting elements. Additionally, the RF spectrum of the final received signal for the SDM case features a signal-to-interference and noise ratio (SINR) of 18 dB, as shown in the measurement in Fig. 9.6e.

In conclusion, three-user uplink communication offering the same aggregate data rate of 0.6 Gbps, being successfully accomplished exploiting either frequency- or spatial-division multiplexing, experimentally reveals that both FDM and SDM can allow for EVM performance within the 3GPP KPI limits for QPSK modulation schemes. This analysis confirms that FDM and transmission over three different carrier frequencies can be activated when a single 5G antenna has to serve multiple users within the same small cell, while the availability of multiple mmWave massive MIMO antennas can be utilized for SDM communication when V-band and IFoF frequency reuse in higher user density 5G small cells have to be employed.

9.2.4 Conclusions

A Fiber-Wireless IFoF/V-band link has been presented and experimentally investigated for hotspot enhanced mobile broadband (eMBB) and dense fixed wireless access (FWA) 5G scenarios. Comprising of a microwave photonics transmission stage and a beam-steerable V-band phased antenna array, experimental results show more than 1 GHz 3 dB bandwidth and capabilities to steer the main beam in a 90° range while meeting the 5G KPIs of 1 Gbps. Additionally, a multi-user transmission was validated in an experimental setup comprising three standalone portable V-band transmitters, exploiting frequency- or spatial-division multiplexing for carrier aggregation or spatial selectivity, respectively.

9.3 Practical Integrated Optical Wireless Architectures and Performance

The optical wireless architecture and performance in both the backhaul and fronthaul part will significantly shape 5G standards and will provide the potential for the requested 1000 times increase in spectral efficiency and 90% reduction of energy consumption. Specifically, in this section, we target our analysis to the design and optimization of small cells for deployment in real hotspot scenarios. In particular, this section aims to provide insights on:

- Comparison between vendor-specific small cells and emulated prototypes such as OpenAirInterface (OAI), which are more suited for testing in the lab and experimental field trials
- Design and analysis of networks based on wireless access and optical fronthaul/backhaul
- Analysis of backhaul systems for small base stations
- Performance analysis by using professional tools like IxChariot in order to measure performance KPIs

9.3.1 Comparative Study: Vendor-Specific vs. Open-Source Small Cells

Mobile operators are spending significant efforts to meet the demands of the market and customers. In this context, 5G is perceived as the enabler to deliver high-speed services and capacity through the deployment of small cell technology to efficiently offload traffic from the macrocell base stations. In fact, the small cell market is increasingly growing from $11.5 billion in sales in 2018, and it is forecasted to reach $ 52 billion in 2025 [3]. This has led to a contention between vendor-specific and open-source small cells.

The leading vendors in the small cell market have invested in high-quality technology and processes to develop leading-edge monitoring and digital triggering activation capabilities. The vendor-specific small cells have become progressively expensive and provide different APIs (Application Programming Interfaces) to support different customer-specific requirements. However, it is clear these small cells may not be viable and may be limited in use for testing and lab purposes – they would indeed be a costly option. On the other hand, open-source small cells are mostly used for testing purposes and are easy to use to implement very complex test configurations. For example, before the advent of software OpenBTS and OpenBSC, many people in the lab environment would have deemed GSM technology to be too complex and misinterpreted. As a result of the efforts of open-source developers and engineers, it is now possible to create a GSM/LTE network with a portion of the cost of standard solutions and to integrate them into open-source GSM/LTE technologies within existing networks, efficiently saving significant cost and many other benefits [4].

In the OAI platform, the focus is on the spectral, algorithmic, and protocol efficiency research not only to implement demonstrators which are based on high-performance embedded architectures but also to test, validate, and analyze wireless systems; on the other hand, OAI does not focus on developing solutions that are deployment-ready [4].

EURECOM the graduate school and research center is leading the development of OAI; the hardware is available at a cost to partners and R&D projects across the globe. From the development of the wireless communication, there is a need for open-source software in 5G. The hardware/software for radio access network (RAN) is made of large numbers of proprietary elements that reduce the scope for innovation and increase the costs for TSPs (telecommunications service providers) to set up new services/applications in an absolute dynamic cellular network. Open-source software requiring some special processors can efficiently reduce cost, can increase flexibility, are capable of simplifying network access, can improve innovation speed, and can expedite the time to market for introducing new services.

There is already a movement going on within the industry on the development of software-defined networking (SDN) concepts to open the proprietary interfaces to control the RAN hardware/software. Recently, Facebook also designed and tested an open-source and cost-effective software-defined wireless access platform called OpenCellular that aims to improve connectivity in remote areas of the world [13].

The Opensource concept has been highly successful in many areas of software. The source code runs on MySQL and Linux – both developed by volunteers from around the world. Source code is published and any improvements made are also shared with the community. Most of the successful projects have a commercial business coordinator that are funded, and they provide support for those organizations that want to pay for it [14]. We use the OAI platform for subsequent experimental studies in Sect. 9.3.4.

9.3.2 Integrated Optical Wireless Fronthaul Networks: Design and Performance Analysis

Fronthaul networks are designed based on the central offices (COs) positions, the COs are responsible for signal and data processing, thus they can be placed centrally, e.g. in regional towns, or in a distributed way, i.e. in each town/village where a small CO is located serving the antenna sites in the very close proximity. In the case of distributed CO placement, the connections from the local telecom nodes carry the appropriate data streams over the existing fiber and transmission infrastructure. 10 Gbit/s and 25 Gbit/s OOK or PAM-4 links can be utilized for the considered distances (20–30 km) due to their characteristics. Conventionally, in 4G wireless networks, the fronthaul links are using CPRI/OBSAI protocol which is between RF and the remaining L1/L2/L3 functions.

The transport of digitized time domain IQ data on the conventional fronthaul requires unreasonably high transport capacities for very-high-capacity applications, such as enhanced mobile broadband (eMBB), or for radio sites with many independent antenna elements (massive MIMO or multi-layer MIMO). This results in transport latencies between the remote unit (RU) and distributed unit (DU)/central unit (CU) of up to a few hundred microseconds. Table 9.1 shows the approximate data rates for time domain IQ data fronthaul (CPRI rates without line coding) needed to support various radiofrequency bandwidths and number of antenna ports in wireless networks using parameter ranges given by 3GPP in [15].

Generally, in 5G there are two deployment scenarios for 5G NR fronthaul networks depending on the location of the DU [15]:

Centralized RAN (C-RAN) when the DU is centralized in a small access room or an access convergence room, as shown in Fig. 9.7a. The distance between DU and RU is around 10 km or less, in this scenario point-to-point fiber is used, and the direct connection may require a large number of trunk fiber resources.

Distributed RAN (D-RAN) when the DU is deployed in the base station room or the RU/DU/CU are integrated and deployed in a base station as shown in Fig. 9.7b. In this scenario, the distance between DU and RU is generally very short such that a direct point-to-point fiber connection is suitable for the fronthaul transmission.

Furthermore, the C-RAN deployment can be divided into two categories, large concentration and small concentration, as shown in Fig. 9.8. For the large concen-

Table 9.1 Required fronthaul data rates in 5G wireless network [15]

	Radio channel bandwidth			
Number of antenna ports	10 MHz	20 MHz	200 MHz	1 GHz
2	1 Gb/s	2 Gb/s	20 Gb/s	100 Gb/s
8	4 Gb/s	8 Gb/s	80 Gb/s	400 Gb/s
64	32 Gb/s	64 Gb/s	640 Gb/s	3200 Gb/s
256	128 Gb/s	256 Gb/s	2560 Gb/s	12,800 Gb/s

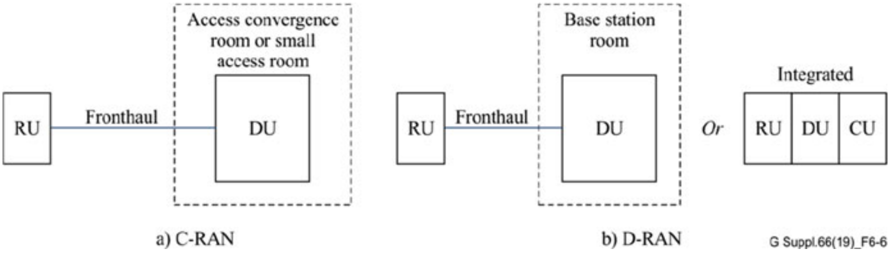


Fig. 9.7 RAN deployment scenario schematic diagram [15]

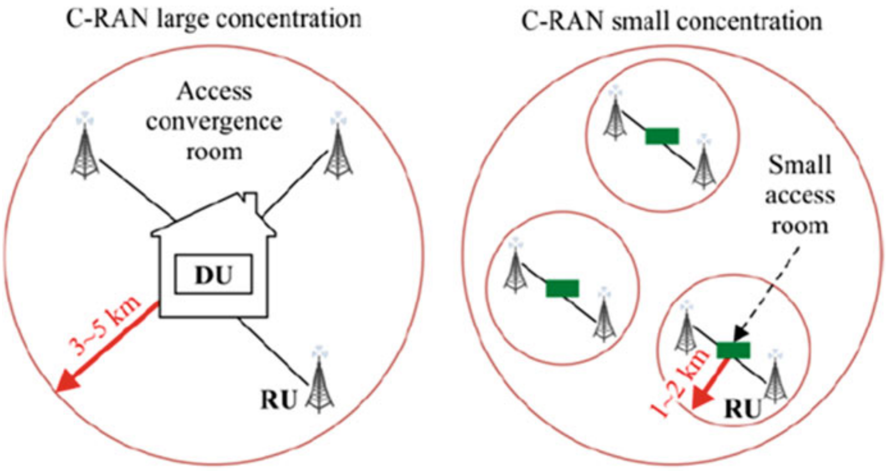


Fig. 9.8 Two categories of C-RAN deployment [15]

tration mode, DU is generally deployed in the access convergence room, whereas for the small concentration mode, DU is in small access rooms.

Fronthaul Design Guidelines for Practical Deployment Scenarios
To highlight the fronthaul design requirements for a typical hotspot coverage scenario, we should refer to a real-life sporting event as a case study that acts as a basis for building more general guidelines. In this context, we refer to a real-life example in the form of football match within a crowed stadium.

9.3.2.1 Sports Event Use-Case Scenario: Football Event

Considering the abovementioned fronthaul design requirements, a real scenario is being employed to justify the high bit rate required for a fronthaul network to carry. This high bit rate is required by deploying a crowded stadium which needs to support a high bit rate by a local telecom company during a football match in Athens, Greece.

The total capacity of the stadium is 32,000 seats; the specific match was attended by $U = 32.000$ spectators. The stadium was almost full. Note that the start time was at 18:00 and the end time was almost at 23:45 (cf. Figure 9.9). Furthermore, it was observed that the maximum number of active users was $U_a = 45,437$ (where the outside users are also included); for this specific assumption, we have to measure the user density long before the match started (14:00), and based on Fig. 9.9, near to 15,000 users are observed around the stadium with specific coverage. Roughly, the active users are close to the number of the spectators. As it is seen from Fig. 9.10, the number of active users peaks at the time when the match is started. In order to fulfill the clients' demand, downlink (DL)/uplink (UL) traffic volume which is

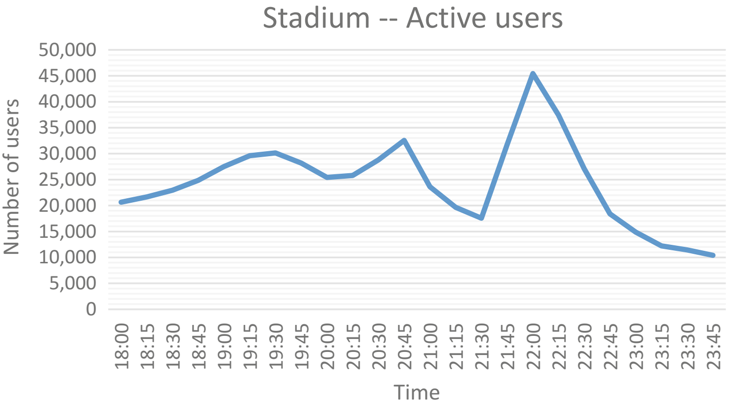


Fig. 9.9 Number of active users during the match inside stadium

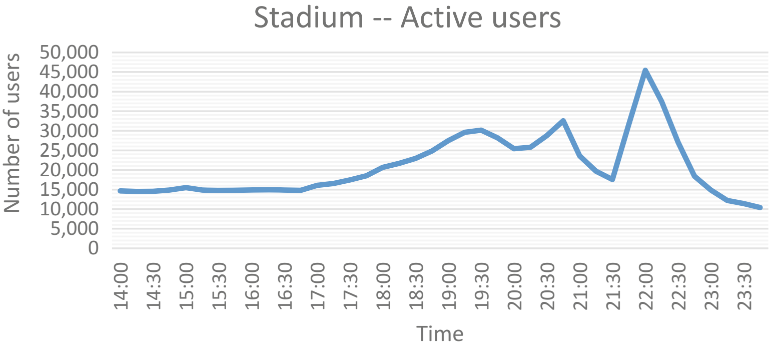


Fig. 9.10 Number of active users outside the stadium before and during the match inside stadium

foreseen to increase by 100 times in the upcoming years aims to offer services like live streaming high-definition (HD) videos and replay moments of the current game. In this context, the available resources cannot adequately serve the clients' needs. Therefore, it is a must to design and deploy an optimized infrastructure considering the fronthaul design requirements for the fronthauling/backhauling in terms of throughput and energy- and cost-efficiency. In this specific example, the 3.5 and 26 GHz bands have been used, and the network architecture is shown in Fig. 9.11. Furthermore, the football stadium with the antenna positions, beams, and fronthaul network placement to provide full stadium coverage is shown in Fig. 9.12.

9.3.2.2 Fronthaul Design Guidelines

Considering large event locations such as stadium and music, sport, and event halls, the crowd number can change from 1000 (music hall) to 100,000 (big football

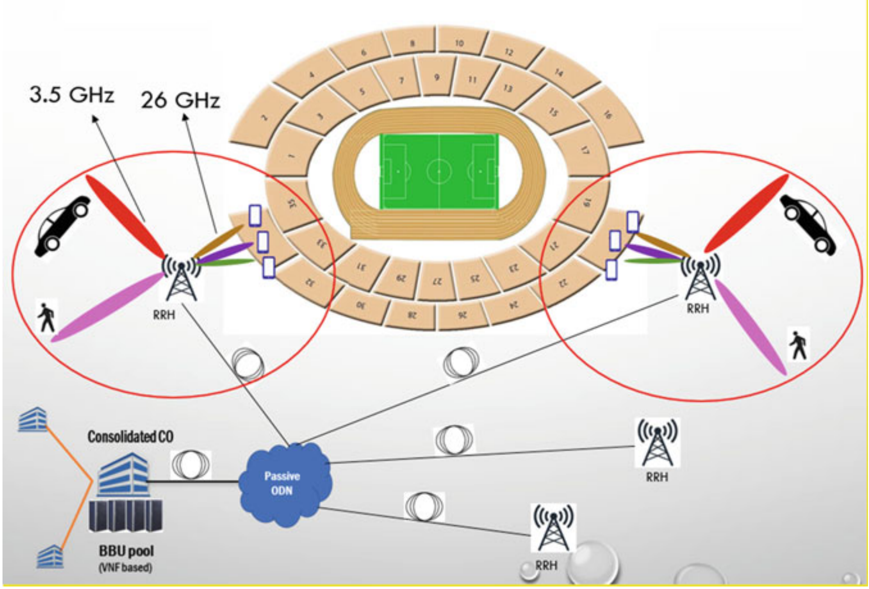


Fig. 9.11 Exemplary football stadium with the antenna and beam [16]

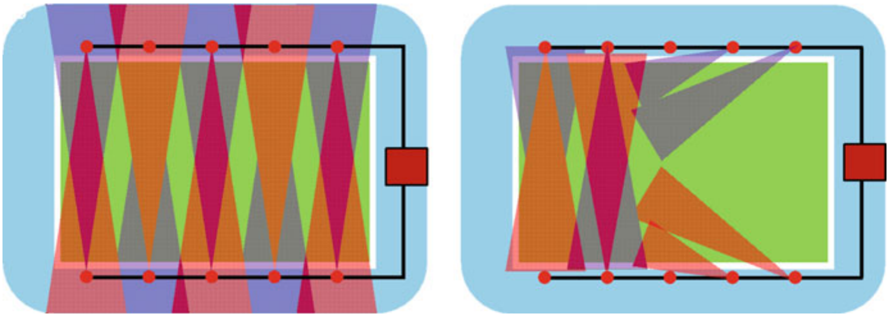


Fig. 9.12 left figure shows our exemplary football stadium with the antenna positions, beams, and fronthaul network placement to support full stadium coverage; the right figure shows the beam adjustment for the concert use-case [16]

stadium), with extreme cases for outdoor concerts and festivals where up to a million people can gather for a single entertainment act. This scenario leads to the following fronthaul design requirements:

- High demand in capacity from the crowd size as well as the high bandwidth application utilization in a limited area requires multiple antenna sites to be installed. Such densely packed antenna sides can be served from the local CO.
- Due to the limited and crowded area, the event venue fronthaul network will be characterized by the very large antenna point number as well as limited in space

distributions. In order to optimize the cost and network performance, it is highly advisable to install the CO at the event locations or nearby [16]. Installation of CO with edge computing capabilities allows the realization of ultra-low latency services, not only for entertainment purposes but also for security services.

- Regarding the choice of fiber, multi-core fiber (MCF) has advantages since it can cover the required distances without active skew compensation and also the high bit rate capacities which are required in crowded areas, which can potentially limit the system complexity and cost, as well as the installed cable thickness; the latter increases cable flexibility that is critical in the case of indoor installations [16].

Moreover, the optical fronthaul design refers to the connection between the CO and several RRHs. Typically, it follows the design rules of a passive optical network (PON), or it can be seen as part of a PON infrastructure (see Fig. 9.13).

Two options are considered for the type of fiber infrastructure of the optical fronthaul:

(i) Multiple single-mode fiber (M-SMF): This type is mostly used by all major operators.
(ii) Multi-core fiber (MCF): This type of the fiber infrastructure links is believed to be the compact high-capacity alternative for future capacity expansions in optical networks.

Each of the above poses different attributes to the network design by determining the type of technology solution that can be implemented, as well as the future expandability in terms of capacity and cost.

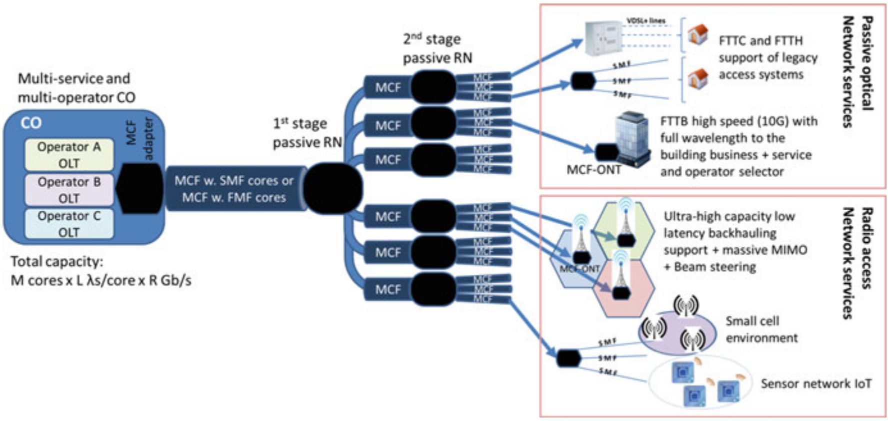


Fig. 9.13 Co-location of radio access and passive optical network services over the SDM-enabled infrastructure [16]

9.3.3 Analyzing the Backhaul System

Next-generation communication technologies and applications will also require high-capacity backhaul systems in order to have a seamless end-to-end delivery. Therefore, we also need to revisit the current backhaul solutions that require different technology options depending on the deployment scenario.

Wireless Backhaul Solutions
Small cells are connected to each other and the supporting core network. In fact, the same infrastructure which provided the operator's macrocell network is being reused by most of the transport network towards the core. A related NGMN study [17] suggests the most common sites will be outdoor locations 3–6 m above street level. Other sites up to rooftop height will also be desirable.

Some wireless solutions for small cell backhaul have been proposed. These solutions can be categorized with similar characteristics, which are, to some extent, dictated by several different key design choices. These include:

- Carrier frequency from ~600 MHz to 80 GHz
- Line-of-sight and non-line-of-sight propagation
- Spectrum licensing arrangement (link licensed, area licensed, light licensed, or license exempt) and dynamic allocation
- Connectivity and topology (point-to-point, point-to-multipoint, forming tree, ring, or mesh)

To categorize wireless solutions for backhauling, the frequency bands can be used where each poses multiple advantages and some disadvantages; these include millimeter wave 60, 70–80 GHz, microwave 6–60 GHz, sub-6 GHz licensed bands, sub-6 GHz unlicensed bands, TV white space (TVWS), and satellite [18].

Wired Backhaul Solutions
For many years, wired backhaul solutions have played a very important role for backhauling traffic in mobile base stations due to their ability to support high data rates, reliability, and availability. New site types, in particular street furniture such as streetlamps and wall mounts for small cell-wired solutions, have all opted for this type of solution. They should also meet economic targets, much lower than those associated with the macrocell backhaul. Satisfying all of these criteria is challenging. However, wired solutions can still play a useful role in the small cell backhaul, and of course the overall backhaul solution can be a hybrid of wired and wireless, with the wired solution delivered to wireless hub sites. Several approaches are possible; one of them is the direct fiber-type connection [19].

Direct Fiber
Point-to-point high-speed data connections can be achieved through fiber with low latency, where virtually any backhaul capacity can be achieved. The solution is usually terminated with network terminating equipment (NTE). In general, direct fiber would acquire a higher connection and rental charge; thus, it would be typically

used as a hybrid solution. A wireless solution would backhaul any wireless hub sites [19].

At this stage, a few points on optical fibers are worth mentioning. Optical fibers include a transparent core and a cladding. Rays of light are kept in the core by total internal refraction. This approach also allows the cable to be twisted in delivery. These fibers that can support multiple propagation paths are known as multi-mode fibers (MMF). These are, in general, wider and used for shorter lengths. Fibers that have a single path are known as single-mode fibers (SMF).

Digital subscriber line (xDSL), asymmetric digital subscriber line (ADSL), and fiber to the x (FTTx) are some other alternative solutions for wired backhauling [20].

9.3.4 Network Performance Evaluation

This section focuses on the assessment of the KPIs using the IxChariot application over a hybrid network with two handsets (UE) connected to the same hybrid network wirelessly. The objective is to test the maximum performance of the eNB using different Layer 4 protocols, i.e., TCP and UDP measurements performed with IxChariot based on real network performance and using the OAI platform.

How the KPI measurement was performed:

The IxChariot platform instantly assesses network performance, including wireless performance by using a simple server-client topology. Usually, at any client (UE in our case) is installed an endpoint executable file or, in the case of Android/iOS, the appropriate application. This endpoint continuously runs as a daemon process and synchronizes with the server side to provide real-time results. Figure 9.14 describes the testbed setup that includes the OAI module.

We have considered to measure the data rate KPI, which is also considered by the 5GPPP Organization (Table 9.2).

The end-to-end (E2E) network value that we have measured by considering a professional measurement tool, the IxChariot, is as follows:

The tabulated results from Table 9.3 can be observed graphically by Fig. 9.15.

The similar test is performed with the OAI network using IxChariot tool. The tabulated results from Table 9.4 can be observed graphically by Fig. 9.16.

9.3.5 Conclusion

It is foreseen that DL/UL traffic volume in cellular networks will increase by 100 times that will constitute services like live streaming, high-definition videos, concerts, etc. The available resources cannot adequately serve the clients' needs. To cater for this demand, emerging 5G technology is considering an integrated optical wireless access network that should be carefully designed in order to provide an

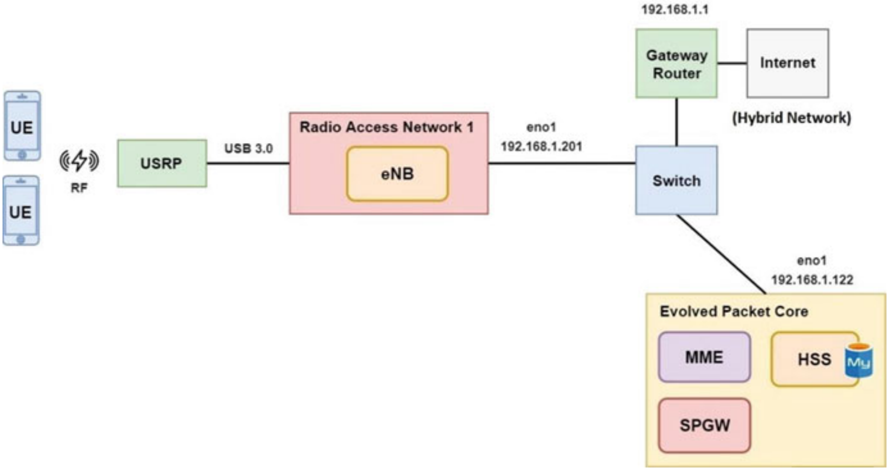


Fig. 9.14 OpenAirInterface network diagram [20]

Table 9.2 KPI values

Title	KPI-user experienced data rate
Description	Data rate as perceived at the application layer. It corresponds to the amount of application data (bits) correctly received within a certain time window
Where to measure	UEs and application server; in communication endpoints, a distinction shall be made between uplink (UL) and downlink (DL) KPI measurements
How to measure	Measurements will follow a passive approach targeting traffic generated by the applications at hand. As such, measurements will focus on sampling the user experienced data rate over a long observation interval (e.g., lasting in about 3 min). The use of synthetic data will also be considered so as to stress the network to its maximum capacity limits

Table 9.3 Data rate KPI measurements

Value no.	Data rate average (Mbps)	Data rate minimum (Mbps)	Data rate maximum (Mbps)
Value 1	22.011	3.361	77.728
Value 2	22.054	3.980	66.667
Value 3	18.636	3.175	72.728
Value 4	18.634	3.704	66.667

optimized infrastructure for both backhauling and fronthauling, in terms of QoS, performance, and energy- and cost-efficiency. The design requirements that should be considered when designing these networks have been addressed here.

Moreover, many open-source platforms have been designed and deployed, which are introducing significant saving in costs and time in the innovation process of cellular communication. OAI, the open-source small cell, which is one of the modest platforms for LTE and 5G, focuses on the spectral, algorithmic, and protocol

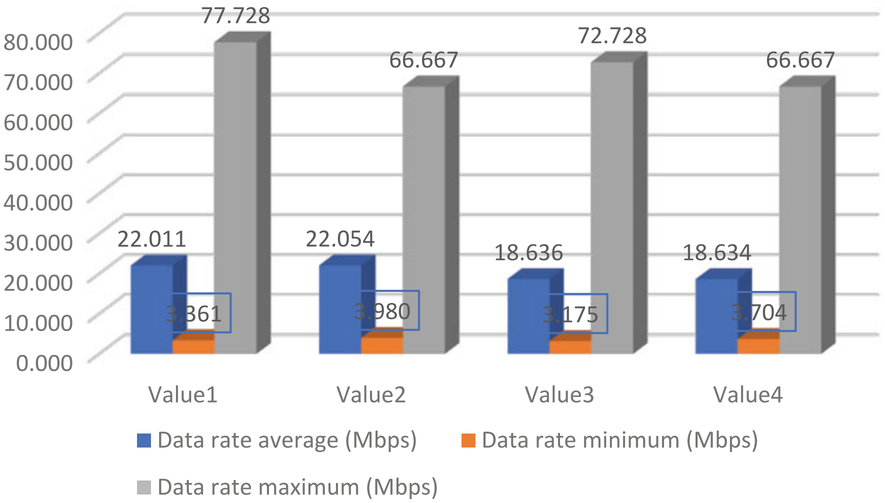


Fig. 9.15 Data rate KPI measurements

Table 9.4 Data rate KPI measurements with the OAI

Value no.	Data rate average (Mbps)	Data rate minimum (Mbps)	Data rate maximum (Mbps)
Value	16.827	13.107	28.311

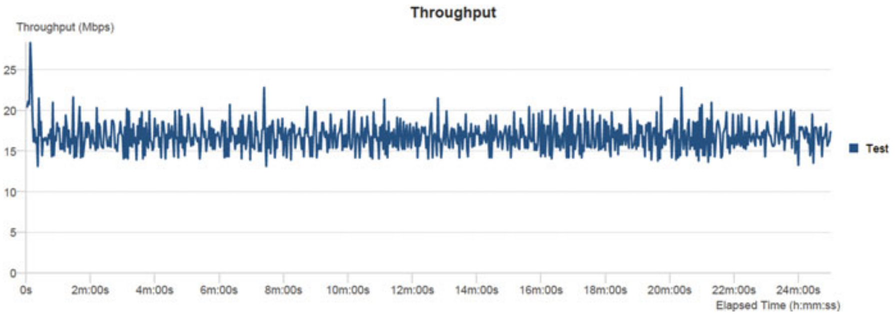


Fig. 9.16 Evolution of "throughput" aggregated from all the simulated users (OAI platform)

efficiency research to enable demonstrators which are not only based on high-performance embedded architectures but also employed towards validating and analyzing wireless systems; OAI does not focus on developing solutions that are deployment-ready.

The network performance evaluations have been conducted in this part with the IxChariot tool, a ready-made network performance evaluation tool. The objective is to test the maximum performance of the eNB using different Layer 4 protocols, i.e., TCP and UDP measurements performed with IxChariot based on real network performance and using the OAI platform. It was concluded that the OAI tool can provide comparable performance with vendor-specific small cell equipment, making it a good candidate for in-house prototype testing.

9.4 Big Data and Cloud Computing for 5G Networks: Stochastic Approach

9.4.1 Introduction

Fifth-generation cellular networks were introduced in 2008 by a partnership of NASA and M2Mi Corp as a coalition to specifically develop 5G. The main advantages compared to its predecessor 4G are the significant increase in network speed by orders of magnitude and the minimized latency, which will render it comparable with broadband networks. These attributes of the newly introduced network generation will make the real-scale application of IoT plausible, while on the other hand will introduce an excessive amount of data that will have to be processed which is one of the major challenges that this technology will have to address. Big data techniques will be exploited in order to gain insight of the network operation as well as perform a key role in machine learning-based network management.

A set of data can be characterized as big data when they can be described by the following characteristics [21, 22]. The first characteristic is “volume” [23] that represents the total data storage requirements. “Variety” is the second characteristic [24] of such data, which implies that they are generated from a multitude of sources and a range of data types, while “velocity” [25] represents the rate with which data are generated and accumulated. The former are considered the basics features of big data, while up to nine different features can be encountered in literature, such as value [25], veracity [26, 27], volatility [28], validity, variability [27, 29], and complexity [25, 30].

As evident, the basic principles of big data can be applied in case of mobile networks and especially 5G technology, as the data generated and exchanged between the network and users can be of enormous size, e.g., video streaming, while at the same time countless users generate and share content with each other which accounts for the velocity property attributable to big data. Finally, variety characteristic exists as well, as the data transferred via a cellular network can be generated by a multitude of sources, such as mobile users, IoT devices, etc. According to [31], these data generated in case of cellular networks can provide the infrastructure provider (InP) with insight on their operation and maintenance needs. Specifically, the network-accumulated data can contribute towards achieving a flexible network and functionality deployment, traffic awareness, and efficient network operation with minimized energy consumption. Having extracted all underlying data of a network, a machine learning agent is then utilized for the automatic management of a cellular network.

Machine learning [32] is a combination of algorithms and computational statistics that enable a computer to perform actions based on sample data. It is split into three major branches, namely, supervised learning [33] which maps inputs to outputs utilizing existing input-output datasets, unsupervised learning [34] which learns by identifying common ground of unclassified existing data, and finally reinforcement learning [35], which advances by taking suitable actions to maximize a reward. The first applications of machine learning for networks were limited to intrusion detection and performance prediction [36], network scheduling [37], and parameter adaptation [38, 39], while deep learning technologies were investigated in [40]. Finally, [41] investigates the ways that machine learning can assist in network design and optimization while at the same time summarizing typical machine learning workflows for application in networks.

In this work, a set of 5G network data concerning network states, traffic patterns, and QoS-related data (e.g., bandwidth) will be considered along with QoE attained by users. A mapping will be generated by utilizing a machine learning procedure, in order to generate a connection between the input and output data and the equivalent network optimization parameters in each case. In the next step, a stochastic process will generate random input data and by utilizing the former trained ML process will provide estimate network parameters in corner cases, thus providing possible network response actions in worst-case scenarios.

This chapter section is organized as follows. Section 9.4.2 will provide the big data processing framework that will be used in order to process the data at hand. After the data analysis step performed, Sect. 9.4.3 will provide the machine learning framework that will link input and output data. Section 9.4.4 introduces the basic stochastic processes that will generate randomized input data. Finally, Sect. 9.4.5 provides results of a case study with data related to a specific network cell, followed by the conclusions in Sect. 9.4.6.

9.4.2 Big Data Network Management

The constantly increasing capability of networks to provide higher bandwidth and increased area coverage, along with the production of mobile applications to cater every possible need, generate enormous amounts of data [42]. These data vary from application activities and user geolocation to operational data concerning slice admission and network optimization parameters. Infrastructure providers leverage the accumulated information by leveraging big data and thus can provide an improved performance of the cellular network while simultaneously maximizing their revenue [43]. As [44] states, the analysis of big data can extract information concerning the user's behavior, network traffic across different periods of the day, as well as their geographic user dispersion and mobility patterns. The basics of big data analytics as introduced in [44, 45] can be represented by a great number N that denote the measurements of n quantities. By grouping the quantities of each measurement into a vector x, all the data can be grouped into a matrix form as follows:

$$X = \left[\underset{(nx1)}{x_1} \underset{(nx1)}{x_2} \; \ldots \; \underset{(nx1)}{x_N} \right] \tag{9.1}$$

The former matrix representation of big data analytics applies only in case data are structured to allow this representation. Unfortunately, this is not always the case, as the data collected are semi-structured or unstructured making this organization impossible. To treat these data stream, NoSQL databases are used, while a pre-processing phase purifies data of unwanted or redundant information and ensures unified consistency [46]. The common regime for processing the derived dataset is by utilizing cloud computing-enabled big data platforms, as the cost for purchasing and maintaining a cluster is practically eradicated. One of the most common platforms for big data analysis is Hadoop [47]. In the case of cellular network design [22], big data analysis application focus on counter-failure issues, network state monitoring, and cache and network optimization.

The final issue of network optimization will be examined in this work. [48] proposed a mobile network optimization framework based on big data which will be adopted using MBrace [49]. For the sake of completeness, a summary of this framework will be provided here. To begin with, the first step of the framework is data accumulation. Data collected for the analysis of a mobile network stem from two different sources. The first source of data is derived from user equipment (UE). This data category provides an in-depth knowledge of users, as they include information such as location, mobility, and network usage patterns. The second data source comes from mobile network operators, which is mostly sourced by the radio access network or the core network. The latter provides data such as network performance, while the former provides cell information pertaining to configuration information, mobility, and network failing incidents. The second step of the process involves purifying and filtering data to remove incomplete or useless data that could provide an error source in the big data regime. Finally, a machine learning processing of the derived data could enable real-time classification and network adjustment. [50] presents several different perspectives of big data analysis. The first perspective presented utilizes divide and conquer strategies as a computing paradigm to address big data problems in the framework of distributed and parallel computing. Feature selection is a second perspective of big data as they play an important role in reducing the complexity and high dimensionality of the uncertain big data characteristics. Classification of big data is another major issue as traditional algorithms do not address big data properly. Traditional classification methods with machine learning are incompatible with big data as directly, and new parallel strategies or classification algorithms must be introduced.

9.4.3 Machine Learning Techniques

According to [50, 51], the increasing complexity of cellular networks and the enormous data stream constantly created render their efficient management a demanding task. The application of machine learning procedures to these data can detect patterns that would be hard to extract from otherwise. Deep learning has gained a wide acceptance among network researchers who are using it to tackle network-related problems [41, 52]. The reason that renders deep learning the ideal tool is the extreme heterogeneity of data. Traditional tools such as shallow neural networks cannot handle a large number of measurements and measured quantities, while excessive data do not boost their performance, which contradicts to the response of deep neural networks. [53, 54] showcase state-of-the-art deep learning techniques applied to mobile networks. Namely, [53] combines deep neural networks with reinforcement learning by asynchronously multiple agents on multiple instances of the environment. This strategy allows a multitude of reinforcement learning (RL) algorithms to be applied to deep neural networks robustly. [35, 55] provide a thorough analysis of such algorithms which will also be summarized here. Q-learning is a model-free reinforcement learning algorithm which can see also be perceived as a method of asynchronous dynamic programming. Agents act optimally in domains without needing to map them and always converge to optimum action values. SARSA was based on TD-learning (compared to its predecessor Q-learning which is based on model-free reinforcement learning) and converges much faster. Q(λ) algorithm is the former Q-learning supported by eligibility traces. They keep records of recently visited state-action pairs and the degree for which each state-action pair is eligible for learning changes. SARSA(λ) applies eligibility traces to the initial SARSA algorithm. In addition, backups are carried out over n-steps. Deep Q networks can combine reinforcement learning with deep neural networks, while deep deterministic policy gradient algorithms are used in case of continuous action spaces. Asynchronous actor-critic algorithm can execute in parallel which enables the algorithm to act on diversified data, and an actor-critic method is used for updates with the aid of learned state-value functions. Normalized advantage function (NAF) is the generalization of Q-learning for continuous domains. The NAF algorithm is simplified as a second actor, meaning that the policy function is avoided. Trust region policy optimization provides the ability to optimize non-linear policies with thousands of parameters, while proximal policy optimization outperforms other policy gradient methods while balancing sample complexity, simplicity, and ease of tuning.

In this work, general regression neural networks (GRNN) [56] and random forests [57] will be applied in order to both assess how the different data dimensions gathered from both UE and network cell data affect QoE and which of the dimensions affect QoE the most. These algorithms are programmed using MBrace in order to exploit the immense power of cloud resources.

9.4.4 Stochastic Processes

In this section, an introduction to stochastic processes will be given based on [58], and random measurement will be generated to examine the data generated by the machine learning process in extreme network functioning cases.

Each of the input measured quantities of the former sections will be handled now as random variables. A random variable is a function that provides a real number to a product of a probabilistic phenomenon. Each random variable is described by two functions: the cumulative distribution function (cdf) that expresses the probability of a value being lower than a number and the probability density function (pdf) which is the derivative of the latter. The cdf takes values ranging from 0 to 1, while pdf can have greater maximum values under the limitation that its integral for the whole real number domain equals to unity. Some of the most important values that describe a distribution functions are its moments. These values are used to qualitatively describe the shape of a function. The nth moment of a function f(x) is given by:

$$\mu_n = \int_{-\infty}^{+\infty} (x - n)^n f(x) dx \tag{9.2}$$

The first moment of a probability distribution function provides its mean value (see Fig. 9.17) and is usually denoted by $\mu \equiv E[x]$ and is usually seen as the average of the value of the distribution in case of a limited set of values.

The second moment of the distribution produces its variance. In probability theory, variance is considered the expectation of the squared deviation of a random variable from its mean. Figure 9.18 illustrates the behavior of the normal distribution function with varying variance. As depicted, the higher the variance becomes, the distribution widens which means that the probability of attaining values further from

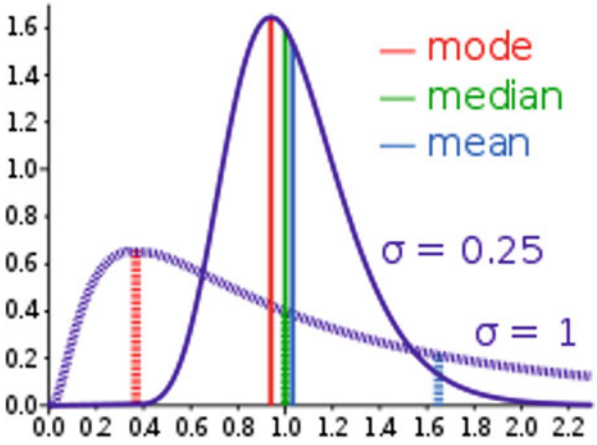


Fig. 9.17 Mean and median values of probability density function

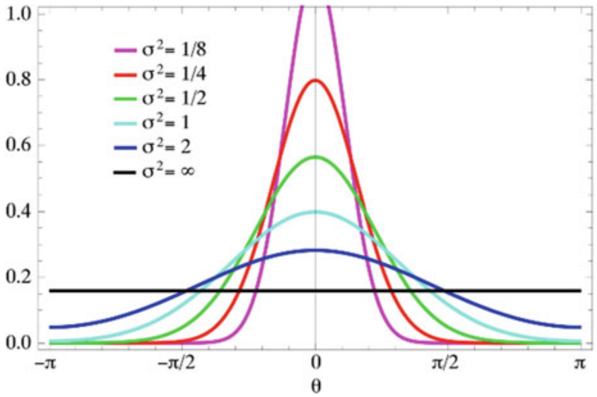


Fig. 9.18 Mean and median values of probability density function

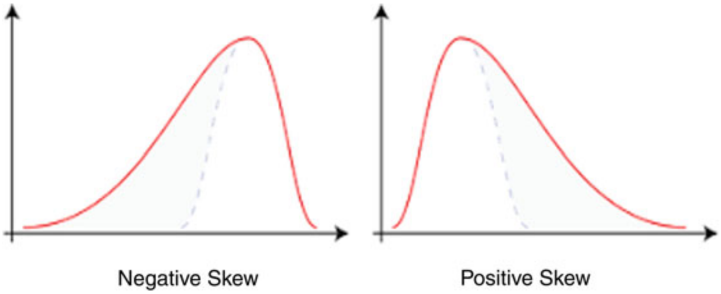


Fig. 9.19 Skewness describing the lopsidedness of the distribution

the mean value increases and in case of an infinite variance, the distribution becomes equivalent to white noise in which case there is an equal probability of attaining any value of the real number line. Finally, two more moments are of great interest in stochasticity, namely, third and fourth moments, which provide the skewness and the kurtosis of the distribution. Skewness describes the lopsidedness of the distribution, while kurtosis is the measure of tailedness of a probability distribution. Figure 9.19 shows a negative skewness gives a right-leaning and a positive skewness a left-leaning curve, equivalently. Finally, the lesser the kurtosis, the more pronounced the tails of the distribution become and vice versa.

Since the pdf of each of the variables concerning the network data have already been derived, a sampling method must be selected. Some of the most used sampling methods are Monte Carlo sampling, Latin hypercube sampling, and stratified sampling. The simplest among them is the Monte Carlo method, where a domain of possible input variables is defined. For each input variable, random values are generated given a probability distribution function, and then a deterministic analysis of the model using the generated inputs is performed. Latin hypercube sampling is a method for generating random values from a multidimensional distribution. Finally, stratified sampling is a method of sampling from a population that can be partitioned into independent subpopulations. All the above sampling methods generate random input data combined with a simulation method such as subset simulation to compute rare events such as failures. In the case of networking data, the subset simulation method will be utilized to assess edge cases of the mobile network system. The basic idea behind subset simulation, as described in [59], is to express the failure probability as a product of larger conditional failure probabilities by introducing intermediate failure events.

9.4.5 *Case Study*

In this work, a metric of assessing QoE is the number of dropped calls. In an effort to perform ad hoc optimizations to resource scheduling of the network, both UE and cell data were collected for a specific network cell, aggregated per call for a period of 6 months, enriched with data collected from users regarding QoE. A summary of the data per category is summarized in Table 9.5.

The data were fed to a random forest algorithm in order to assess the importance of each category when calls are being dropped. The data gathered are depicted in Fig. 9.21.

It is obvious from Fig. 9.20 that for the network cell considered, all factors pertaining to congestion (usage metrics) and bandwidth are well covered with radio performance being the dominant factor for dropped calls. This is confirmed by both UE data showing erratic communication to the specific cell and handover attempts. It is important to state here that handover failures are also a significant factor, contributing to dropped calls.

Table 9.5 Data collected per category

Category	Handover	RRC/MME	Radio	Usage	Bandwidth	Other
No of variables	12	10	8	3	3	7

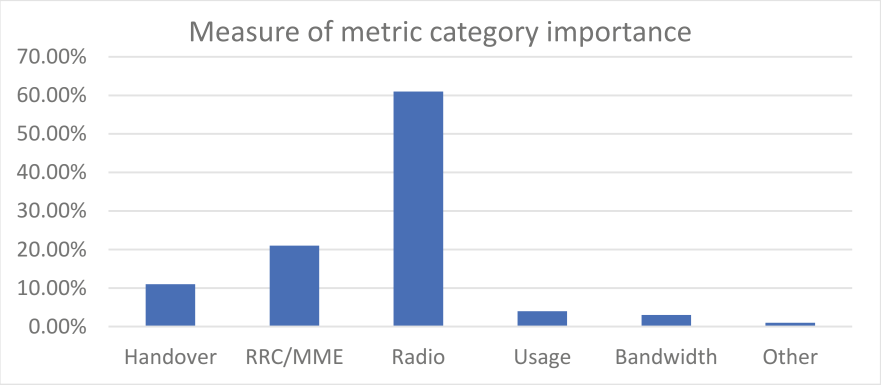


Fig. 9.20 Category importance for dropped calls

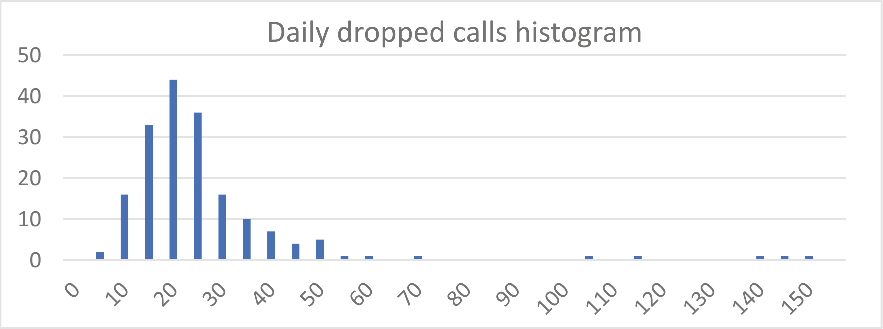


Fig. 9.21 Dropped calls per day frequency

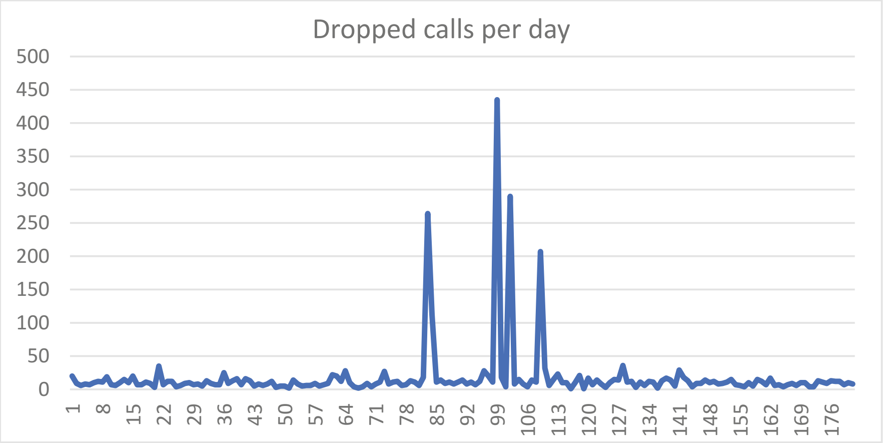


Fig. 9.22 Dropped calls per day

In Fig. 9.21, a daily dropped call histogram is depicted. Considering the shape of this histogram, a log-normal distribution is proposed for describing its stochastic properties. The trend of dropped calls is shown in Fig. 9.22, while actual user complaints pertaining to QoE are shown in Fig. 9.23.

In order to assess the importance of each metric category to QoE, the combination of UE and network cell data was fed to a random forest algorithm along with the QoE data. The results are depicted in Fig. 9.24. There is an obvious shift of importance when QoE is taken under consideration. Now bandwidth issues seem to play an important role, while the most prominent reason for inferior QoE seems to be RRC and MME issues. Virtual resource balancing can come into play here, providing more resources to MME-related tasks in order to limit the time needed for UE session initiation.

Having a clearer picture of the factors contributing to a perceived decrease of QoE, a GRNN was trained with a subset of the data used for the random forest

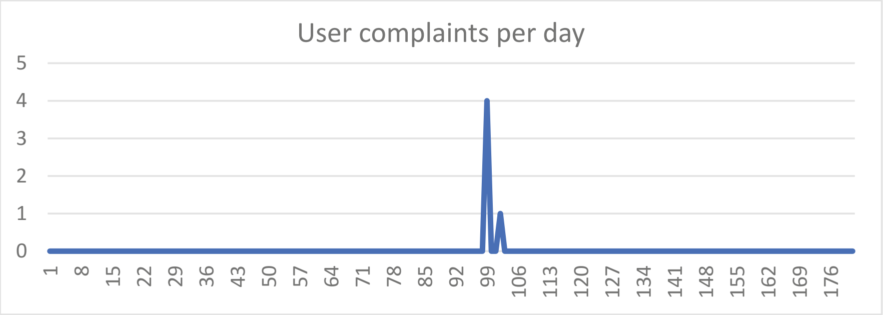


Fig. 9.23 User complaints logged per day

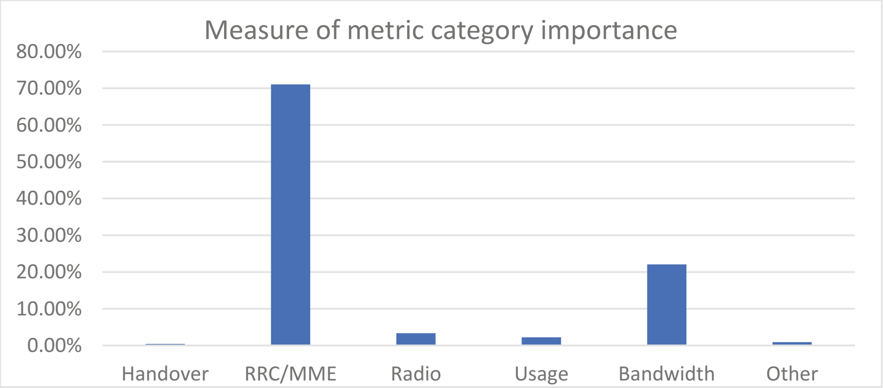


Fig. 9.24 Category importance for QoE

algorithm. For this data, the GRNN exhibited a 14% error, rendering it acceptable for predicting the perceived QoE when fed with actual UE and network cell data.

The trained GRNN was used in order to assess perceived QoE when conditions were 30% better (optimistic scenario) and 30% worse (pessimistic scenario). This was accomplished with the aid of MSolve. Stochastic [60] by generating two sets of corresponding artificial datasets for a duration of 2 years, adhering to log-normal distributions with mean values and standard deviations being decreased and increased by 30%, respectively. Dropped calls are depicted in Figs. 9.25 and 9.26, while user complaints are depicted in Figs. 9.27 and 9.28.

From this data, it is clear that perceived QoE and actual dropped calls do not exhibit a rational correlation (i.e., the more the dropped calls, the less the perceived QoE). According to the opinion of the authors, this stems from the fact that dropped calls due to poor signal offer a “rational” explanation for the users as to why the calls were dropped, while dropped calls with no radio issues (e.g., poor signal strength) are perceived as more disturbing.

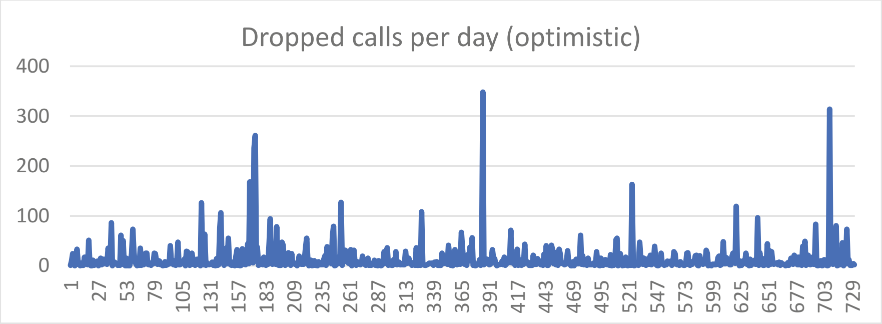


Fig. 9.25 Dropped calls per day for optimistic scenario

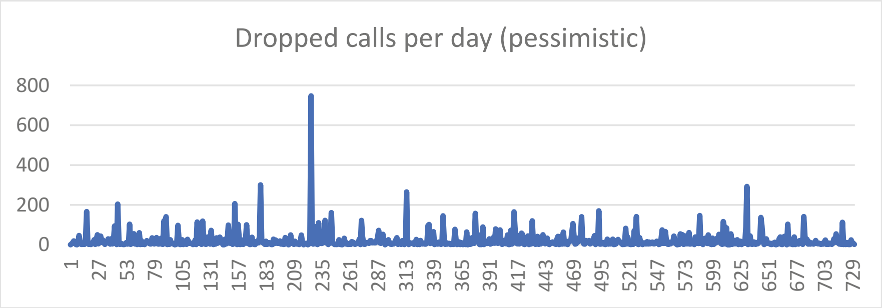


Fig. 9.26 Dropped calls per day for pessimistic scenario

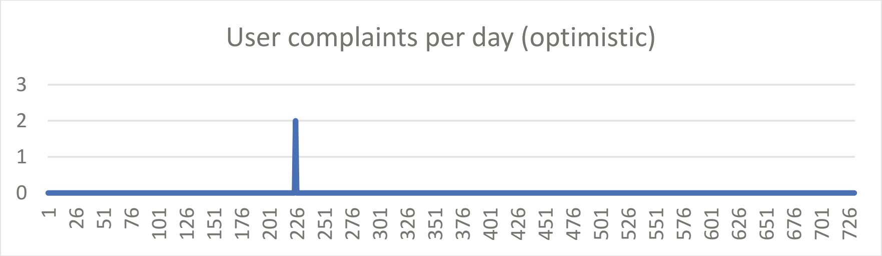


Fig. 9.27 User complaints predicted per day for optimistic scenario

9.4.6 Conclusions

Virtual resource management offers a vast array of possibilities with respect to near real-time optimization of cellular networks. In this context, a rationalization of multidimensional UE and network cell data was attempted in order to establish a cause-and-effect relationship of gathered data and perceived QoE. Due to the vast volume, heterogeneity, and multidimensionality of the data considered, the

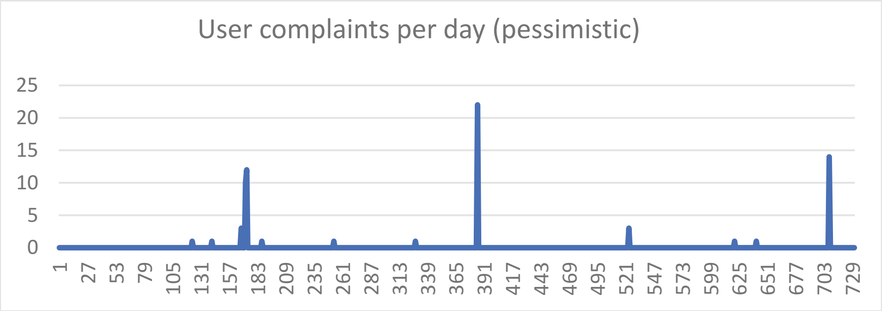


Fig. 9.28 User complaints predicted per day for pessimistic scenario

MBrace big data framework was utilized for programming and executing various AI algorithms. In order to assess the sensitivity of the perceived QoE to various network parameters that resulted in dropped calls, a stochastic analysis was performed by exploiting the MSolve. The *stochastic* tool was able to predict the perceived QoE using the GANN ML approach. The results provided a valuable insight as to how various factors contributing to dropped calls affected the perceived QoE differently and could provide valuable guideline as to how ad hoc network optimizations could be performed.

9.5 Conclusion

In this chapter, V-band Fiber-Wireless physical layer experimental performance assessments are initially presented. The integrated communication platform was able to demonstrate more than 1 GHz of available spectrum bandwidth with negligible degradation while beamsteered across a 90° sector. Then, following linearity measurements on the channel and 1 Gbps 16-QAM transmission, two multi-user cases are presented exploiting either frequency- or spatial-division multiplexing, respectively, demonstrating frequency aggregation and frequency reuse over the same network infrastructure, with performance below the 3GPP-defined threshold of 17.5% EVM figure [12], in both scenarios.

Following the discussion on physical layer performance of the integrated optical wireless link, the authors also address the small cell deployment and experimentation perspective. In the first instance, a case is made for open-source platforms towards promoting enhanced savings in terms of research effort and cost. The so-called OAI was introduced as a typical open-source platform that is widely used by the research community for fast and low-cost prototyping. Moreover, this chapter considered the deployment strategy for integrated fronthaul-backhaul network targeting highly dense hotspot coverage areas. A case study was given that showed the intricacies of designing and deploying ARoF-based architectures for live football event that led to a general set of design guidelines for more generic scenarios. Migrating from the integrated fronthaul-backhaul technologies towards a more global perspective of network performance, the authors also provide a holistic view on network performance using the IxChariot application. The objective is not only to assess the viability of IxChariot as an experimental tool but to compare the OAI with vendor-specific equipment.

Not only we investigate network architectures and practical deployment approaches, but we also address their optimization based on ML approaches. Virtual resource management and big data represent prominent 5G paradigms that offer new opportunities in terms of network management and optimization. In this chapter, we consider how ML can play a major role in network optimization by bringing intelligence to the limelight and by investigating how learned patterns or relationships between network states and QoE can be used towards intelligent prediction and optimization. As such, a stochastic analysis is presented towards predicting the perceived QoE in mobile networks, resulting in valuable insights on how various factors contributing to dropped calls affected the perceived QoE differently and as to how ad hoc network optimizations could be performed.

Acknowledgments The work is supported by H2020 5GPPP Phase II project 5G-PHOS (Contract No. 761989) and 5G STEP-FWD Marie Curie EU project (Contract No. 722429). The authors would like to acknowledge Keysight for supporting the experiments with measurement equipment. The authors would like to acknowledge the contribution of MGroup (http://mgroup.ntua.gr) to the efforts of this work along with the valuable insight and help of its member Dimitris Tsapetis. This work was funded by the Spotlight Marie Curie EU program.

References

1. Next Generation Mobile Networks Alliance, "5G White Paper," 2015. [Online] Available: https://www.ngmn.org/wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf
2. ETSI White Paper No. 9, "E-Band and V-Band - Survey on status of worldwide regulation", first edition – June 2015, ISBN No. 979-10- 92620-06-1. [Online] Available: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp9_e_band_and_v_band_survey_20150629.pdf
3. https://www.businesswire.com/news/home/20190621005318/en/Worldwide-52-Bn-Small-Cells-Market%2D%2D
4. https://myriadrf.org/news/open-source-lte/5/6
5. ITU-T Series G/Supplement 66 '5G wireless fronthaul requirements in a passive optical network context' 07/2019.
6. Brown, G. *Exploring 5G new radio: Use cases, capabilities & timeline*, Qualcomm White Paper, Sept. 2016.
7. eCPRI specifications V2.0 (2019-05-10) [Online] Available: https://www.gigalight.com/downloads/standards/ecpri-specification.pdf
8. Lim, C., et al. (2019). Evolution of radio-over- Fiber technology. *Journal of Lightwave Technology, 37*(6), 1647–1656.
9. Chih-Lin, Y., Liu, S., Han, S. W., & Liu, G. (2015). *On big data analytics for greener and softer RAN*. IEEE Access. https://doi.org/10.1109/ACCESS.2015.2469737

10. Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 44*, 206–226.
11. Ackerman, E. I., & Cox, C. H. (2001). *IEEE Microwave Magazine, 2*(4), 50–58.
12. 3GPP TS 38.104, "5G; NR; Base Station (BS) radio transmission and reception", v. 15.2.0, 2018-7
13. https://www.openairinterface.org/?page_id=72
14. www.thinksmallcell.com/opensource
15. ITU-T Series G/Supplement 66 '5G wireless fronthaul requirements in a passive optical network context' 07/2019.
16. blueSPACE - H2020-ICT-2016 context and open source documents.
17. "Small cell backhaul requirements", NGMN Alliance, June 2012, http://goo.gl/eHHtx
18. "60 GHz Technology for Gbps WLAN and WPAN: From Theory to Practice" Su-Khiong (SK) Yong et al, Wiley 2010, http://goo.gl/aqkPI
19. Report title: Backhaul technologies for small cells, 14 Feb 2013 Version: 049.07.02, P-61.
20. Bonding and vectoring rates: http://www.alcatel -lucent.com/wps/PA_1_A_9C1/DocumentDownloadFormServlet?LMSG_CABINET=Docs_and_Resource_Ctr&LMSG_CONTENT_FILE=White_Papers/Leveraging_VDSL2_for_Mobile_Backhaul_SWP.pdf&lu_lang_code=en_WW
21. Kaur, N., & Sood, S. K. (2017). *Dynamic resource allocation for big data streams based on data characteristics (5Vs)*. https://doi.org/10.1002/nem.1978
22. Hadi, M. S., Lawey, A. Q., El-Gorashi, T. E. H., & Elmirghani, J. M. H. (2018). *Big data analytics for wireless and wired network design: A survey*. Computer Networks. https://doi.org/10.1016/j.comnet.2018.01.016
23. Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data, 2*, 24. https://doi.org/10.1186/s40537-015-0032-1
24. Baek, H., & Park, S. K. (2015). Sustainable development plan for Korea through expansion of green IT: Policy issues for the effective utilization of big data. *Sustainability, 7*, 1308–1328. https://doi.org/10.3390/su7021308
25. Kaisler, S., Armour, F., Espinosa, J. A., & Money, W.. (2013). *Big data: Issues and challenges moving forward.*https://doi.org/10.1109/HICSS.2013.645
26. Demchenko, Y., Grosso, P., Laat, C. De, & Membrey, P.. (2013). *Addressing big data issues in scientific data infrastructure.*https://doi.org/10.1109/CTS.2013.6567203
27. Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C., & Yang, G. Z. (2015). Big data for health. *IEEE Journal of Biomedical and Health Informatics, 19*, 1193–1208. https://doi.org/10.1109/JBHI.2015.2450362
28. Zhang, L.. (2014). *A framework to specify big data driven complex cyber physical control systems.*https://doi.org/10.1109/ICInfA.2014.6932715
29. Almeida, P. D. C. De, & Bernardino, J. (2015). *Big data open source platforms*. https://doi.org/10.1109/BigDataCongress.2015.45
30. Gani, A., Siddiqa, A., Shamshirband, S., & Hanum, F. (2016). A survey on indexing techniques for big data: Taxonomy and performance evaluation. *Knowledge and Information Systems, 46*, 241–284. https://doi.org/10.1007/s10115-015-0830-y
31. Chih-Lin, I., Liu, Y., Han, S., Wang, S., & Liu, G. (2015). On big data analytics for greener and softer RAN. *IEEE Access, 3*, 3068–3075. https://doi.org/10.1109/ACCESS.2015.2469737
32. Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 44*, 206–226.
33. S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," Informatica (Ljubljana). 2007.
34. Francis, L. (2014). Unsupervised learning. In *Predictive modeling applications in actuarial science: Volume I: Predictive modeling techniques*. Cambridge University Press.
35. Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research, 4*, 237–285.

36. Sun, Y., et al. (2016). *CS2P: Improving video bitrate selection and adaptation with data-driven throughput prediction*. https://doi.org/10.1145/2934872.2934898
37. Mao, B., et al. (2017). Routing or computing? The paradigm shift towards intelligent computer network packet transmission based on deep learning. *IEEE Transactions on Computers, 66*, 1946–1960. https://doi.org/10.1109/TC.2017.2709742
38. Winstein, K., & Balakrishnan, H. (2013). *TCP ex machina: Computer-generated congestion control*. https://doi.org/10.1145/2534169.2486020
39. Dong, M., Li, Q., Zarchy, D., Godfrey, P. B., & Schapira, M. (2015). *PCC: Re-architecting congestion control for consistent high performance*.
40. Fadlullah, Z. M., et al. (2017). State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems. *IEEE Communications Surveys & Tutorials, 19*, 2432–2455. https://doi.org/10.1109/COMST.2017.2707140
41. Wang, M., Cui, Y., Wang, X., Xiao, S., & Jiang, J. (2018). Machine learning for networking: Workflow, advances and opportunities. *IEEE Network, 32*, 92–99. https://doi.org/10.1109/MNET.2017.1700200
42. Liu, J., Liu, F., & Ansari, N. (2014). Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop. *IEEE Network, 28*, 32–39. https://doi.org/10.1109/MNET.2014.6863129
43. Bi, S., Zhang, R., Ding, Z., & Cui, S. (2015). Wireless communications in the era of big data. *IEEE Communications Magazine, 53*, 190–199. https://doi.org/10.1109/MCOM.2015.7295483
44. He, Y., Yu, F. R., Zhao, N., Yin, H., Yao, H., & Qiu, R. C. (2016). Big data analytics in mobile cellular networks. *IEEE Access, 4*, 1985–1996. https://doi.org/10.1109/ACCESS.2016.2540520
45. Qiu, R. C., Hu, Z., Li, H., & Wicks, M. C. (2012). *Cognitive radio communication and networking: Principles and practice*.
46. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access, 2*, 652–687. https://doi.org/10.1109/ACCESS.2014.2332453
47. Meng, X., et al. (2016). MLlib: Machine learning in apache spark. *Journal of Machine Learning Research, 17*, 1235–1241.
48. Zheng, K., Yang, Z., Zhang, K., Chatzimisios, P., Yang, K., & Xiang, W. (2016). Big data-driven optimization for mobile networks toward 5G. *IEEE Network, 30*, 44–51. https://doi.org/10.1109/MNET.2016.7389830
49. Dzik, J., Palladinos, N., Rontogiannis, K., Tsarpalis, E., & Vathis, N. (2013). *MBrace: Cloud computing with monads*. https://doi.org/10.1145/2525528.2525531.
50. Xie, J., et al. (2018). A survey on machine learning-based mobile big data analysis: Challenges and applications. *Wireless Communications and Mobile Computing, 2018*. https://doi.org/10.1155/2018/8738613
51. Zhang, C., Patras, P., & Haddadi, H. Deep learning in mobile and wireless networking: A survey. *IEEE Communications Surveys & Tutorials, 21*, 2224–2287. https://doi.org/10.1109/comst.2019.2904897
52. Zhu, H., Zhang, Y., Li, M., Ashok, A., & Ota, K. (2018). Exploring deep learning for efficient and reliable mobile sensing. *IEEE Network, 32*, 6–7. https://doi.org/10.1109/MNET.2018.8425293
53. Mnih, V. et al. (2016). *Asynchronous methods for deep reinforcement learning*.
54. Arjovsky, M., Chintala, S., & Bottou, L.. (2017) *Wasserstein generative adversarial networks*.
55. Andrew, A. M. (1998). Reinforcement learning: An introduction. *Kybernetes, 27*, 1093–1096. https://doi.org/10.1108/k.1998.27.9.1093.3
56. Isabona, J., & Osaigbovo, A. I. (2019). Investigating predictive capabilities of RBFNN, MLPNN and GRNN models for LTE cellular network radio signal power datasets. *FUOYE Journal of Engineering and Technology, 4*. https://doi.org/10.46792/fuoyejet.v4i1.339
57. Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

58. Papadopoulos, V., & Giovanis, D. G.. (2018). Stochastic finite element method. In *Mathematical engineering*.
59. Au, S. K., & Beck, J. L. (2001). Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics, 16*, 263–277. https://doi.org/10.1016/S0266-8920(01)00019-4
60. MGroup, *MSolve.Stochastic GitHub repo*. https://github.com/mgroupntua/MSolve.Stochastic

Part IV
Cloud Based UDNs for Beyond 5G

Chapter 10
Virtual Networking for Lowering Cost of Ownership

Fatma Marzouk, Maryam Lashgari, João Paulo Barraca, Ayman Radwan, Lena Wosinska, Paolo Monti, and Jonathan Rodriguez

Abstract 5G and beyond mobile networks hold the promise of supporting a vast emergence of new services and increased traffic growth. This represents a challenge for mobile networks operators, which are faced with the pressure of providing a variety of these services according to the stringent requirements of future mobile network generations, while still being able to (i) preserve service resilience, (ii) sustain profitability by reducing costs, and (iii) ensuring minimal energy consumption in the infrastructure. Fortunately, emerging 5G and beyond networks are expected to adopt increasingly prominent technological drivers that can tackle the above challenges by pushing the planning and management operations logic to the limit.

10.1 Introduction

It is agreed that 5G and beyond mobile networks generations would increasingly include key paradigms such as virtualization and autonomous management, which promise more flexibility and reactivity for mobile network operators (MNOs). Along with this technical and architectural turning point, governing business models have also evolved to reinvent roles and relationships between players, while opening new opportunities for innovation.

F. Marzouk (✉)
University of Aveiro, Aveiro, Portugal
e-mail: fatma.marzouk@ua.pt

M. Lashgari · L. Wosinska · P. Monti
Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden
e-mail: maryaml@chalmers.se

J. P. Barraca · A. Radwan
Instituto de Telecomunicações, Campus Universitário de Santiago, Aveiro, Portugal

J. Rodriguez
Instituto de Telecomunicações, Campus Universitário de Santiago, Aveiro, Portugal

Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, UK

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_10

These technological key paradigms have been either adopted or recognized to be relevant by the 5G 3GPP; however, approaches as to how to exploit their application for RAN technology in a bid to promote further efficiency for MNOs are still in their infancy.

To this extent, we devote this chapter to first highlight the benefits of virtualization and autonomous technologies, when applied to the RAN infrastructure. Particularly, we highlight how they have driven technical and business models for RAN sharing. In this context, we initially target meeting performance reliability in a cost-efficient way based on a proposed network-planning strategy and subsequent performance evaluation. Then, we extend this study to investigate the trade-off between the savings introduced from the centralization of service processing and the additional cost due to the addition of backup paths/resources. Thereafter, we focus on energy efficiency RA (resource allocation) design. A literature review is conducted on existing approaches for energy-efficient resource allocation, highlighting the existing gaps and open research challenges. Finally, we propose a design for hybrid resource allocation aiming at improving energy efficiency (EE) on the C-RAN, where we consider the optimal use of both computational and radio resources.

10.2 RAN Virtualization and Autonomous Management

The adoption of virtualization and autonomous management technologies represents one of the most promising approaches to handle the increased complexity in future RAN management operations.

These technologies are expected to coexist in synergy within the 5G landscape and beyond. In this section, we provide an overview of the main RAN virtualization and autonomous paradigms and highlight their impact on future mobile networking.

10.2.1 Enabling Paradigms

10.2.1.1 NFV

With virtualization techniques, networking functions are implemented in software that is able to run independently of the underlying hardware. As such, all devices can be virtualized by being abstracted to a virtualized functionality using network function virtualization, with the exception of the ones that handle the reception/transmission of wireless signal. The latter can be virtualized using software-defined networking concept. Although rising from the computing world, the concept of network function virtualization is today increasingly applied to wireless mobile networking, particularly toward radio access networks – baseband processing pooling is one form of enabling NFV, which is commonly referred to as Cloud RAN or virtual RANs. Compared to legacy RANs, the baseband processing and scheduling tasks in C-RANs are migrated to the cloud. By generalizing the application of NFV in future generation RANs, the capacity of every network function would be rendered available for expansion and reduction through an increase/decrease of virtual resources according to the network load conditions, which would enhance the elasticity of the network, improves resources utilization efficiency, and leads to CAPEX (capital expenditure)-OPEX (operational expenditure) savings. When virtualization techniques including NFV and SDR are applied beyond the RAN functionality coverage, to include either infrastructure, spectrum, or air interface virtualization, we refer to this as wireless network virtualization (WNV). WNV is the main enabler for future RAN sharing between multiple mobile virtual network operators (MVNOs), as it would allow them to share a common RAN on demand, potentially provided by a neutral operator.

10.2.1.2 SDN

To face the rapid evolution of networks, network configuration approaches had to evolve to include more innovation in the way complex networks are controlled. Since some years ago, research initiatives in that sense reached a theoretical concept answering this requirement. One fruitful outcome of these efforts is software-defined networking (SDN). The essence behind the SDN logic is decoupling the control plane of networking devices from the forwarding plane. The Open Networking Foundation presents the main body responsible for the standardization of SDN. The proposed SDN architecture by ONF is composed of three planes: data plane, control plane, and application plane. In the data plane, the SDN-enabled networking devices are abstracted to their simple data forwarding functions. The control plane contains the SDN controller that represents the control logic. The application plane hosts the various SDN-based business-specific applications. Thanks to the abstraction of the underlying devices, new applications can easily be deployed at this layer through simple programming. The interfaces between the three different planes in SDN are open. The southbound is responsible for communications with the data plane, while the northbound is responsible for the communications with the application plane.

The application of SDN on the RAN presents a key enabler for addressing futures RAN complexity in terms of management and control operations. Indeed, SDN can complement C-RAN architectures by offering a software-defined RAN (SD-RAN) control plane that is programmable and configurable via an innovative and advanced SD-RAN application plane. The implemented application can include a panoply of evolved decision maker modules for the management and optimization of the complex RAN resources/operations. The output decisions at this layer will be ultimately translated via the SD-RAN control plane to a set of configurations to be adjusted by the data plane via the SDN southbound interface.

10.2.1.3 SON and Learning Based Management

Self-organization (SO) is a concept that was first developed in chemistry and physics and then applied to biology, physics, social sciences, computer science, and finally mobile networking systems [1]. With respect to the first field of application, the SO concept is defined by the emergence of pattern and order in a system by internal processes, rather than external constraints or forces [2]. Following the same basic principles, the application of SO to the field of mobile networking systems leads to self-organizing networks (SONs), where traditional network management procedures are improved thanks to the automation capability, brought by the SO concepts and SON technology. SON-driven automation is enabled by adding more intelligence to the network following the concept of self-configuration, self-optimization, and self-healing. For MNO, SON technology leads to (i) simplification of operational and management tasks in HetNets, (ii) improvement of network performance, (iii) reduction of time to market of new services, and (iv) OPEX savings. Benefits include more than 50% reduction in dropped calls, OPEX savings of more than 30%, and an increase in service revenue by 5–10% [40].

Research in the area of SON, especially in the scope of HetNets, has attracted the attention of a large segment of the research community over the past decade. The contributions of the research efforts have spanned different challenges, including specifying SON challenges [3–5], proposing algorithms in support of a given self-function [6–9], and surveying the state of the art [5, 10, 11].

The first instance of SONs can be described as adaptive and autonomous systems, based on control loops and threshold comparison. In order to handle more complex scenarios, the current state of the art is investigating the application of advanced techniques, such as machine learning (ML) and data mining to SON [12–16]. ML algorithms can be categorized in multiple ways, with a stronger focus on supervised learning, unsupervised learning, and reinforcement learning. Several efforts of applying ML to SON are available in the literature to enhance mobility robustness optimization (MRO), mobility load balancing (MLB), capacity and coverage optimization (CCO), self-healing, resource allocation, and energy saving.

SON in future networks will be considered as an integral part of the RAN, rather than a complementary part. Indeed, the evolved 5G and beyond Cloud RAN will require not only faster operation of SON but also new, innovative, and proactive operations. Hence, the coverage map of SON algorithms within the ongoing set of 5G and beyond standards will be extended to cater for new use cases. These include self-protection to cater for automated security [17–20], SON for mmWave [21–23], SON for MIMO to enable adaptability to the different propagation characteristics of the mmWave links [24, 25], SON for NFV-based networking mainly to ensure optimal NFV placement and traffic steering problem [26, 27], and SON for multi-RAT optimization and spectrum sharing, toward overall improved networks performance [28]. Ultimately, SON for EE radio management is another use case of SON application for 5G networks that is currently attracting lots of interest.

10.2.2 Impacts on Future Mobile Networking

In line with growing costs and declining revenues for mobile operators, network sharing is emerging as a disruptive mechanism that can recover significant OPEX/CAPEX costs, by creating new sources of revenues and new cost reduction solutions. Indeed, network sharing would allow an operator that does not have the infrastructure nor spectrum resources, to dynamically share the physical networks operated with other mobile network operators and hence maximize resource utilization efficiency.

10.2.2.1 New Business Model

The concept of resource sharing has evolved over time and has recently experienced a major wave of revolution. Indeed, with the emergence of softwarization and autonomic management technologies, resource sharing concept has transitioned from only hardware-based resource sharing to overall softwarized mobile network-based resource sharing. The first form of hardware-based sharing has appeared with 3GPP Rel.99, allowing operators to share non-active assets of the RAN such as site locations or physical supporting infrastructure of radio equipment. Later, with the advent of 3GPP Rel-6 (UMTS) [29], Rel-8 (LTE) [30], and Rel-10 (LTE-A) [31], active sharing appeared, where operators share BS elements like the RF chains, antenna, or even Radio Network Controllers (RNC). LTE brought also a growing interest among operators to additionally enable spectrum-based sharing to maximize spectrum efficiency. LTE spectrum sharing technologies consider three spectrum segments including the TV white space channels, the frequently unused service-dedicated 3.5GHz, and the 5GHz unlicensed band.

The aforementioned hardware and spectrum-based sharing schemes are based on fixed contractual agreements/sharing framework over long time periods (typically on a monthly/yearly basis) and entail sharing partial part of the Infrastructure provider/MNO resources or the available unlicensed spectrum bands. This type of sharing complies with the rational of the traditional business model consisting of two entities: the infrastructure provider (InP), which has resources but no subscribers, and the MNO that in contrast has subscribers but has no infrastructure resources. The InP is the responsible entity of virtualizing resources to be used/shared by the different coexisting MNOs, with management operations of virtualized resources performed via interactions between both entities.

The expected shift to fully virtualized mobile networks with the ongoing and upcoming mobile network generations presents the key enabler for full sharing among coexisting mobile network operators. This would entail an evolution of the governing business model to cater for new business opportunities for telecom/network operators, manufacturers, and solution providers as well as for a range of new stakeholders. Indeed, compared to the traditional two-level business model, the MNO in the evolved three-tier business model can be further separated into

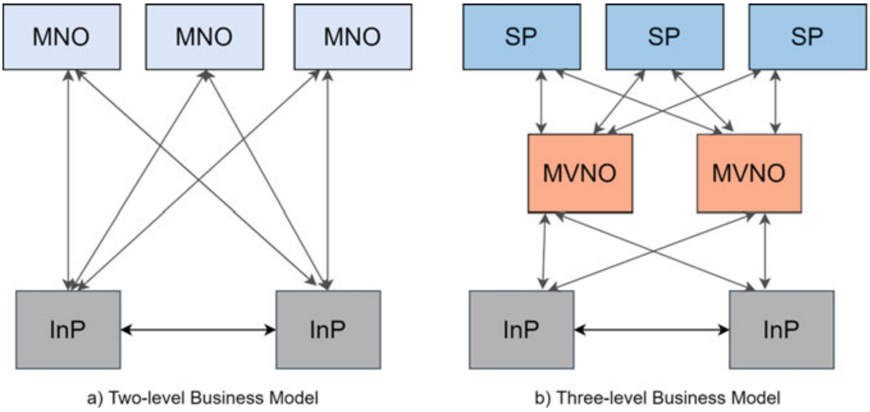


Fig. 10.1 Business model for virtualized mobile networks

two different specialized categories: the service provider (SP) and the MVNO, as shown by Fig. 10.1. The SP is the entity that has subscribers. The MVNO is the entity responsible for leasing the resource from one or multiple InP to satisfy the accumulated requests from each SP and hence the entity evolved in creating virtual resources based on these requests. The established sharing paradigm can be extended with other roles in future multi-tenants systems such as vertical segments/industries that lack network infrastructure but opportunistically or periodically need to reach their customers or enable services orthogonal to the telecommunication industry. The multi-tenant resource allocation operation should ensure the SLA (service-level agreement)/QoS (quality of service) of the different slices and cater for the adaptive capacity allocation, to enable opportunistic sharing of the mobile network infrastructure between the different vertical segments and services providers on time scales shorter than the contract agreement.

10.2.2.2 Enabling RAN Adaptive Sharing

The technical model for adaptive RAN sharing involves the main building blocks softwarization and autonomous management technologies, working in synergy for enabling adaptive RAN sharing/network slicing. Figure 10.2 depicts these building blocks. In this architecture, the RAN is fully virtualized thanks to the application of NFV for the virtualization of the RAN functions, SDR to slice the remote radio heads (RRHs), and SDN to manage the networking device. The RAN is controlled by a software-defined unified control plane (SD-UCP). NFV/SDR enables RAN sharing by different tenants MVNOs, while the SD-UCP translates the decisions of the enhanced SO-VRM algorithms back to the radio physical nodes, to enable the dynamic allocation and the flexible management of resources according to SLA and load from each MVNO. The unified control plane would allow that all

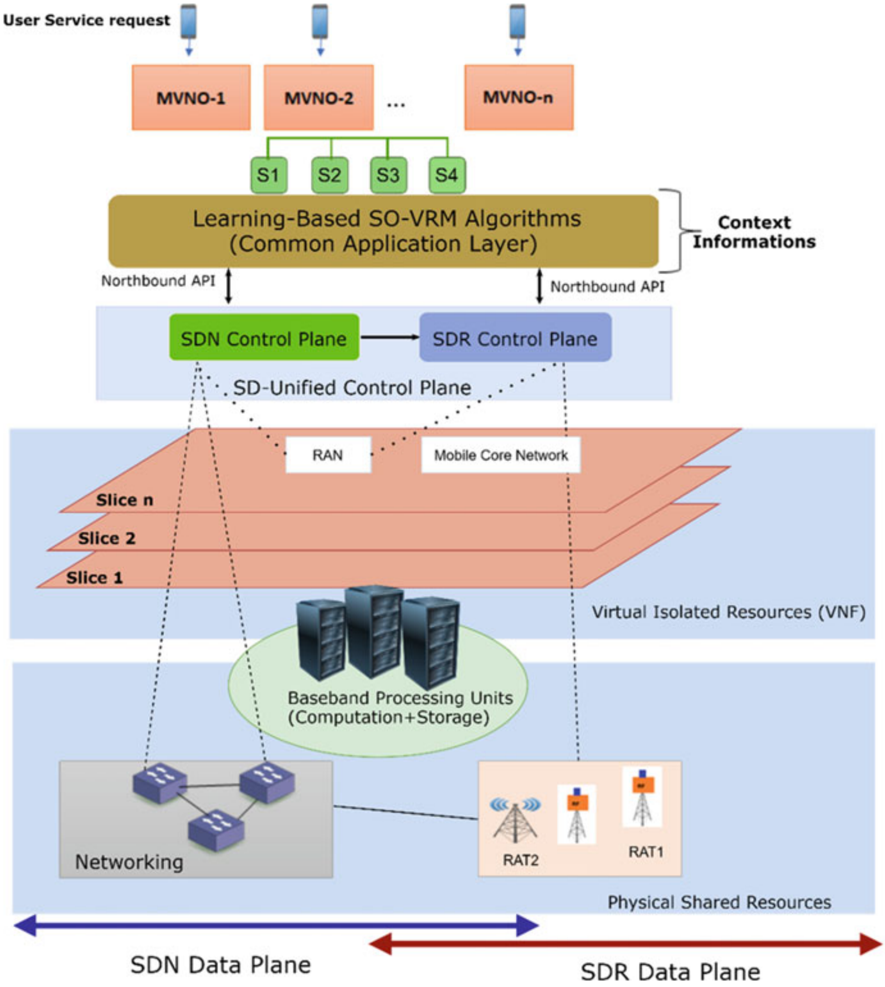


Fig. 10.2 Architecture for adaptive RAN sharing and application specific network slice. (© [2020] IEEE. Reprinted, with permission, from [32])

established slices c share the available bandwidth on the different existing RATs depending on the MNO policy/MVNO SLA. Moreover, it allows greater efficiency, enabling rational use of resources. To ensure isolation of the traffic from the different MVNOs, dynamic resource allocation should include a minimum throughput for each isolated MVNO's application specific slice. Allocation of resources reflects the MVNO's policy and slice QoS, as well as the MNO's preference to optimize a certain metric, such as cost saving, energy efficiency, or a trade-off between both.

10.3 Cost-Saving Design Strategies for 5G Network Infrastructures

Many 5G services have strict specifications in terms of latency and reliability performance [33]. Meeting these demanding requirements and designing a resilient network in a cost-efficient way are great challenges for the network operators. The hybrid cloud radio access network (H-CRAN) architecture depicted in Fig. 10.3 is a promising option to meet 5G service constraints.

An H-CRAN architecture consists of three tiers: remote radio units (RRUs), radio aggregation units (RAUs), and radio cloud centers (RCCs). The RAU node is responsible for serving all the RRUs connected to it, and the RAU nodes are connected to the RCC. Part of the baseband processing is implemented at the RAU, and the rest of it is done at the RCC. The network segment connecting the RAU and RCC is called the *midhaul*, and the segment between the RRU and the RAU is referred to as *fronthaul*. The next-generation fronthaul interface (NGFI) [35] and common public radio interface (CPRI/eCPRI) [36] are two options to transmit data over the fronthaul and midhaul segments.

There are a number of aspects that operators need to consider when designing their network infrastructures. Among them are resiliency and cost. More specifically, an H-CRAN architecture should support resiliency against the failure of the components or entire network nodes, while the cost of network deployment is minimized [34]. Additionally, from a cost perspective, it is beneficial for the operators to deploy services at centralized computing locations to be able to leverage the economy of scale of large data center sites [37]. However, meeting the service

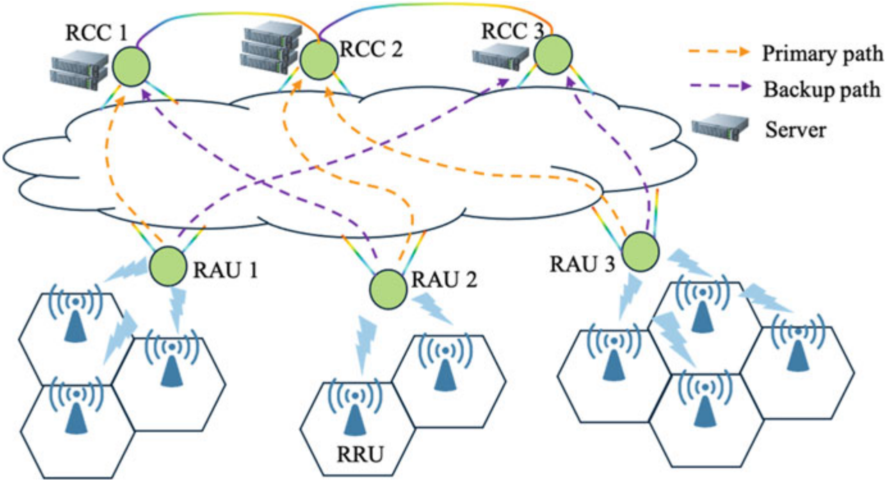


Fig. 10.3 An illustration of H-CRAN architecture with millimeter wave fronthaul and wavelength division multiplexing midhaul [34]

latency and availability requirements are challenges in a centralized deployment that are elaborated in the following.

In an H-CRAN architecture, a failure might happen at the RCC, RAU, RRU, or in any node/link in the fronthaul and midhaul segments. The number of users affected by a failure depends on the failure location. If the failure happens in the RCC, a large number of users will be affected, which makes significant impact on network reliability performance. Likewise, the failure in a midhaul node/link may cause service interruption for a subset of users associated with the RAU. Therefore, in order to improve the reliability performance, an H-CRAN architecture should be designed to minimize disruption due to failure of a server in the RCC, the whole RCC (due to a catastrophic event), and any midhaul node/link.

A possible way to ensure the survivability of services is to provision backup resources for connections between the RAUs and RCC in a *dedicated* or *shared* fashion. Dedicated protection methods fall into two categories: (a) 1 + 1, where backup resources are active and two live connections exist between the source and destination, and (b) 1:1, in which the backup resources are not active until a failure occurs in the primary path [38]. On the other hand, meeting the reliability performance requirement of a service by duplicating network and compute resources can be very expensive. Therefore, when possible, it is preferable to use more cost-efficient solutions, i.e., shared protection where backup resources can be shared among multiple services.

The cost efficiency of network infrastructure can be further increased by centralizing service processing in a few large data centers (DCs). Some studies promote distributed RAN architectures, which can take a step toward satisfying the quality-of-service constraints (i.e., in terms of latency and reliability performance). However, this approach is losing benefits introduced by centralization, i.e., cost reduction and easy deployment of RAN features [39]. Indeed, processing services in large-scale DCs will cost less than in the small and distributed DCs because of the economy of scale.

The main challenge to reach large and centralized DCs is guaranteeing the latency and reliability performance requirements which may be difficult due to the long distances to the large DCs. The only way to meet a latency constraint, using a specific technology (e.g., optical transport, millimeter wave), is to choose a DC close enough to the end user. However, the reliability performance can be improved by adding a redundant midhaul path. Unfortunately, the benefits of adding a protection path in terms of reliability performance will also introduce additional costs due to introducing backup resources that might adversely affect the overall cost savings of centralizing the service processing. For this reason, its impact needs to be analyzed carefully.

The research community has been looking into the aforementioned reliability and cost efficiency challenges. A number of works in the literature studied resilient design methods and cost-saving strategies [40–42]. The work in [40] considers centralized and distributed algorithms to place baseband unit (BBU) hotels in cloud radio access network (C-RAN) to ensure service continuity in case of single BBU hotel failure and compares their performance, scalability, and adaptability to changes in the network topology. The work in [40] shows that a distributed approach helps to off-load the SDN orchestrator and is able to cope with the evolution of C-RAN topology, whereas the changes in the original placement are limited. The authors in [41] bring up the benefits of "centralization" in terms of computational resources and power savings, as well as the importance of designing a survivable C-RAN network in case of failure. They proposed three approaches for survivable BBU hotel placement: (1) dedicated path protection, (2) dedicated BBU protection, and (3) dedicated BBU and path protection. Most of the interest in the literature is focused on designing a resilient C-RAN architecture. To provide survivability, the existing works either duplicate resources or consider sharing of the BBU ports in the BBU hotels. In order to improve cost savings, the authors in [42] assumed a given outage probability and used spatial traffic model and queuing theory to find the required number of transceivers in the considered scenario. Indeed, the transceivers in the fronthaul are used to work also at peak load, but the peak load conditions can be relatively rare. Therefore, by accepting a reasonable outage probability, the work in [42] shows that the fronthaul can be designed with lower capacity and fewer transceivers, which leads to cost and energy savings.

In order to guarantee survivability of services in the event of a failure in the RCC or any node/link in the midhaul in a cost-efficient manner, this chapter introduces a strategy called shared-path shared-compute planning (SPSCP). The proposed strategy decides on the location of a primary and backup RCC for each RAU and respective midhaul paths while allowing the sharing of the backup connectivity and computing resources. The SPSCP strategy has lower cost than equivalent approaches that use dedicated computing and midhaul connectivity resources for the backup and shows a cost improvement compared to those approaches where sharing is not encouraged while deciding on the location of the primary RCC nodes and midhaul path [34].

In addition, to benefit from cost improvements of centralized network deployment, this chapter investigates the trade-off between the savings derived from centralizing service processing and the additional cost due to the protection path used to meet the availability requirement of a given service. The performance evaluation shows that by centralizing service processing with the help of a protection path to meet the service reliability requirement, savings in overall infrastructure cost can be achieved [37] while not violating any latency constraint.

In Sect. 10.3.1, the network-planning strategy to meet reliability performance requirement is presented. The system architecture and use case are discussed in Sect. 10.3.1.1, while the proposed approach to design a resilient H-CRAN architecture and assessment of the results are presented in Sect. 10.3.1.2. Section 10.3.2 discusses the cost benefits of centralizing service processing. In particular, Sect. 10.3.2.1 presents the system architecture, latency and availability requirements, and cost model, and Sect. 10.3.2.2 discusses the economy of scale benefits and simulation results. Finally, conclusions are drawn in Sect. 10.3.3.

10.3.1 Shared-Path Shared-Compute Network-Planning Strategy

In order to meet the quality-of-service requirements of 5G services, a resilient network infrastructure should be provided. One possible method is adding backup resources although it increases the network deployment cost. A possible solution to reduce this cost is maximizing sharing of the backup connectivity and computing resources. The intuition behind this section is to derive a cost-efficient resilient network design strategy by exploiting the potential of sharing backup resources.

10.3.1.1 System Architecture and Use Case Description

We consider an H-CRAN architecture with a mesh wavelength division multiplexing (WDM) network for the midhaul segment as shown in Fig. 10.3.

We assume a single failure scenario, i.e., at most one failure can happen in the network at a time. An RRU failure can be handled by handover to other RRUs and is defined by the operators' handover policy. Further, we assume that the RRUs are dual homed to two different RAUs. Accordingly, a failure of RRU or RAU will not affect the overall availability. We assume that a failure can happen in a server in the RCC or the whole RCC can be down because of a catastrophic event. Also, a failure might happen in any node or link in the midhaul segment of the network. In order to satisfy the latency requirement, the number of hops between an RAU and its RCC node cannot exceed a given value, denoted by h, which is dictated by the latency requirements.

The target is to design a resilient H-CRAN architecture. In order to achieve this goal, one primary and one backup RCC should be assigned to each RAU, and the primary and backup connectivity paths between the RAU and RCC should be found. The design of the resilient network and allocation of the resources should be done with the objective of minimizing the network deployment cost. The cost is the summation of the total deployment cost of RCC nodes, server units within the RCC, and connectivity units, which is defined as:

$$C = N_{RCC}.C_{RCC} + N_{Ser}.C_{Ser} + N_{Conn}.C_{Conn} \tag{10.1}$$

where N_{RCC}, N_{Conn}, and N_{Ser} are the number of RCC nodes, connectivity units, and server units, respectively. C_{RCC}, C_{Conn}, and C_{Ser} are the cost of deploying one RCC node, one connectivity unit, and one server unit, respectively.

A key method to reduce the resilient network deployment cost is sharing the connectivity resources in the backup path and computing resources in the backup RCC. Two conditions should be met to enable the sharing of the computing resources, referred to as the *server sharing condition*. The RAUs can share a backup server in an RCC node if (1) their primary servers are located in different RCC nodes and (2) the paths to their primary RCC are node disjoint. Moreover, to share connectivity resources in the midhaul backup path, two conditions should be satisfied, referred to as the *connectivity sharing condition*. The RAUs can share connectivity resources in the backup path if (1) their primary servers are placed in different RCC nodes and (2) the primary paths to their RCC nodes are node disjoint.

Therefore, the primary and backup RCC nodes and the midhaul paths should be found by considering the above sharing constraints to minimize the overall cost. The details of the proposed strategy are described in the next section.

10.3.1.2 General Approach and Performance Evaluation

This section presents a strategy referred to as shared-path shared-compute planning (SPSCP) used to find the primary and backup RCC nodes of each RAU together with their connectivity paths.

Network-Planning Strategy

The strategy is a heuristic algorithm that chooses the primary and backup RCC with the lowest cost for each RAU [34]. In this algorithm, first, all RAUs are sorted based on the increasing value of the nodal degree of the midhaul nodes where they are located. This set is called $\mathcal{A}_s$, which is used to choose the primary and backup RCC. The midhaul nodes without RAU, i.e., set denoted by $\mathcal{G}$, can be chosen to place the RCC; thus, we assign a tag to these nodes called combined degree. We define the combined degree of each midhaul node as the summation of the nodal degree of the node and the number of RAUs that are within h hops from that midhaul node. We choose the midhaul nodes that are within h hops from the RAU, i.e., set $\mathcal{P}$, and sort them based on the decreasing value of the combined degree and get set $\mathcal{P}'$. Then, we evaluate the total cost of the network for all different options of choosing primary and backup RCC and select the option with the lowest cost. The shortest path algorithm is used to find the primary and backup connectivity paths. To calculate the cost, we use the cost function (10.1) described in Sect. 10.3.1.1 where we consider the possibility of sharing backup connectivity and computing resources. The cost of required resources for the backup is zero if they can be shared with backup resources of other RAUs. We repeat this procedure for all RAUs until we find primary and backup RCCs and their connectivity paths. The detailed steps of the algorithm are presented in Fig. 10.4 for a given value of the number of allowable hops (h).

Performance Evaluation

In this section, the performance of the SPSCP strategy is evaluated via simulations. We assume the same network parameters as the ones described in [34]. To obtain

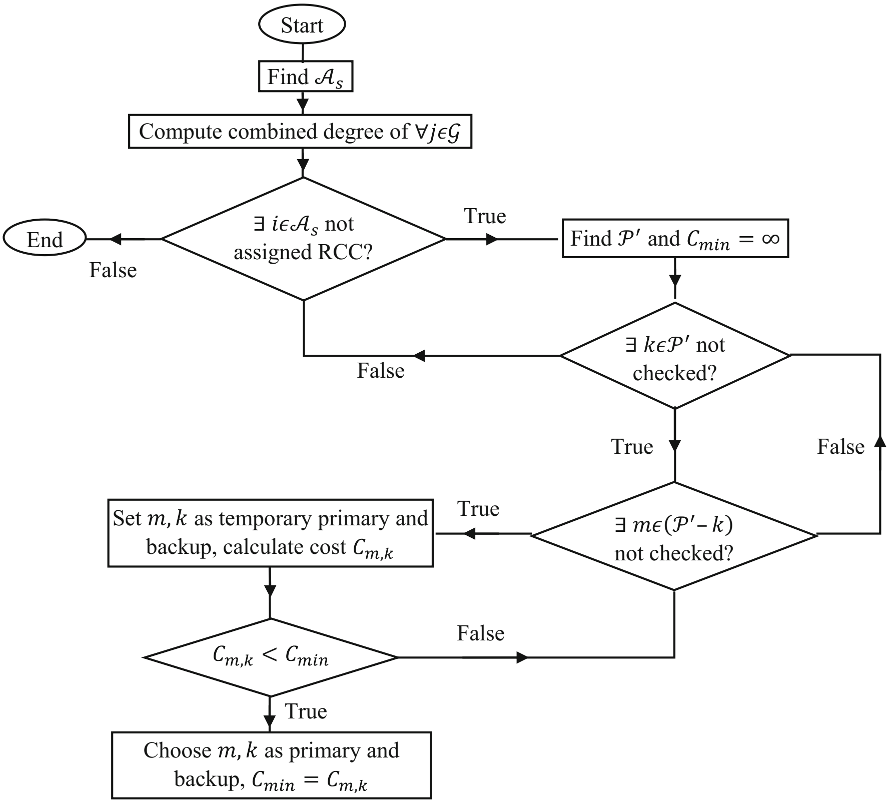

Start
Find $\mathcal{A}_s$
Compute combined degree of $\forall j \epsilon \mathcal{G}$
$\exists\ i \epsilon \mathcal{A}_s$ not assigned RCC?
End
False
True
Find $\mathcal{P}'$ and $C_{min} = \infty$
$\exists\ k \epsilon \mathcal{P}'$ not checked?
False
True
False
$\exists\ m \epsilon (\mathcal{P}' - k)$ not checked?
True
Set m, k as temporary primary and backup, calculate cost $C_{m,k}$
$C_{m,k} < C_{min}$
False
True
Choose m, k as primary and backup, $C_{min} = C_{m,k}$

Fig. 10.4 The flowchart of SPSCP strategy

the results, we set $C_{RCC} = 120$ cost units [CU], and connectivity and computing cost are changed to show their impact on the total cost.

The performance is evaluated against three benchmark algorithms. The first one is referred to as resource duplication (RD) where connectivity and computing resources are duplicated for the backup. The second one is referred to as preliminary resource sharing (PRS). PRS works exactly as RD but tries a posteriori to share backup resources where possible, i.e., without changing the pairing between RAUs and RCCs. The third one is called reconfiguration and improved resource sharing (RIRS). RIRS aims at improving the cost performance of the network designed according to RD. RIRS revisits the pairing between RAUs and their backup RCC nodes and the connectivity paths in order to maximize sharing of the backup resources.

The total cost of the network as a function of the allowable number of hops h for $C_{Conn} = 1$, $C_{Ser} = 6$ [CU] is shown in Fig. 10.5a. By relaxing the hop count constraint, i.e., by increasing the number of allowable hops, the cost decreases. The

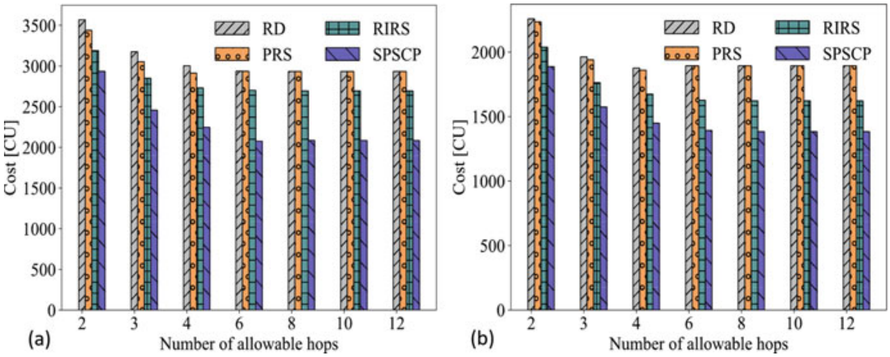


Fig. 10.5 Overall cost of network deployment as a function of number of allowable hops between RAU and RCC: (**a**) $C_{Conn} = 1$, $C_{Ser} = 6$ [CU], (**b**) $C_{Conn} = 6$, $C_{Ser} = 1$ [CU]

breakdown of the total cost of SPSCP shows that the cost of RCC deployment and backup servers are decreasing with the increasing number of hops between an RAU and its RCC node. This is the direct result of concentrating RCC nodes on a few midhaul nodes, although SPSCP considers the potential of sharing backup resources when it chooses the place of deploying the primary and backup RCC nodes. However, RD, PRS, and RIRS try to deploy RCC on fewer midhaul nodes without considering the shareability potential, which results in a lower cost reduction than SPSCP for higher h. Therefore, SPSCP shows better cost savings, which increases with increasing values of h.

The total cost of the network as a function of the number of allowable hops for $C_{Conn} = 6$, $C_{Ser} = 1$ [CU] is shown in Fig. 10.5b. The results show a similar trend as in Fig. 10.5a when the value of h increases. However, the cost savings of SPSCP with respect to RD, PRS, and RIRS are smaller than in the case shown in Fig. 10.5a. This is because in the scenario under investigation, the connectivity resources are the bottleneck, and the opportunity of sharing connectivity resources is limited. Therefore, increasing the connectivity unit cost results in lower cost savings. For the sake of comparison, the cost savings of SPSCP with respect to three other approaches for low and high values of h are shown in Table 10.1. It is evident that the cost savings of SPSCP with respect to all strategies when $C_{Conn} = 6$, $C_{Ser} = 1$ [CU] are lower compared to the case when $C_{Conn} = 1$, $C_{Ser} = 6$ [CU].

10.3.2 Cost Benefits of Centralizing Service Processing

Processing services in large-scale data centers is more cost-efficient than using distributed small computing nodes. On the other hand, large-scale data centers may be placed far from the users in centralized locations. Therefore, the potentially long propagation delay should be considered in the provisioning phase in order to make

Table 10.1 Cost savings of SPSCP with respect to RD, PRS, and RIRS for different connectivity and computing unit cost

Allowable hop count	Method	SPSCP cost saving, $C_{Conn} = 1$, $C_{Ser} = 6$	SPSCP cost saving, $C_{Conn} = 6$, $C_{Ser} = 1$
$h = 3$	RD	22.63%	19.79%
	PRS	19.55%	18.86%
	RIRS	13.81%	10.71%
$h = 12$	RD	28.91%	26.95%
	PRS	28.91%	26.95%
	RIRS	22.61%	14.71%

sure that the latency requirements are met. In addition, the availability requirements of services should be met, regardless of which computing nodes are used for service processing. In this section, we leverage upon the cost-effectiveness of the centralized deployment and propose a strategy to maximize the involvement of large-scale data centers, which also guarantees the quality of service in terms of latency and availability requirements.

10.3.2.1 Infrastructure Model, Service Requirements, and Cost

In this subsection, the network architecture, latency, availability, and cost models are presented.

Network Architecture

The considered network architecture is presented in Fig. 10.6. The RRU and RAU functionalities are placed in the same network element, referred to as access point (AP), while the RCC functionalities are deployed in the DC. The user equipment (UE) and AP are connected through wireless links. It is assumed that the AP is connected to a number of servers residing in the DC through an optical transport network. The transport network is composed of three segments with different transmission capacities. In this architecture, we have two points for aggregating and grooming traffic. The first aggregation point is on the boundary of the local and province segments. The second aggregation point is on the boundary of the province and regional segments.

We assume to have four types of DCs located in various segments of the network and with different characteristics in terms of size (amount of computing resources and the number of users that can be served) and cost efficiency, (the overall DC cost vs. the total number of users that can be supported). The DC types are the following: local, province, regional, and national. Local DCs are small, while the national DCs have the largest scale and are the most cost-efficient among other types.

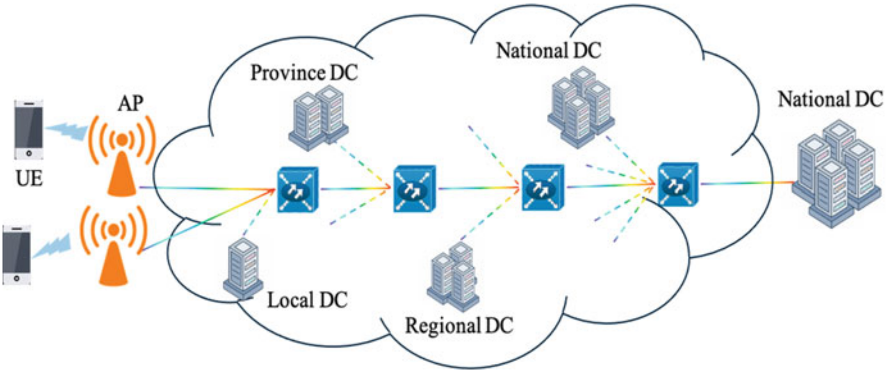


Fig. 10.6 The network architecture with three segments in the transport network [37]

Latency and Availability Requirements

The number of APs connected to small DCs is lower than the number of APs served by the large DCs. Small DCs are often close to the end user, and a lower number of nodes and links should be traversed to reach them. This option is offering lower latency and higher availability compared to processing services in large DCs. On the other hand, by centralizing service processing and using large-scale DCs, a higher number of APs can be served at one location, which has cost savings because of the opportunity to leverage on a better economy of scale. However, in this case, meeting the latency and availability requirements of the services can be challenging because large DCs are normally deployed far from the users, and services must traverse more components to reach those DCs. Therefore, the latency and availability constraints should be considered when designing the network.

The latency is the summation of latency of the UE, RAN, server, propagation, and switching latency of links and devices in the transport network. The availability is modeled as the product of availability of UE, RAN, all the nodes and links along the path from the AP to the DC, and the server. The latency value can be decreased only by choosing a server in a DC close to the UE. However, the availability can be improved by adding a protection path, referred to as protected (P) scenario, while in the unprotected (UP) scenario, only one path between AP and DC can be used. More details on the latency and availability computation are provided in [37].

Cost Computation

Our cost model includes the cost of computing, i.e., server in the DC, and connectivity resources in the transport network. We do not model the cost of the radio access network.

The total computing cost is a function of the required number of DCs, the number of servers in each DC, and the cost scaling factor (CSF) of a DC. CSF is used to calculate how many cost units need to be spent for the DC infrastructure (i.e., cooling, power, and networking equipment) out of each cost unit spent on servers. Clearly, the CSF of the national DC is the lowest among all types of DCs because of their economy of scale. Two different categories of switching nodes exist in the transport network. The cost of the nodes which are not performing traffic aggregation is modeled as the cost of their optical cross connects (OXCs), multiplexers (MUXs), and demultiplexers (DeMUXs). For the switching nodes performing traffic aggregation and grooming, the cost of packet switches and transceivers are added to the OXC, MUX, and DeMUX costs. Furthermore, the cost difference for transceivers with different transmission rates is reflected in the cost model. More details on the cost modeling assumptions can be found in [37].

10.3.2.2 Trade-off Evaluation

In this section, the trade-off between cost savings of centralizing service processing and the extra cost of providing a protection path to meet availability requirements is investigated.

Characteristics of the Network and Devices

We consider five use cases corresponding to five types of services along with their different latency and availability requirements as included in Table 10.2. For each use case, the maximum distance between the AP and DC should be calculated, according to the latency and availability requirements.

Table 10.3 provides the value of the key parameters for each type of DC considered in the study, i.e., CFS, DC distance from the AP, number of deployed servers in the DC, and the number of APs that can be connected to a DC (referred to as service density). The value of the availability of the RAN is assumed to be 99.999%. For the information on the value of the simulation parameters and the cost of the devices in the network infrastructure, we refer to [37].

Table 10.2 Considered use cases and their requirements

Use Case	Description	Latency	Availability	Reference
1	Augmented reality, collaborative gaming	12 ms	99.9%	[43]
2	Remote control for smart manufacturing	5.5 ms	99.99%	[43]
3	Discrete automation	20 ms	99.99%	[44]
4	Process automation/monitoring	20 ms	99.9%	[44]
5	V2X for short term environment modeling	10 ms	99.99%	[45]

Table 10.3 DC characteristic. The CSF of the regional, province, and local DCs are a function of η, i.e., national DC cost scaling factor. d_{CN} is the distance between DC and AP

DC type	d_{CN} range [km]	Service density	Num. server	Cost scaling factor
National	$d_{CN} > 100$	1000	250	$N_{EF} = \eta$
Regional	$100 \geq d_{CN} > 10$	100	25	$R_{EF} = 3 \times N_{EF}$
Province	$10 \geq d_{CN} > 1$	10	3	$P_{EF} = 2 \times R_{EF}$
Local	$1 \geq d_{CN}$	2	1	$L_{EF} = 2 \times P_{EF}$

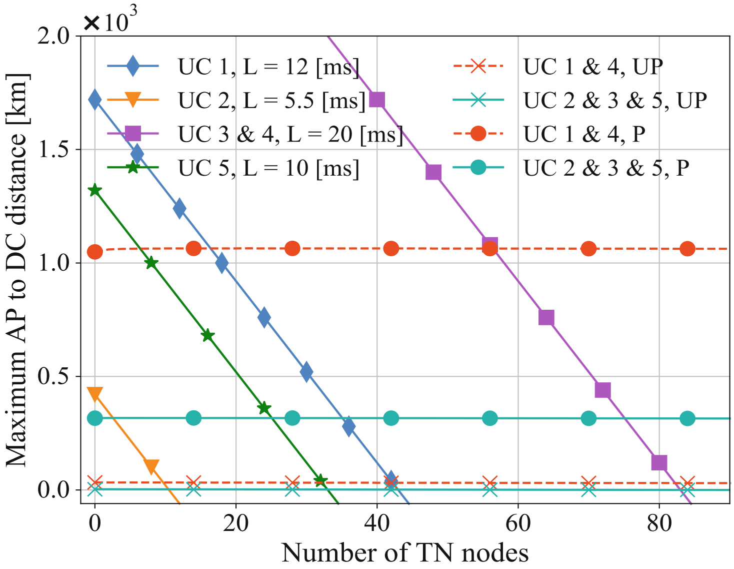


Fig. 10.7 Maximum distance between AP and DC for protected and unprotected cases

Cost Assessment and Results

Figure 10.7 shows the maximum distance between AP and DC as a function of number of transport network (TN) nodes for the different use cases listed in Table 10.2.

By adding one protection path, the maximum distance between the AP and DC for use cases 1 and 4 can be increased by up to 300 km while still meeting their latency requirements. For use cases 2, 3, and 5 in the protected scenario, a service can be deployed in DCs even further from the AP, i.e., in the 1000 km range.

Figure 10.8 depicts cost saving offered by a more centralized service processing. Use case 3 presents the highest cost saving (up to 63%) because it has a relaxed latency requirement which allows centralization, whereas the strict availability requirement can be met by adding a protection path. In addition, use cases 3 and 4 have the same latency requirements, but the cost saving of use case 3 is higher which shows availability requirement is the determining factor for the achieved gain. This

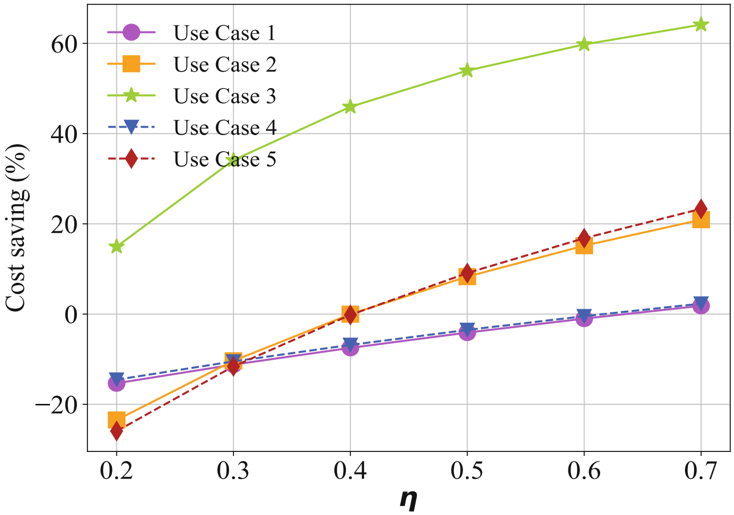


Fig. 10.8 Cost savings by adding a protection path in TN as a function of national DC CSF

result is also evident by comparing cost savings of use cases 1 and 4 since they have the same availability requirements.

10.3.3 Conclusion

In this chapter, cost-saving strategies considering the latency and reliability performance requirements in 5G network were proposed.

In the first scenario, the goal was to design a resilient H-CRAN architecture in which the failure can happen in any server of the RCC nodes, the whole RCC, or any link or node in the midhaul segment. The proposed cost-efficient strategy, referred to as shared-path shared-compute planning (SPSCP), assigns a primary and a backup RCC node to each RAU. To decrease the overall cost of the network, the SPSCP strategy tries to maximize sharing of the backup connectivity and computing resources. Adopting SPSCP strategy results in 26.9% cost savings compared to the approach that uses dedicated resources for the backup, and 14.7% cost savings compared to the method that does not consider shareability potential of the backup resources when assigning the primary RCC.

In the second scenario, in order to increase cost efficiency of the 5G network infrastructure, we investigated the benefits of processing services in large data centers in contrast to the small, distributed edge computing nodes. By considering services with different constraints, we guaranteed that all reliability and latency requirements of the deployed services are met. One protection path was added in the transport network to meet the availability requirements and processing services in centralized and large DCs. In spite of the extra cost of the redundant path, it still leads up to 63% cost savings because of the economy of scale offered by centralized service processing.

10.4 Energy-Efficient Virtual Resource Management for 5G and Beyond

We devote this section to energy-efficient virtual resources management approaches for 5G RANs and beyond. After a thorough analysis of existing research works, we shed the light on some perspectives and some opportunities arising in that sense. Interestingly, we present a performance evaluation of an EE efficient virtual resource management that tackles some of the identified challenges.

10.4.1 Review on Energy-Efficient Resource Allocation

Energy efficiency has always been in the spotlight of the cellular networking, especially with the shift of the design and planning logic from traditional rigid planning considering the peak traffic demand to more adaptive schemes leveraging from the flexibility of the virtualization or self-organization to bring more energy efficiency to the RAN. Most of the research contributions in that sense focused on the optimization of the RRH resource elements. Other works targeted the optimization of EE through the design of computational resource allocation schemes to map load from RRHs to an optimal number of baseband units in the cloud. The joint radio and computational resource allocation has recently been addressed by several research works, aiming to cater for the hybrid nature of resources in Cloud-based RAN environments.

10.4.1.1 EE Radio Resources Allocation

Radio resource management research works include dynamic radio resource allocation on physical resource blocks (PRBs) and/or a power allocation to users, depending on their channel state information and the required data rate, with the aim of optimizing EE. Other radio-related tasks include user pairing to an optimal RRH or to a set of RRH antenna resources, in a coordinated multipoint (CoMP) and/or beamforming design. In particular, recent research works on CoMP [46–50] provide evidence on the traditional benefit of interference mitigation where interfering signals from neighboring RRHs are used constructively to provide diversity gain enhancing the reliability of the received signal. Moreover, applying self-organization to the RRHs by selectively and automatically powering on/off according to the load variation has also been extensively addressed. In particular, the allocation tasks have been formulated into mixed combinatorial or non-convex optimization problems and solved after decomposition to elementary steps. For instance, RRH selection and RRH on/off problems have been solved via heuristic algorithms, such as greedy activation [51–53]. Power/bandwidth allocation has been solved by relaxation and decomposition techniques [54] or game theory [55]. Table 10.4 provides a summary of these EE radio resource allocation works.

10.4.1.2 EE Computational Resources Allocation

A second existing approach for energy efficiency improvement is to minimize the power consumed at the BBU pool by minimizing the number of BBUs at the cloud, considering RRH that can be grouped into a cluster and mapped to one BBU. In the literature, the problem has been mostly formulated as bin packing minimization (BPM) problem and solved via [69, 70], or meta-heuristic [71]. Although these efforts proved to be efficient for minimizing the active number of BBUs, they do not account for the user QoS requirement and the level of interference in the network, when forming RRHs clusters [72–74]. Different from the constraint-oblivious behavior of these approaches, recent works included the QoS constraint by formulating the problem to a modified BPM [75] or to set the partitioning problem (SPP) [76]. Other research efforts considered service time constraint along with the power minimization problem through the design of a BBU workload-scheduling scheme [77], while [78] presents another research strand toward more realistic BBU resource allocation by reshaping the problem into a virtual BBU minimization problem. Table 10.5 provides a summary of these BBU resource allocation works.

10.4.1.3 EE Hybrid Resources Allocation

In contrast to the aforementioned approaches that considered computational or radio optimization aspects independently, recent approaches have targeted the design of EE multi-resource allocation taking into account both resources, as summarized in Table 10.6. We refer to these collectively as hybrid resource allocation, e.g., [84] aims to minimize the overall system power consumption for C-RAN by optimizing the number of virtualized BBU, set of selected RRHs, and the beamforming vector at active RRHs. It is worthy to note that none of the hybrid research works considered the distributed antenna system behavior of a set of RRH mapped to the same BBU and consequently the associated relationship between the radio and computational resource allocation. QoS constraints consideration has also been overlooked. Furthermore, none of these schemes considered leveraging on the baseband servers' virtualization gain and including this in the cross-layer optimization problem by enabling the dimensioning of the optimal number of virtual BBUs as a trade-off between QoS and reduced energy consumption.

Table 10.4 EE radio resource allocations works

Ref	[46], 2014	[47],2014	[56], 2014	[57],2015	[58].2015	[59], 2016	[51],2016	[60], 2016	[61], 2016	[52], 2016	[53], 2016	[62], 2016	[50], 2016	[55], 2016	[63], 2017	[64],2017	[65], 2017	[66], 2017	[67], 2017	[68], 2018	[48], 2018	[54],2018
Subchannel/PRB allocation			✓			✓	✓		✓	✓					✓	✓						✓
Power allocation		✓	✓						✓					✓	✓					✓		✓
CoMP	✓	✓															✓				✓	
Beamforming				✓				✓			✓	✓	✓				✓		✓		✓	
User-RRH pairing			✓	✓	✓		✓	✓		✓	✓			✓	✓	✓		✓		✓		✓
RRH *on/off*	✓				✓		✓			✓	✓		✓				✓	✓	✓	✓		
InP/MVNO																			✓			
Simulation-based evaluation	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓			✓
Evaluation-based implementation																						

Table 10.5 EE BBU resource allocations works

Ref	[79], 2012	[72],2013	[71], 2015	[80], 2015	[69], 2015	[70], 2016	[81],2017	[73], 2017	[77],2017	[74],2017	[82],2018	[83], 2018	[78], 2018
BBU scheduling		✓							✓				
RRH-BBU mapping	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	
Virtualized BBUs		✓	✓					✓		✓			
Load-dependent BPS power consumption	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
BBU *on/off*				✓	✓	✓		✓	✓	✓	✓		✓
InP/MVNO													
Simulation-based evaluation	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		
Implementation based evaluation		✓											

© [2020] IEEE. Reprinted, with permission, from [32]

Table 10.6 EE hybrid resource allocations works

Ref	[85], 2014	[84], 2015	[86], 2016	[87], 2016	[88],2016	[89], 2017	[90], 2017	[91], 2017	[92], 2018	[93], 2018	[94],2019
Subchannel/PRP allocation	✓		✓							✓	✓
Power allocation		✓	✓								✓
CoMP							✓	✓			
Beamforming		✓			✓		✓			✓	
User-RRH pairing		✓	✓		✓		✓	✓	✓	✓	✓
BBU scheduling		✓			✓						
RRH-BBU mapping			✓				✓		✓		✓
Virtualized BBUs	✓	✓	✓	✓	✓		✓	✓		✓	✓
Load-dependent BPS power consumption			✓	✓	✓	✓	✓		✓	✓	✓
RRH *on*/*off*		✓	✓	✓				✓	✓	✓	
BBU *on*/*off*	✓			✓			✓	✓	✓		✓
InP/MVNO											
Simulation-based evaluation	✓	✓	✓	✓	✓		✓	✓		✓	✓
Implementation evaluation-based									✓		

© [2020] IEEE. Reprinted, with permission, from [32]

10.4.2 Challenges Toward Virtual Resource Management for C-RANs

Despite the valuable contributions that the research community brought in terms of hybrid resource allocations, there are still some open challenges that need to be addressed. One primary challenge is related the design of EE multi-operator (EE-MO) hybrid RA. It is worthy to note that [67] presents the only work about multi-operator resource allocation. The scheme considers cooperating MVNOs with inter-band noncontiguous carrier aggregation, by segmenting the licensed spectrum of each into private and shared bands, in which UE access is mutually exclusive. Results of the proposal evaluation proved an improvement in terms of energy efficiency and spectrum efficiency. The work, however, caters for RRH antenna resource sharing among operators solely, and not for the BBU sharing. In this context, an interesting challenge here lies in the design of energy aware multi-operators RAN resource management operations that relies on powerful SO algorithms and the fully virtualized RAN infrastructure to allow multiple operators to coexist and adaptively share the RAN resources while reducing the energy consumption on the RAN. Precisely, the unified software-defined control plane would leverage big data and ML algorithms and use as input context information collected from all network resources while considering each of the MVNOs' SLA, instantaneous load, and minimum required QoS. Using ML capability, the collected data will be analyzed, and operation-specific optimizations models developed. Ultimately, these optimization models will be applied, and the output is a set of optimal parameters to be adapted by the different radio and virtual RAN resources.

Moreover, the aforementioned approaches should cater for the hybrid nature (radio and computational) of the resource and include new optimization schemes that cater for not only intra-operators' scale but also inter-operators' scale. Regarding the multi-operator radio resource operations, a promising idea to improve energy efficiency and coverage is the use of UEs belonging to a given MVNO to act as a small cell upon coverage whole detection [95]. As for the multi-operator computational resource operations, it would ensure VNF placement at the various physical network locations (for NFV-enabled resources), and their efficient migration from underutilized and high energy cost to low-energy physical locations, targeting the minimization of the energy consumption. This requires proactive ML frameworks that can predict the future traffic among MVNO's users and consequently pre-organize optimal NFV.

10.4.3 EE Hybrid Resource Allocation for C-RAN

Given the aforementioned state of the art, we identified several open research challenges, among others, the need to explore more EE hybrid resource allocations that cater for both types of resources and maintain QoS aware behavior when performing RRH to computational BBU mapping. To this extent, we propose an EE resource allocation scheme that aims to minimize the power consumption at the BBU pool when mapping RRH group to BBUs, while considering users' QoS and BBU capacity constraints. The objective of our RRH group-based mapping (RGBM) scheme is twofold. First, it aims to improve the throughput of users experiencing bad radio conditions, while meeting a minimal required throughput for all users in the network. Second, it targets the minimization of the number of BBU units aiming to reduce power consumption and increase the spectral efficiency, while maintaining the minimal throughput required for users. To achieve this, the proposed scheme uses two key steps: (i) the formation of cooperative RRH groups aimed at improving the QoS of weak users and (ii) the formation of RRH cluster to be mapped to a minimal number of BBUs without violating the minimum user QoS demands.

The considered system is a C-RAN composed of a set of N distributed RRHs, R where, $R = \{r_1, r_2, .. \, r_N\}$ connected to a BBU pool through high-performance links, for centralized baseband processing and resource block scheduling tasks.

The cooperative RRH grouping performed in the first step of our scheme targets the improvement of weak users' conditions. Formed groups are denoted as the set $G = \{g_1, g_2, .. \, g_N\}$. The joint transmission coordinated multipoint (JT-CoMP) and transmission point selection (TP selection) is considered among a group of cooperative RRHs. The group formation procedure starts by identifying weak users, i.e., users experiencing a signal to noise ratio (SINR) below a fixed threshold. Then, it considers for every weak user a list of the most interfering RRHs for group formation. An RRH candidate on that list would potentially join the group if the grouping conditions with respect to the minimal required throughput are met for other users.

Once the groups are formed, the second step aims to form clusters of RRH groups. The set of formed clusters is denoted $B = \{b_1, b, .. \, b_N\}$. These clusters are mapped according to a one-by-one association relation to a set of baseband units at the pool level. Within one cluster, members are RRHs groups including singleton RRH groups.

We consider also that all RRHs start operating with a frequency reuse of one. Then, the scheme relies on attributing the distributed antenna system (DAS) behavior to formed groups and then formed clusters. That is, attempts for bandwidth sharing are first considered among small-scale groups, which consist of cooperative RRHs following the RRH group formations steps, and then for larger scale groups which are composed of RRHs cluster (BBU units) during the clustering formation procedure.

10.4.3.1 Defined as a Linear Programming Problem

In this section, we provide the mathematical modeling of the RGBM scheme.

We start by deriving the throughput, after and before the group formation procedure for weak and normal users.

Initially, the SINR of a weak user u_e connected to an RRH r on resource block rb is:

$$\Gamma_{e,r}^{rb} = \frac{Pt_r \; l_{e,r} \; h_{e,r,rb}}{\sum_{r' \in R, r \neq r'} Pt_{r'} \; l_{e,r' h_{e,r,rb}} + N_0} \tag{10.2}$$

where Pt_r, $l_{e,\,r}$ $h_{e,\,r,\,rb}$ denotes the transmission power of RRH r, the path loss, and the small-scale fading between user u_e and RRH r, respectively.

The SINR and throughput of a weak user being served by a formed group of cooperative RRHs (g) would be improved according to (10.3 and 10.4), respectively.

$$\Gamma_{e,r}^{rb} = \frac{\sum_{r \in g} P\mathrm{t}_r \; l_{e,r} \; h_{e,r,rb}}{\sum_{r' \in g', g \neq g'} Pt_{r'} \; l_{e,r' h_{e,r',rb}} + N_0} \tag{10.3}$$

$$Th_{e,r} = B_0 \log_2 \left(1 + \Gamma_{n,g,r}^{rb}\right) \tag{10.4}$$

where B_0 denotes the subchannel bandwidth.

The SINR and throughput of a normal user part of a formed group of cooperative RRHs, g, are provided by Eq. 10.5 and 10.6, respectively.

$$\Gamma_{n,r}^{rb} = \frac{Pt_r \; l_{e,r} \; h_{e,r,rb}}{\sum_{r' \in g', g \neq g'} Pt_{r'} \; l_{e,r' h_{e,r',rb}} + N_0} \tag{10.5}$$

$$Th_{n,r} = B_0 \log_2 \left(1 + \Gamma_{n,g,r}^{rb}\right) \tag{10.6}$$

We assume that the initial bandwidth allocated to each RRH is B_g and that upon the group formation procedure, cooperative RRHs part of the group share this bandwidth. We assume that there are a number RB_g of resource blocks per TTI to be assigned to all users of group, U_g given the bandwidth B_g.

The cumulative throughput experienced by a group g is hence:

$$Th(g) = \sum_{u \in U_g} \sum_{r \in RB_g} Th_{u,r} \tag{10.7}$$

where

$$Th_{u,r} = \begin{cases} Th_{e,r} \text{ if u is weak user} \\ Th_{n,r} \text{ if u is normal user} \end{cases} \tag{10.8}$$

We use a binary variable denoted $x_{b,\,g}$ which determines the association of a group g to a BBU b such as:

$$x_{b,g} = \begin{cases} 1 \text{ if group } g \text{ is associated to BBU } b \\ 0 \; otherwise \end{cases} \tag{10.9}$$

That said, the throughput experienced by cluster b is defined by (10.10):

$$Th(b) = \sum\nolimits_{g \in G} x_{b,g} \; Th(g) \tag{10.10}$$

Regarding the power model at the BBU pool, we consider a power consumption at each BBU that varies linearly as a function of the load processed in terms of offered throughput. The power consumed by a BBU b is, hence, provided by Eq. 10.11:

$$PC(b) = \tau + \mu \; Th(b) \tag{10.11}$$

where τ and μ reflect the power consumption by active BBU b with no traffic and the coefficient varying with the traffic, respectively.

We formulate in (10.12) the optimization utility function (UT) for cluster formation reflecting the targets of the RGBM scheme in terms of power minimization and meeting the minimum required throughput for all users:

$$\text{Minimize}(UT) = y_b \sum_{b \in B} PC(b) \tag{10.12}$$

where

$$y_b = \begin{cases} 1 \text{ if cluser } b \text{ is chosen} \\ 0 \; otherwise \end{cases} \tag{10.13}$$

Subject to:

$$\text{C1} : \sum\nolimits_{b \in B} y_b x_{b,g} = 1 \tag{10.14}$$

$$\text{C2} : \frac{Th(b)}{N_b} \geq Th_{min} \tag{10.15}$$

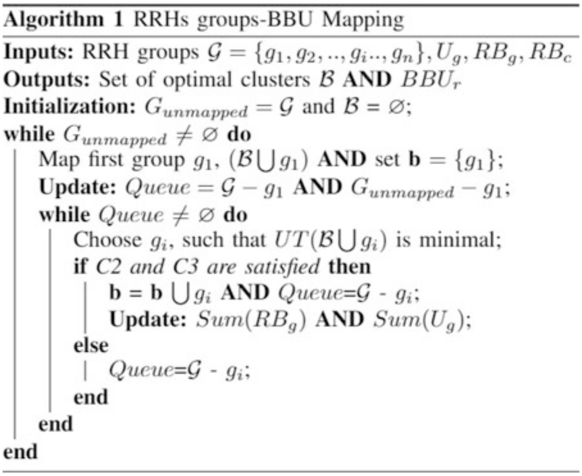
Constraint C1 ensures that a group g can be mapped to only one cluster b. Constraint C2 reflects that the number of users N_b processed by the same cluster b must satisfy a minimum required throughput.

This problem formulation belongs to the integer linear programming class and is NP-Hard.

10.4.3.2 Optimization vs. Heuristics

In this section, we provide a low complexity solution to the problem defined in the previous section. The solution relies on the use of an efficient greedy approach for the RRH group-based mapping [96]. Algorithm 1 depicts the proposed solution. The inputs for the algorithms are the set of groups G, the number of users at each group U_g, the number of resource blocks used by each group, and the number of resource blocks available at each cluster, R_c.

At each iteration, the RGBM algorithm selects an RRH group that minimizes the utility function (UT). The group in only mapped to the currently filled cluster b, if it meets C1 and C2.



```
Algorithm 1 RRHs groups-BBU Mapping
Inputs: RRH groups $\mathcal{G} = \{g_1, g_2, .., g_i .., g_n\}, U_g, RB_g, RB_c$
Outputs: Set of optimal clusters $\mathcal{B}$ AND $BBU_r$
Initialization: $G_{unmapped} = \mathcal{G}$ and $\mathcal{B} = \varnothing$;
while $G_{unmapped} \neq \varnothing$ do
    Map first group $g_1$, $(\mathcal{B} \bigcup g_1)$ AND set b = $\{g_1\}$;
    Update: $Queue = \mathcal{G} - g_1$ AND $G_{unmapped} - g_1$;
    while $Queue \neq \varnothing$ do
        Choose $g_i$, such that $UT(\mathcal{B} \bigcup g_i)$ is minimal;
        if $C2$ and $C3$ are satisfied then
            b = b $\bigcup g_i$ AND $Queue$=$\mathcal{G}$ - $g_i$;
            Update: $Sum(RB_g)$ AND $Sum(U_g)$;
        else
            $Queue$=$\mathcal{G}$ - $g_i$;
        end
    end
end
```

10.4.3.3 Performance Evaluation

We evaluated the performance of the RGBM approach with MATLAB-based simulations, comparing it to two state-of-the-art schemes for mapping, those being the bin packing minimization (BPM)-based mapping and the conventional one-to-one (OTO) mapping. The BPM represents the classical resolution method of the state of the art which accounts only for a fixed per BBU capacity as a constraint and overlooks the users' QoS requirement. On the other hand, the OTO represents the conventional scheme used in distributed RAN where each RRH is mapped to one BBU. The considered C-RAN architecture comprises 19 RRHs along with a uniform distribution of users [96]. The performance evaluation accounts for different performance metrics, being the number of required BBUs, total power consumption, average users' throughput, and energy efficiency.

Figure 10.9 depicts the comparison in terms of required BBUs as a function of the number of UEs per RRH. As can be shown, the OTO scheme exhibits the highest number of BBUs. This number reflects the total number of RRHs in the network. The use of the RGBM (proposed approach) leads to the lowest number

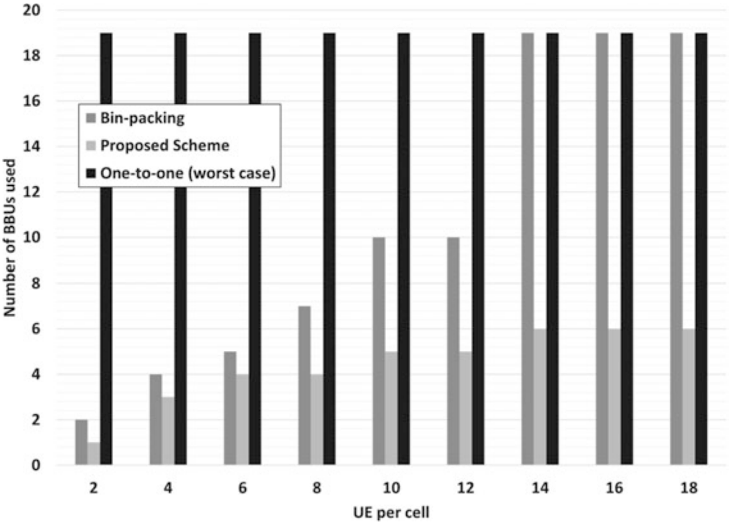


Fig. 10.9 Comparison of number of required BBUs. (© [2020] IEEE. Reprinted, with permission, from [96])

of BBUs, considering all users densities, whereas the BPM-based mapping shows intermediate usage that increasingly worsens as the user density per cell increases. Particularly for user densities equal to or higher than 14 UEs/cell, the BPM reaches its limit by activating one BBU for each RRH similar to the OTO scheme. The outperformance of the RGBM scheme in maintaining the lowest BBUs can be justified by its interference and QoS aware nature that tends to maximize the number of RRHs groups belonging to the same BBU, as long as the minimum required throughput of processed users is satisfied. Hence, our scheme succeeds to maximize the spectral efficiency by maximizing the share of the BBU, which leads to the use of the lowest number of BBUs.

Figure 10.10 illustrates the comparison between the three schemes in terms of induced BBU pool power consumption as a function of the number of UEs per RRHs. As shown by the figure, the results obtained with RGBM and BPM are well in accordance with the linear power consumption model used and explained in Sect. 10.4.3.1. Indeed, for both schemes, the total power consumption increases as the number of UEs increases. Nevertheless, the power consumption is fixed for the nonadaptive scheme OTO, where each RRH is assigned to a BBU operating at its maximal power. The results illustrated in this figure prove that our proposed scheme succeeds to maintain the lowest power consumption for all user densities, followed by the BPM and the OTO. The better power-saving capability of our proposed scheme directly emanates from the ability to effectively reduce the number of instantiated BBUs. In particular, for user densities higher than 12 UE/cell, the adaptive BPM scheme reaches the limit in terms of power savings compared to the conventional OTO. This is due to the activation of all BBU as reported in Fig.

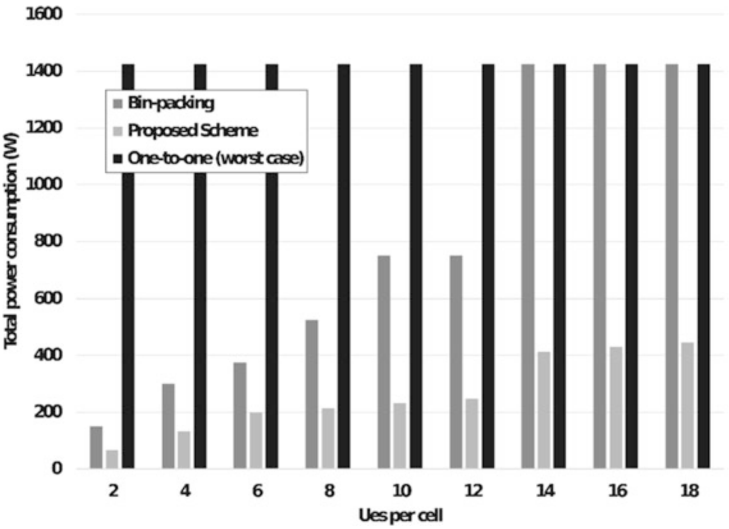


Fig. 10.10 Comparison of total power consumption. (© [2020] IEEE. Reprinted, with permission, from [96])

10.8. In contrast, our proposed mapping scheme succeeds to bring significant power savings, even for this high-density scenario, in contrast to the two other schemes.

Figure 10.11 shows the comparison results with respect to the average user throughput. As one can deduct, all schemes succeed to maintain an average throughput equal to or higher than the required minimal throughput, with OTO showing the highest throughput followed by BPM, and lastly the RGBM. Indeed, the increase in the spectral reuse and energy saving achieved by our scheme comes at a cost, in terms of a slight reduction in the average user throughput while still satisfying users' target QoS.

Figure 10.12 comes to quantify the trade-off between acceptable throughput addressing the user QoS requirements and achieving low power consumption. An ideal metric for capturing this trade-off is energy efficiency. In this context, this metric (bits per joule) reveals that the OTO is the least energy-efficient due to its extreme power consumption. Both optimized mapping schemes (RGBM and OTO) show competitive results for low user densities, with our scheme achieving a distinctive trade-off for medium and high user densities.

It is worth noting that the minimization of the number of required BBUs does not only impact the overall power consumption and energy efficiency on the C-RAN but also creates less overhead in the front haul of the network. Indeed, according to the RGBM approach, groups are formed in a distributed way and the information required for cluster formation would only be sent per group, instead of per RRH.

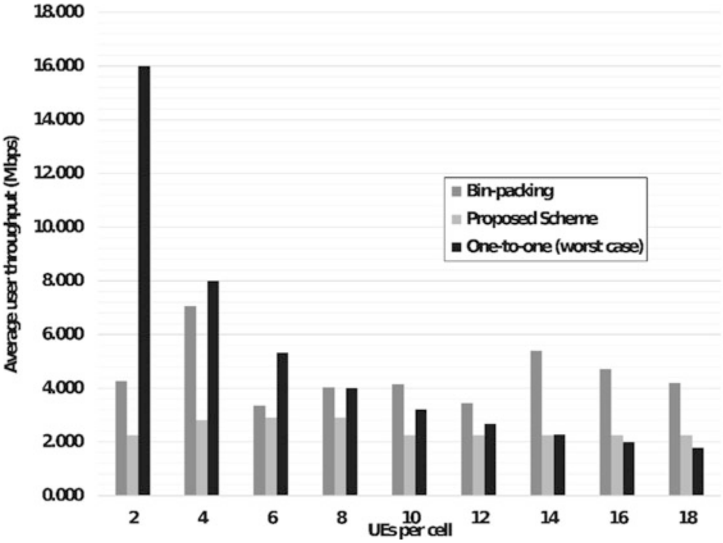


Fig. 10.11 Comparison of average user throughput. (© [2020] IEEE. Reprinted, with permission, from [96])

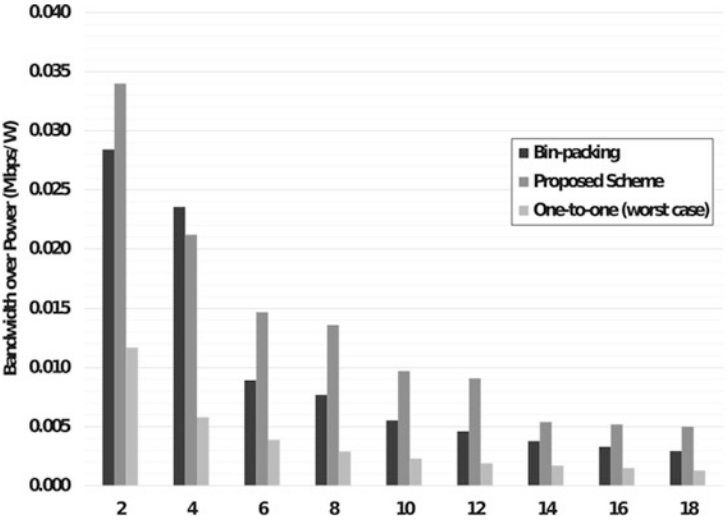


Fig. 10.12 Comparison of bandwidth over power. (© [2020] IEEE. Reprinted, with permission, from [96])

10.4.3.4 Conclusions

In this section, we presented a review on efficient resource allocations works with respect to energy efficiency. Based on the presented review, we identified some of the open research challenges and future research directions. Then, we proposed a group-based RRH to BBU mapping (RGBM) for minimizing the number of BBUs instantiated in the C-RAN. Different from state-of-the-art approaches, the BBU minimization heuristic is further optimized, thanks to the first stage of cooperative RRH groups formation based on QoS requirement of weak users. By considering the DAS behavior on formed BBU (clusters), we can establish that our proposed solution leads to the formations of more optimal clusters, that is, a better adjustment of the level of interference on the network, when compared to other SoTA schemes. This leads to a better improvement of the initially detected weak users' radio and hence to the capability of consolidating more RRHs to a BBU while still satisfying the individual users QoS, when compared to the interference and QoS oblivious approaches for mappings. Simulations results have demonstrated that the aforementioned features endow our proposed solution with a higher gain in terms of power-saving capability and energy efficiency, when compared to two SoTA schemes. The presented results prove that catering for the users' radio quality conditions as well as their QoS requirement in the RRH to BBU mapping can bring considerable power savings and energy efficiency to the C-RAN.

10.5 Conclusion

Softwarization and autonomous management technologies are expected to play major roles in future emerging mobile networks (5G and beyond). In this chapter, we have provided an overview of these technologies and elaborated on how they can be harnessed to provide in a broad sense greater revenue for mobile stakeholders. Moreover, the chapter elaborated design targets for the virtual infrastructure, which include greater resiliency, cost saving, and energy efficiency.

Toward designing a resilient and cost-efficient H-CRAN architecture, we presented two novel network-planning strategies. In the first instance, we proposed the SPSCP strategy, as a resilient design strategy that assigns a primary and a backup RCC node to each RAU. Besides providing resiliency, the strategy is also capable of achieving cost efficiency thanks to the maximization of the sharing in the backup connectivity and in the computing resources. Simulation results have shown that our proposed strategy demonstrates 26.9% and 14.7% of cost savings compared to the dedicated resources and the non-shareable potential of the backup resources, respectively, when assigning the primary RCC. Secondly, we investigated the benefits of processing services in large-scale data centers in contrast to the small scale and proved that the addition of a protection path in the transport network, besides providing reliability, had the potential to provide 63% cost savings thanks to the economy of scales obtained from centralized service processing.

Toward pushing further energy saving gains in the network, we investigated EE-MO-RA considering both computational and radio resources. In this context, we proposed an RRH group-based mapping (RGBM) scheme that aims to first improve weak users' radio conditions through the formation of cooperative RRH groups, followed by the subsequent minimization of the C-RAN power consumption through an efficient greedy heuristic for mapping. The simulation results presented that factoring in the level of interference in the network and the user QoS leads to a considerable gain in terms of power saving and energy efficiency when compared to the baseline schemes (BPM-based mapping and the conventional OTO).

The set of presented solutions throughout this chapter provides a foundation toward more efficient mobile network planning and resource allocation solutions for B5G systems in terms of resiliency, cost, power consumption, and energy efficiency.

Acknowledgments This work was funded by the MSCA-ITN project 5G STEP FWD with funding from the European Union's Horizon 2020 research under grant agreement number 722429.

The research leading to these results has also received funding from the European Union's Horizon 2020 research and innovation program under grant agreement H2020-MSCA-ITN-2016 SECRET-722424.

References

1. Karsenti, E. (2008). Self-organization in cell biology: A brief history. *Nature Reviews. Molecular Cell Biology, 9*(3), 255–262.
2. Yates, F. E. (1983). *What is self-organization?* Princeton University Press.
3. Marwangi, M. M. S. et al. (2011). Challenges and practical implementation of self-organizing networks in LTE/LTE-Advanced systems. In *Proceedings of international conference information technology multimedia* (pp. 94–100).
4. Gacanin, H., & Ligata, A. (2017). Wi-fi self-organizing networks: Challenges and use cases. *IEEE Communications Magazine, 55*, 158–164.
5. Marchetti, N., Prasad, N. R., Johansson, J., & Cai, T. (2010). Self-Organizing Networks: State-of-the-art, challenges and perspectives. In *Proceedings of international conference communication* (pp. 503–508).
6. Alsedairy, T., Qi, Y., Imran, A., Imran, M. A., & Evans, B. (2015). Self organising cloud cells: A resource efficient network densification strategy: T. Alsedairy et al. *Transactions on Emerging Telecommunications Technologies, 26*(8), 1096–1107.
7. Park, J., & Lim, Y. (2016). Adaptive access class barring method for machine generated communications. *Mobile Information Systems, 2016*, 6.
8. Ramiro, J., & Hamied, K. (Eds.). (2011). *Self-organizing networks: Self-planning, self-optimization and self-healing for GSM, UMTS and LTE*. Wiley.
9. Fernández-Segovia, J. A., Luna-Ramírez, S., Toril, M., & Úbeda, C. (2016). A fast self-planning approach for fractional uplink power control parameters in LTE networks. *Mobile Information Systems, 2016*, 11.
10. Aliu, O. G., Imran, A., Imran, M. A., & Evans, B. (2013). A survey of self organisation in future cellular networks. *IEEE Communications Surveys & Tutorials, 15*(1), 336–361.
11. Choi, H.-H., Kwon, S.-C., Ko, Y., & Lee, J.-R. (2016). *Self-organization in mobile networking systems*.
12. Moysen, J., & Giupponi, L. (2018). From 4G to 5G: Self-organized network management meets machine learning. *Computer Communications, 129*, 248–268.
13. Fadlullah, Z. M. et al. State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems. *IEEE Communications Surveys & Tutorials, 19*(4), 2432–2455, Fourth quarter 2017.

14. Kibria, M. G., Nguyen, K., Villardi, G. P., Zhao, O., Ishizu, K., & Kojima, F. (2018). Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. *IEEE Access, 6*, 32328–32338.
15. Imran, A., Zoha, A., & Abu-Dayya, A. (2014). Challenges in 5G: How to empower SON with big data for enabling 5G. *IEEE Network, 28*(6), 27–33.
16. Sultan, K., & Ali, H. (2017). Where big data meets 5G? In *Proceedings of international conference internet things data cloud computing* (pp. 1–4).
17. Chen, J., Cheng, X., Du, R., Hu, L., & Wang, C. (2017). BotGuard: Lightweight real-time botnet detection in software defined networks. *Wuhan University, 22*(2), 103–113.
18. Celdrán, A. H., Pérez, M. G., Clemente, F. J. G., & Pérez, G. M. (2018). Towards the autonomous provision of self-protection capabilities in 5G networks. *Journal of Ambient Intelligence and Humanized Computing*, 1–14.
19. Machado, C. C., Granville, L. Z., & Schaeffer-Filho, A. (2016). ANSwer: Combining NFV and SDN features for network resilience strategies. In *Proceedings of IEEE symposium computers and communications* (pp. 391–396).
20. Pérez, M. G., et al. (2017). Dynamic reconfiguration in 5G Mobile networks to proactively detect and mitigate botnets. *IEEE Internet Computing, 21*(5), 28–36.
21. Ahmed, F., Deng, J., & Tirkkonen, O. (2016). Self-organizing networks for 5G: Directional cell search in mmW networks. In *Proceedings of IEEE international symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)* (pp. 1–5).
22. Arana, J. M., Han, J. P., & Cho, Y. S. (2016). Random-access technique for self-organization of 5G millimeter-wave cellular communications. *Mobile Information Systems, 2016*, 11.
23. Amiri, R., & Mehrpouyan, H. (2018). *Self-organizing mm wave networks: A power allocation scheme based on machine learning* (pp. 1–4).
24. Adeshina Busari, S., Mohammed Saidul Huq, K., Felfel, G., & Rodriguez, J. (2018). Adaptive resource allocation for energy-efficient millimeter-wave massive MIMO networks. In *Proceedings of IEEE GLOBECOM* (pp. 1–6).
25. Chen, N., Rong, B., Zhang, X., & Kadoch, M. (2017). Scalable and flexible massive MIMO precoding for 5G H-CRAN. *IEEE Wireless Communications, 24*(1), 46–52.
26. Mohammadkhan, A., Ghapani, S., Liu, G., Zhang, W., Ramakrishnan, K. K., & Wood, T. (2015). Virtual function placement and traffic steering in flexible and dynamic software defined networks. In *Proceedings of IEEE international workshop Local Metropolitan AreaNetwork* (pp. 1–6).
27. Tasiopoulos, A. G. et al. (2017). DRENCH: A semi-distributed resource management framework for NFV based service function chaining. In *Proceedings of IFIP networking conference (IFIP network) workshops* (pp. 1–9).
28. Nekovee, M., Qi, Y., & Wang, Y. (2017). Self-organized beam scheduling as an enabler for coexistence in 5G unlicensed bands. *IEICE Transactions on Communications, 100*(8), 1181–1189.
29. *Network Sharing; Architecture and Functional Description (Release 6)*, 3GPP Standard TS TS 36.104, 2009.
30. *Network Sharing; Architecture and Functional Description (Release 8)*, 3GPP Standard TS 23.251, 2010.
31. 3GPP TS 23.251, “Network sharing; Architecture and functional description (Release 11)”, 2011.
32. Marzouk, F., Barraca, J. P., & Radwan, A. On energy efficient resource allocation in shared RANs: Survey and qualitative analysis. *IEEE Communications Surveys & Tutorials, 22*(3), 1515–1538, thirdquarter 2020.
33. Elayoubi, S. E., Fallgren, M., Spapis, P., Zimmermann, G., Martín-Sacristán, D., Yang, C., Jeux, S., Agyapong, P., Campoy, L., Qi, Y., & Singh, S. (2016). 5G service requirements and operational use cases: Analysis and METIS II vision. In *European Conference on Networks and Communications (EuCNC)* (pp. 158–162).

34. Lashgari, M., Wosinska, L., & Monti, P. (2019). A shared-path shared-compute planning strategy for a resilient hybrid C-RAN. In *Proceedings of International Conference on Transparent Optical Networks (ICTON)* (pp. 1–6).
35. Ericsson AB, Huawei Technologies Co. Ltd, NEC Corporation and Nokia, "eCPRI Specification", May 2019, Version 2.0.
36. Yuan, C. I. Y., Huang, J., Ma, S., Cui, C., & Duan, R. (2015). Rethink fronthaul for soft RAN. *IEEE Communications Magazine, 53*(9), 82–88.
37. Lashgari, M., Natalino, C., Contreras, L. M., Wosinska, L., & Monti, P. (2019). Cost benefits of centralizing service processing in 5G network infrastructures. In *Proceedings of Asia Communications and Photonics Conference (ACP)* (pp. 1–3).
38. Simmons, J. M. (2014). *Chapter 7: Optical network design and planning* (2nd ed.). Springer.
39. Jaber, M., Owens, D., Imran, M. A., Tafazolli, R., & Tukmanov, A. (2016). A joint backhaul and RAN perspective on the benefits of centralised RAN functions. In *Proceedings of IEEE International Conference on Communications Workshops (ICC)* (pp. 226–231).
40. Khorsandi, B. M., Tonini, F., & Raffaelli, C. (2019). Centralized vs. distributed algorithms for resilient 5G access networks. *Photonic Network Communications, 37*(3), 376–387.
41. Shehata, M., Musumeci, F., & Tornatore, M. (2019). Resilient BBU placement in 5G C-RAN over optical aggregation networks. *Photonic Network Communications, 37*(3), 388–398.
42. Chaudhary, J. K., Zou, J., & Fettweis, G. Cost saving analysis in capacity-constrained C-RAN fronthaul. I*n Proceedings of 2018 IEEE Globecom Workshops (GC Wkshps)* (Vol. 8, pp. 1–7), Dec. 201.
43. NGMN Alliance: 5G extreme requirements: End-to-end considerations. *White Paper* Version 2.5, (2019).
44. 3GPP: Service requirements for the 5G system; Stage 1. TS22.261, Version 16.8.0, (2019).
45. Alliance, N. (2016, June). Perspectives on vertical industries and implications for 5G. *White Paper*.
46. Zhang, Q., Yang, C., Haas, H., & Thompson, J. S. (2014). Energy efficient downlink cooperative transmission with BS and antenna switching off. *IEEE Transactions on Wireless Communications, 13*(9), 5183–5195.
47. Li, J., Peng, M., Cheng, A., Yu, Y., & Wang, C. (2017). Resource allocation optimization for delay-sensitive traffic in fronthaul constrained cloud radio access networks. *IEEE Systems Journal, 11*(4), 2267–2278.
48. Nguyen, K.-G., Tervo, O., Vu, Q.-D., Tran, L.-N., & Juntti, M. (2018). Energy-efficient transmission strategies for CoMP downlink—Overview, extension, and numerical comparison. *EURASIP Journal on Wireless Communications and Networking, 2018*(1).
49. Chand, P., Mahapatra, R., & Prakash, R. (2013). Energy efficient coordinated multipoint transmission and reception techniques-a survey. *International Journal of Computer Networks and Wireless Communications, 3*(4), 370–379.
50. Li, J., Wu, J., Peng, M., & Zhang, P. (2016). Queue-aware energy-efficient joint remote radio head activation and beamforming in cloud radio access networks. *IEEE Transactions on Wireless Communications, 15*(6), 3880–3894.
51. Lee, Y. L., Wang, L.-C., Chuah, T. C., & Loo, J. (2016). Joint resource allocation and user association for heterogeneous cloud radio access networks. In *Proceedings of International Teletraffic Congress (ITC)* (pp. 87–93).
52. Soliman, H. M., & Leon-Garcia, A. (2016). QoS-aware Joint RRH activation and clustering in cloud-RANs. In *2016 IEEE wireless communications and networking conference* (pp. 1–6).
53. Luong, P., Tran, L., Despins, C., & Gagnon, F. (2016). Joint beamforming and remote radio head selection in limited fronthaul C-RAN. In *Proceedings IEEE VTC-Fall* (pp. 1–6).
54. Kim, T., & Chang, J. M. (2018). QoS-aware energy-efficient association and resource scheduling for HetNets. *IEEE Transactions on Vehicular Technology, 67*(1), 650–664.
55. Zhou, Z., Dong, M., Ota, K., Wang, G., & Yang, L. T. (2016). Energy-efficient resource allocation for D2D communications underlaying cloud-RAN-based LTE-A networks. *IEEE Internet of Things Journal, 3*(3), 428–438.

56. Peng, M., Zhang, K., Jiang, J., Wang, J., & Wang, W. (2015). Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. *IEEE Transactions on Vehicular Technology, 64*(11), 5275–5287.
57. Luo, S., Zhang, R., & Lim, T. J. (2015). Downlink and uplink energy minimization through user association and beamforming in C-RAN. *IEEE Transactions on Wireless Communications, 14*(1), 494–508.
58. Tran, G. K., Shimodaira, H., Rezagah, R. E., Sakaguchi, K., & Araki, K. (2015). Dynamic cell activation and user association for green 5G heterogeneous cellular networks. In *Proceedings of IEEE PIMRC* (pp. 2364–2368).
59. Wang, K., Yang, K., & Magurawalage, C. S. (2018). Joint energy minimization and resource allocation in C-RAN with mobile cloud. *IEEE Transactions on Cloud Computing, 6*(3), 760–770.
60. Yu, Z., Wang, K., Ji, H., Li, X., & Zhang, H. (2016). Joint user association and downlink beamforming for green Cloud-RANs with limited fronthaul. In *Proceedings of IEEE GLOBECOM* (pp. 1–6).
61. Fan, B., Tian, H., & Yan, X. (2016). Resource allocation in a generalized LTE air interface virtualization framework exploiting user behavior. *EURASIP Journal on Wireless Communications and Networking, 2016*(1), 107.
62. Peng, M., Yu, Y., Xiang, H., & Poor, H. V. (2016). Energy-efficient resource allocation optimization for multimedia heterogeneous cloud radio access networks. *IEEE Transactions on Multimedia, 18*(5), 879–892.
63. Wang, S., & Sun, Y. (2017). Enhancing performance of heterogeneous cloud radio access networks with efficient user association. In *Proceedings of IEEE ICC* (pp. 1–6).
64. Wang, B., Yang, Q., Yang, L. T., & Zhu, C. (2017). On minimizing energy consumption cost in green heterogeneous wireless networks. *Computer Networks, 129*, 522–535.
65. Zeng, D., Zhang, J., Guo, S., Gu, L., & Wang, K. (2017). Take renewable energy into CRAN toward green wireless access networks. *IEEE Network, 31*(4), 62–68.
66. Pan, C., Zhu, H., Gomes, N. J., & Wang, J. (2017). Joint precoding and RRH selection for user-centric green MIMO C-RAN. *IEEE Transactions on Wireless Communications, 16*(5), 2891–2906.
67. Opadere, J., Liu, Q., & Han, T. (2017). Energy-efficient RRH sleep mode for virtual radio access networks. In *Proceedings of IEEE Globecom* (pp. 1–6).
68. Chughtai, N. A., Ali, M., Qaisar, S., Imran, M., Naeem, M., & Qamar, F. (2018). Energy efficient resource allocation for energy harvesting aided H-CRAN. *IEEE Access, 6*, 43990–44001.
69. Sigwele, T., Alam, A. S., Pillai, P., & Hu, Y. F. (2015). Evaluating energy-efficient cloud radio access networks for 5G. In *Proceedings of IEEE conference data science data intensive systems* (pp. 362–367).
70. Guo, H., Wang, K., Ji, H., & Leung, V. C. M. (2016). Energy saving in C-RAN based on BBU switching scheme. In *Proceedings of IEEE International Conference Network Infrastructure Digit Content (IC-NIDC)* (pp. 44–49).
71. Qian, M., Hardjawana, W., Shi, J., & Vucetic, B. (2015). Baseband processing units virtualization for cloud radio access networks. *IEEE Wireless Communications Letters, 4*(2), 189–192.
72. Kong, Z., Gong, J., Xu, C., Wang, K., & Rao, J. (2013). eBase: A baseband unit cluster testbed to improve energy-efficiency for cloud radio access network. In *Proceedings of IEEE ICC* (pp. 4222–4227).
73. Aldaeabool, S. R., & Abbod, M. F. (2017). Reducing power consumption by dynamic BBUs-RRHs allocation in C-RAN. In *Proceedings of Telecommunication Forum (TELFOR)* (pp. 1–4).
74. Al-Zubaedi, W., & Al-Raweshidy, H. S. (2017). A parameterized and optimized BBU pool virtualization power model for C-RAN architecture. *In Proceedings of IEEE EUROCON* (pp. 38–43).

75. Boulos, K., El Helou, M., & Lahoud, S. (2015). RRH clustering in cloud radio access networks. In *Proceedings of International Conference Application. research. Computer science engineering. (ICAR)* (pp. 1–5).
76. Boulos, K., Helou, M. E., Ibrahim, M., Khawam, K., Sawaya, H., & Martin, S. (2017). Interference-aware clustering in cloud radio access networks. In *Proceedings of IEEE International Conference Cloud Network (CloudNet)* (pp. 83–88).
77. Ferdouse, L., Ejaz, W., Anpalagan, A., & Khattak, A. M. (2017). Joint Workload Scheduling and BBU Allocation in cloud-RAN for 5G Networks. In *Proceedings of symposium applied computing* (pp. 621–627).
78. Alhumaima, R. S., Ahmed, R. K., & Al-Raweshidy, H. S. (2018). Maximizing the energy efficiency of virtualized C-RAN via optimizing the number of virtual machines. *IEEE Transactions on Green Communications and Networking, 2*(4), 992–1001.
79. Namba, S., Warabino, T., & Kaneko, S. (2012). BBU-RRH switching schemes for centralized RAN. In *Proceedings of international conference communications network China* (pp. 762–766).
80. Khan, M., Alhumaima, R. S., & Al-Raweshidy, H. S. (2015). Reducing energy consumption by dynamic resource allocation in C-RAN. In *Proceedings of the European conference on Networks Communications (EuCNC)* (pp. 169–174).
81. Lyazidi, M. Y., Giupponi, L., Mangues-Bafalluy, J., Aitsaadi, N., & Langar, R. (2017). A novel optimization framework for C-RAN BBU selection based on resiliency and price. In *Proceedings of VTC-Fall* (pp. 1–6).
82. Liu, J., & Falowo, O. E. (2018). Traffic-aware heuristic BBU-RRH switching scheme to enhance QoS and reduce complexity. In *Proceedings of IEEE PIMRC* (pp. 1–7).
83. Guo, S., Zeng, D., Gu, L., & Luo, J. (2018). When green energy meets cloud radio access network: Joint optimization towards Brown energy minimization. *Mobile Networks and Applications*, 1–9.
84. Tang, J., Tay, W. P., & Quek, T. Q. S. (2015). Cross-layer resource allocation with elastic service scaling in cloud radio access network. *IEEE Transactions on Wireless Communications, 14*(9), 5068–5081.
85. Wang, K., Zhao, M., & Zhou, W. (2014). Traffic-aware graph-based dynamic frequency reuse for heterogeneous Cloud-RAN. In *Proceedings of IEEE Globecom* (pp. 2308–2313).
86. Lyazidi, M. Y., Aitsaadi, N., & Langar, R. (2016). Dynamic resource allocation for Cloud-RAN in LTE with real-time BBU/RRH assignment. In *Proceedings of IEEE ICC* (pp. 1–6).
87. Al-Dulaimi, A., Al-Rubaye, S., & Ni, Q. (2018). Energy efficiency using cloud management of LTE networks employing fronthaul and virtualized baseband processing pool. *IEEE Transactions on Cloud Computing*, 1–1.
88. Wang, K., Zhou, W., & Mao, S. (2016). Energy efficient joint resource scheduling for delay-aware traffic in Cloud-RAN. In *Proceedings of IEEE GLOBECOM* (pp. 1–6).
89. Vincenzi, M., Antonopoulos, A., Kartsakli, E., Vardakas, J., Alonso, L., & Verikoukis, C. (2017). Cooperation incentives for multi-operator C-RAN energy efficient sharing. In *Proceedings of IEEE ICC* (pp. 1–6).
90. Wang, K., Zhou, W., & Mao, S. (2017). On joint BBU/RRH resource allocation in heterogeneous cloud-RANs. *IEEE Internet of Things Journal, 4*(3), 749–759.
91. Iardella, N. et al. (2017). Flexible dynamic coordinated scheduling in virtual-RAN deployments. In *Proceedings of IEEE ICC Workshops* (pp. 126–131).
92. Yao, J., & Ansari, N. (2018). QoS-aware joint BBU-RRH mapping and user association in cloud-RANs. *IEEE Transactions on Green Communications and Networking, 2*(4), 881–889.
93. Liu, Q., Han, T., & Ansari, N. (2018). Energy-efficient on-demand cloud radio access networks virtualization. In *Proceedings of IEEE Globecom* (pp. 1–6).
94. Amani, N., Pedram, H., Taheri, H., & Parsaeefard, S. (2019). Energy-efficient resource allocation in heterogeneous cloud radio access networks via BBU offloading. *IEEE Transactions on Vehicular Technology, 68*(2), 1365–1377.

95. Rodriguez, J. et al. (2017). SECRET — Secure network coding for reduced energy next generation mobile small cells: A European Training Network in wireless communications and networking for 5G," A European Training Network in wireless communications and networking for 5G. In *Proceedings of Internet Technology Applications (ITA)* (pp. 329–333).
96. Marzouk, F., Akhtar, T., Politis, I., Barraca, J. P., & Radwan, A. (2020). Power minimizing BBU-RRH group based mapping in C-RAN with constrained devices. In *Proceedings of IEEE ICC* (pp. 1–7).

Chapter 11
Advanced Cloud-Based Network Management for 5G C-RAN

Massimiliano Maule, Ojaghi Kohjogh, and Farhad Rezazadeh

Abstract Cloud technology offers new innovative alternatives for radio access network (RAN) deployments complementing existing proven purpose-built solutions, able to fulfill the diversity of requirements of 5G services in a cost-effective manner. Using new emerging technologies, such as (i) software-defined networking (SDN), (ii) network function virtualization (NFV), and (iii) network slicing, cloud RAN functionalities are applied over a general-purpose platform, where the baseband units (BBU) are migrated to the cloud for a centralized processing and management. In this chapter, we propose three advanced approaches for the orchestration and monitoring of cloud RAN resources, exploiting the latest optimization techniques applied to 5G RAN networks.

11.1 Introduction

The radio access network (RAN) has seen an incremental evolution over the years and will require a major step forward as we enter the new age of mobile telecommunications. With 5G, research entities have suggested that the RAN network architecture needs to be restructured beyond the evolution of the 3GPP LTE releases. The novel RAN structure requires a new approach in light of the new use cases, services, and traffic types that 5G introduces, paying special attention to configurability and flexibility.

M. Maule (✉)
Iquadrat Informatica S.L., Barcelona, Spain
e-mail: mmaule@iquadrat.com

O. Kohjogh (✉)
Wireless Networks Research Lab (WINE), Universitat Oberta de Catalunya (UOC), Castelldefels, Spain
e-mail: bojaghi@uoc.edu

F. Rezazadeh
Centre Tecnològic Telecomunicacions Catalunya, Castelldefels, Spain
e-mail: farhad.rezazadeh@cttc.cat

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_11

Cloud radio access network (C-RAN) is introduced as an innovative new architecture that tries to meet such needs by centralizing the base stations and providing a cooperative solution between multiple operators. The radio functionalities are decoupled into BBU and remote radio head (RRH) and then centralizing the BBUs from multiple sites into a single geographical point such as a cloud data center, using cloud computing and virtualization techniques. Such technology comes with minimal cost, high energy efficiency, and centralized network architecture that attracted a lot of attention in both academia and industry.

Although C-RAN appears to be a promising access architecture, the efficient management of the resources in C-RAN to satisfy traffic demand is a significant challenge due to the user mobility and dynamic environment. From this consideration, this chapter illustrates different resource management approaches able to maximize the service quality and radio resources utilization efficiency through the application of network slicing enablers over C-RAN.

Three types of applications are illustrated in this chapter. In Sect. 11.2, the author presents a joint slice-based C-RAN solution by exploring different functional splits between central and distributed units. Using these two 5G enablers, the proposed framework enhances the fronthaul (FH) network infrastructure scalability, providing superior QoS (quality of service). In Sect. 11.3, a novel real-time RAN network slicing approach is presented. Given the maximum capacity of the radio interface, the author presents a novel radio resources management algorithm where the dimension of each slice is dynamically defined through the joint evaluation of the tenant's Service Level Agreement (SLA) and the real-time traffic performance. The framework is tested on 5G research testbed using real radio and user equipment (UE). To conclude, as automation and AI technologies improve, Sect. 11.4 presents an AI application to zero-touch networks on top of C-RAN architecture. The overall aim of zero-touch networks is for machines to learn how to become more autonomous so that we can delegate complex, mundane tasks to them. This subsection illustrates the major role of AI to assist operators to automate RAN operations, boost network performance, improve resource management, and shorten the time to market for new features.

11.2 SlicedRAN: Service-Aware Network Slicing Framework for 5G RAN

11.2.1 5G Cloud-RAN Architecture

Introduced by China Mobile [1], the C-RAN architecture has recently gained momentum technologies in the 5G era. As proposed by 3GPP in [2], the C-RAN architecture decomposed into three main parts:

- **Central units (CUs)** – composed of higher flexible and programmable processors

- **Remote radio heads (RRHs)** – located at the remote site and controlled by CU
- **Fronthaul network (FH)** – low-latency high-bandwidth optical or wireless network, connecting CU and RRHs

The C-RAN is the paramount candidate to meet the International Telecommunications Union's (ITU) IMT-Advanced 5G mobile network services requirements [3], which include the following:

(i) Enhanced mobile broadband (eMBB) is the service that needs higher bandwidth, such as high-definition (HD) videos, virtual reality (VR), and augmented reality (AR).
(ii) Ultra-reliable and low-latency communications (uRLLC) is characterized as services demanding low-latency and more reliable mobile services, such as industrial Internet, remote surgery, and assisted and automated driving.
(iii) Massive machine-type communications (mMTC) is designed for services that include high requirements for connection density, such as smart city and smart agriculture.

In order to be able to serve this traffic, virtualization emerges as an essential component at the network edge, namely, the virtual partitioning of the mobile RAN. Consequently, it is of great importance to investigate all the new enabling architectural elements of these networks. 5G New Radio (5G NR) is the evolution to LTE Advanced and LTE Advanced Pro wireless technologies which is defined in 3GPP Release 15 and beyond. In the novel architecture, base station (BS), known as eNB within the LTE network architecture, is fully shifted from the RRHs into a centralized CU (equivalently BBU) [4]. CUs are connected to the Evolved Packet Core (EPC) through a Backhaul (BH) network, and all RRHs are connected to CUs through a FH network, typically transmitting radio signals using current specifications of serial line interfaces like CPRI (Common Public Radio Interface) or OBSAI (Open Base Station Architecture Initiative) [4]. This segregation allows the possibility of sharing base-band processing between numerous RRHs, thus enabling higher utilization of resources, better coordination, and reduced deployment costs. Moreover, the introduction of virtualizing techniques into the RAN architecture qualifies virtualization of BS functions, which will be further discussed. This architecture is capable of improving spectrum efficiency and energy efficiency and reducing deployment costs (due to pooling gains) [5]. Nevertheless, its advantages do not guarantee for realistic large-scale deployments due to the stringent requirements on the FH for 5G. For example, in C-RAN where all BS functionality is centralized except the RF function which is located at the RRH, thus transmitting IQ samples through the FH will require bandwidth of up to 2.5 Gbps and a very low delay of 0.25 ms.

Figure 11.1 shows the structure of 5G NR which is connected to the core network (CN) and Internet. In this figure, there exists three RRHs with Virtualized Network Functions (VNF). All RRHs are connected to a CU via the transport network (FH/BH), and the CU is connected to the CN via the BH link. Finally, the core is connected to the Internet. This architecture helps mobile network operators to support the various demands in 5G networks due to the flexibility of deploying Network Functions (NFs) and enabling the customized functional split per use cases which brings advantages in both network systems and user experience part.

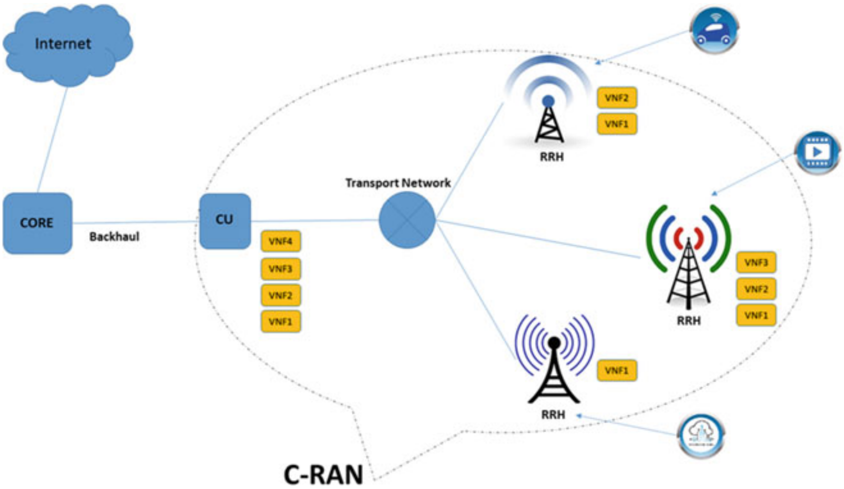


Fig. 11.1 The schematic overview of 5G NR with CN and Internet

11.2.1.1 Functional Splits

5G networks are expected to support various applications with a high flexibility meeting diversity of requirements in terms of latency, data rates, and massive connectivity. A 5G NF supplies a particular capability to support communication through a 5G network. NFs are normally virtualized, but some functions may need an addition of more specialized hardware. NFs can be the functions that are common functions which are essential for all applications, for example, authentication and identity management NFs. On the other hand, there exist some other functions which might not be useful for all the use cases. For instance, a mobility management function such as handover can only be used for the eMBB applications [6], or an uRLLC user needs higher decentralized functions to reduce the hybrid automatic repeat request (HARQ) delay and guarantee low delay.

The aim of the aforementioned C-RAN architecture is to centralize all functions into a BBU pool (i.e., CU) [2, 7]. The functionalities that can be decoupled comprise channel encoding and error correction decoding, modulation/demodulation, resource mapping/demapping, channel estimation and equalization, fast Fourier transform (FFT) and its inverse, analog-to-digital and digital-to-analog conversion, and antenna radio transmission and reception. To reduce the FH traffic amount, some modules can be migrated to the RRH side and other functions shifted to central unit (CU). However, the functionality at RRHs can be just as basic signal and analog processing known as Distributed RAN (D-RAN) [7].

The main advantages of C-RAN architecture are:

- Simplifies the structure of RRHs.
- Reduces operation and deployment costs.

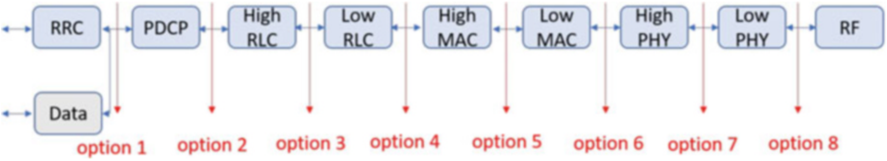


Fig. 11.2 Functional split options [2]

- Enables virtualization and slicing of RAN.
- Allows the flexible allocation of a pool of radio and computational resources.
- Controls the transitions from distributed BSs to a centralized RAN.
- Migrates a hardware-defined infrastructure to a software-defined environment.
- Maximizes spectrum efficiency and hardware usage.

As illustrated in Fig. 11.2, different functional splits options between CU and RRHs are introduced by 3GPP. Among them, four options are widely deployed:

- **PHY-layer split (option 6**): This physical (PHY) split known as C-RAN achieves the highest centralization and coordination which enables a more efficient resource management and can be realized only with an ideal FH which consumes very high bandwidth and has very low delay bounds.
- **MAC-layer split (option 4):** The medium access control (MAC) layer and the layers above it are pooled within the CU with centralized scheduling (as MAC is in CU) for several RRHs. This split allows synchronized multi-cell coordination for coordinated multi-point (CoMP) and enhanced inter-cell interference coordination (eICIC) but requires a low-latency FH and has significant traffic overheads.
- **RLC-layer option (option 3):** The Radio Link Control (RLC) layer and other layers above it are virtualized at the BBU. The failure over transport network may also be recovered using the end-to-end Automatic repeat request mechanism at CU. This may provide protection for critical data.
- **PDCP-layer option (option 2):** This option runs the Packet Data Convergence Protocol (PDCP) functions at the BBU and may use any type of FH network. The main advantage of this option is the possibility to have an aggregation of different RRH technologies (e.g., 5G, LTE, and WiFi).

11.2.1.2 Integrated FH/BH Network

As previously explained, there are multiple-layer split options for C-RAN, and accordingly ongoing discussions on the lower layer split for the 3GPP 5G networks, which states a clear inquiry for protocols to support such integration of FH/BH known as crosshaul network [8, 9]. This data will be packetized in the coexistence of FH/BH traffic within the same transport network. This requirement has already been recognized and accepted by the industry that is investigating new protocols such as

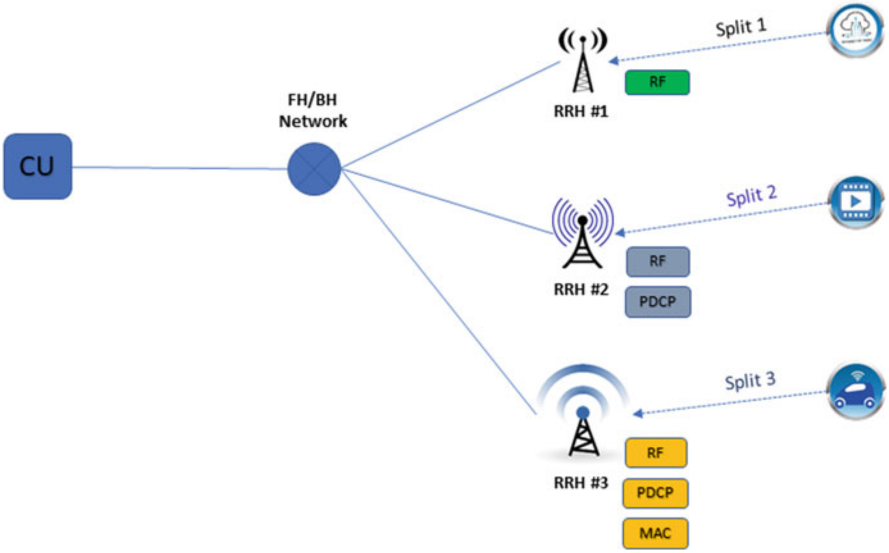


Fig. 11.3 C-RAN with FH/BH Network supporting different layers of split

the Enhanced Common Public Radio Interface (eCPRI [10]) and Next Generation Fronthaul Interface (NGFI) [11]. The integrated FH/BH in 5G will act as a transport network providing another degree of freedom for load balancing and enabling a flexible and software-defined reconfiguration of all networking elements in a multi-tenant and service-oriented unified management environment. Furthermore, for cost reasons, the heterogeneity of transport network equipment must be tackled by unifying data, control, and management plane across all technologies as much as possible. The coexistence of FH/BH in 5G network envisioned will consist of high-capacity switches and heterogeneous transmission links (e.g., fiber or wireless optics, high-capacity copper, mmWave) interconnecting RRHs, thus supporting different functional split (see Fig. 11.3).

11.2.2 Virtualization Enablers

In this section, two promising enablers of virtualization, namely, SDN and NFV, for 5G mobile networks are discussed.

11.2.2.1 Software-Defined Networking (SDN)

The emergence of SDN [12] as one of the key technologies provides a basis for introducing a uniform QoS networking approach in the context of evolving mobile networks domain. The main idea of SDN is the separation of the control and the data plane through a well-defined API (e.g., OpenFlow). In this approach, a software-based controller has an overview of the whole network and is responsible for the decision-making, while the hardware is simply in charge of forwarding packets to destination following packet-handling rules. The flexibility introduced with SDN permits to address different challenges of the current and future mobile networks.

The RAN as a complex and costly part of mobile networks infrastructure will benefit more from the SDN concept. Since the RAN part is composed by multiple Radio Access Technologies (RATs) (i.e., LTE, WiFi), SDN-based solution is deployed to coordinate and synchronize the BSs. Furthermore, softwarization in the RAN through programmability enables flexibility and authorizes a broad range of applications and novel technologies such as virtualization [13] and slicing [14] which will be further discussed.

11.2.2.2 Network Function Virtualization (NFV)

The NFV technology is a carrier-driven initiative with the goal to adapt the way that operators design networks by using virtualization technologies to virtualize NFs. NFV is responsible for forwarding NFs as a software, capable of running as virtualized rules, and allowing to be deployed at required locations in the network without needing to install an equipment for each new rule. NFV is applicable to any NF in both mobile and fixed networks. NFV technology can be used to implement RRH upgrades for lower splits in software [15]. This solution can be an enabler for a network with flexible functional splits, where the NFs are adapted according to a certain set of requirements and enabled when required by the NFV. In [16], the FH is known as the main part of the SDN-based mobile network architecture which shows the significant role of SDN/NFV in this part of the mobile network.

11.2.3 SlicedRAN: Design and Implementation

This subsection presents the work that has been done so far regarding RAN slicing along with functional split while considering different slices [17]. RAN slicing allows a customized functional split deployment per slice and optimization of the available resources such as network capacity. We analyze joint slicing and functional split optimization in the future RAN approach. The protocol stack in an eNB consists of several layers, each one responsible for a specific function or a set of functions. Indeed, the whole operation of the eNB can be modelled as a chain of these functions [18]. In this context, without precluding any of the granularity levels proposed by 3GPP [2], in our work, we focus on the slice-based functional split, assuming one slice for each service. As discussed in [2], the network layers PDCP, RLC (high and low sublayers), MAC (high and low sublayers), and PHY (high and low sublayers) can be allocated either in the CU or in the RRH. Accordingly, each

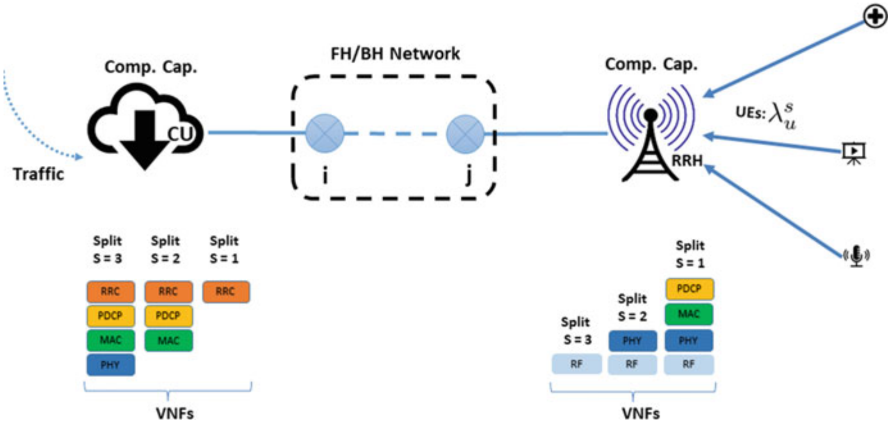


Fig. 11.4 Scheme of SlicedRAN with virtualized functional splits

Table 11.1 Functions' allocation and FH/BH bandwidth requirements for a traffic denoted by λ^{u}_{s}, for UE u and service s, with 20 MHz bandwidth; Downlink: MCS (modulation and coding scheme) index 28, 2×2 MIMO replicated from [20]

Split	Bandwidth (b/s)	Function at CU	Functions at RRH
1	λ_u^s	f_3	f_0, f_1, f_2
2	$1.02\lambda_u^s + 1.5 \bullet 10^6$	$f_2 \bullet f_3$	f_0, f_1
3	$2.5 \bullet 10^9$	$f_1 \bullet f_2 \bullet f_3$	f_0

functional split will be defined by the set of functions allocated in the CU and the set of functions allocated in the RRH.

As described in [2] and implemented in [19], functional isolation is assumed at the eNB. This means that each slice of an eNB can have a different functional split. As shown schematically in Fig. 11.4, RAN slicing allows a customized functional split deployment per slice, and optimization of the available resources, e.g., transport network capacity and RRH or CU computational capacity. Figure 11.4 conveys how different control functions are allocated either at the RRH or at the CU for each slice.

In the following, we will assume a set of four NFs, denoted as F = {f_0, f_1, f_2, f_3}, where f_0 is the low-layer NF (RF, signal and analog processing, etc.), which is always placed in the RRH; f_1 serves all PHY functions except for function f_0; f_2 corresponds to RLC and MAC; and f_3 is the high-layer NF (e.g., PDCP and above layers). Depending on the functional split, these functions will be allocated either at the RRH or at the CU, thus defining the FH/BH bandwidth requirements between the CU and RRHs. Table 11.1 includes a summary of the allocation of functions and the associated FH/BH bandwidth requirements for each split [20].

In principle, regardless of the adopted functional split, f_0 is always placed in RRH, and f_3 is in CU, thus generating three different functional split options, namely, split 1, split 2, and split 3. Split 1 is a completely decentralized functional split that accommodates all functions except f_3 at the RRH. That is, all layers below PDCP run in the RRH. Given the allocation of functions, this split does not have traffic overhead, and the required FH/BH capacity can be approximated by the aggregate UEs' traffic. In split 2, f_2 is moved from the RRH to the CU, thus leaving only f_0 and f_1 in the RRH. This allows a higher degree of coordination among eNBs sharing the same CU, enabling a better utilization of resources with techniques such as CoMP, frame alignment, and centralized HARQ. However, split 2 allocation imposes higher traffic overhead than split 1. Finally, in split 3, only the RF function is located at the RRH, while the rest of functions are moved to CU (complete centralization), transmitting IQ samples through the FH/BH. In this case, samples are usually encapsulated with Common Public Radio Interface (CPRI) [21], and the required FH capacity depends on the bandwidth allocated to the eNB, the number of antennas, etc. That is, FH capacity requirement does not depend on the UEs' traffic for split 3. The main advantage of split 3 is that the centralization achieves the highest coordination degree among eNBs.

Given the described scenario, the network creates different slices on top of the physical RAN to serve the traffic. Let us assume a UE u with service s connected to RRH r. The slice tailored to serve this traffic with the appropriate QoS consists of a specific functional split and the path between the CU and the RRH r, always taking into account the network constraints, such as the available PRBs in the RRH, the computational capacity in both the RRH and the CU, and the capacity of the links of the FH/BH network.

For the sake of simplicity, we present the analysis of a small network with 3 RRHs (i.e., $R = 3$) and 12 UEs (i.e., $U = 12$) (cf. Figs. 11.5 and 11.6), to clarify the advantages of leveraging virtualization in SlicedRAN. In our simulation setup, we assume a bandwidth of 20 MHz for each BS (i.e., 100 PRBs) and a link capacity ranging from 100 Mb/s to 25 Gb/s.

As can be seen in Fig. 11.5, state-of-the-art (SoA) fails to support the QoS of all UEs due to the restriction in the capacity of RRHs (here RRH-3) and also the limitation in using only a single functional split in the RRHs.

Conversely, in Fig. 11.6, SlicedRAN serves the QoS of those UEs who were not satisfied in SoA since in SlicedRAN RRHs are virtualized, which enables them to use different functional splits in order to meet the QoS of different UEs. For example, RRH-1 serves three different UEs by supporting three-different functional splits as illustrated in Fig. 11.6.

We then study the throughput performance of SlicedRAN comparing with SoA, where each RRH is allowed to use only a single split of 1, 2, or 3. As it can be observed in Fig. 11.7, SlicedRAN achieves more throughput due to serving more UEs with different types of services. This effect is pronounced for the utilization of slicing (i.e., supporting different functional splits per RRH) in SlicedRAN, which

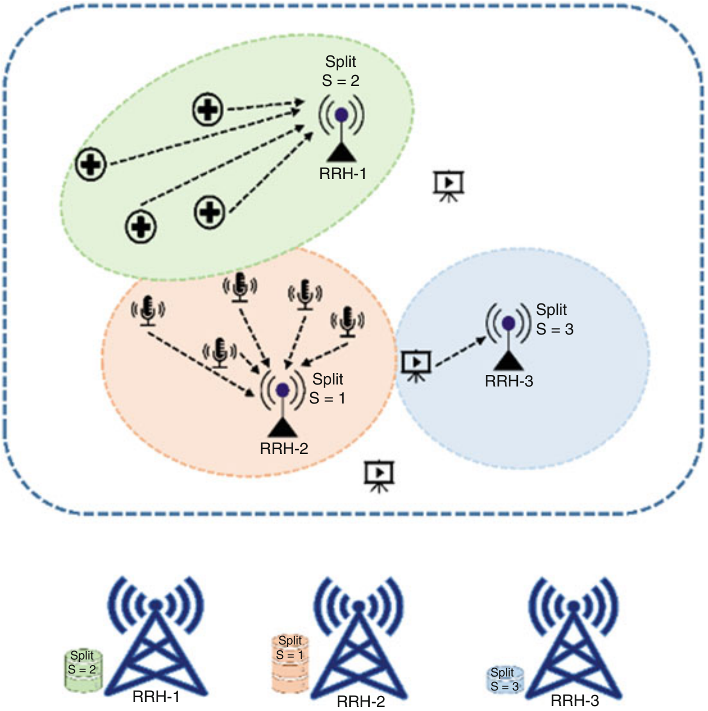


Fig. 11.5 Scheme of UE association in SoA, where only a single functional split is allowed per eNB

better uses the resources and serves the different QoS requirements of UEs. In contrast, each of split 1, 2, and 3 is configured for the RRHs. In this context, the UEs that cannot be served with the functional split deployed in the RRH will have to be dropped.

To conclude, RAN slicing and functional split are two major concepts of future 5G networks. In this work, we proposed *SlicedRAN*, which is a joint slicing and functional split optimization framework for 5G not yet fully investigated. In this work, we elucidate how to solve the problem of providing isolated and tailored slices for different services with customized functional splits per slice when CU and RRHs are connected through a FH/BH network. Whereas existing proposals assume a single functional split per RRH [20], SlicedRAN leverages virtualization to create multiple slices in RRHs, each one with the most appropriate functional split to meet the requirements of the slice. Slices are created based on the QoS of traffic demand and the set of RAN constraints, such as RRHs' computational capacity, capacity of the FH/BH network, spectrum availability per RRH, etc. Not all services support the whole range of possible functional splits. For instance, high transmission rates usually require a high degree of centralization to implement efficient CoMP. We take this into account in the functional split optimization.

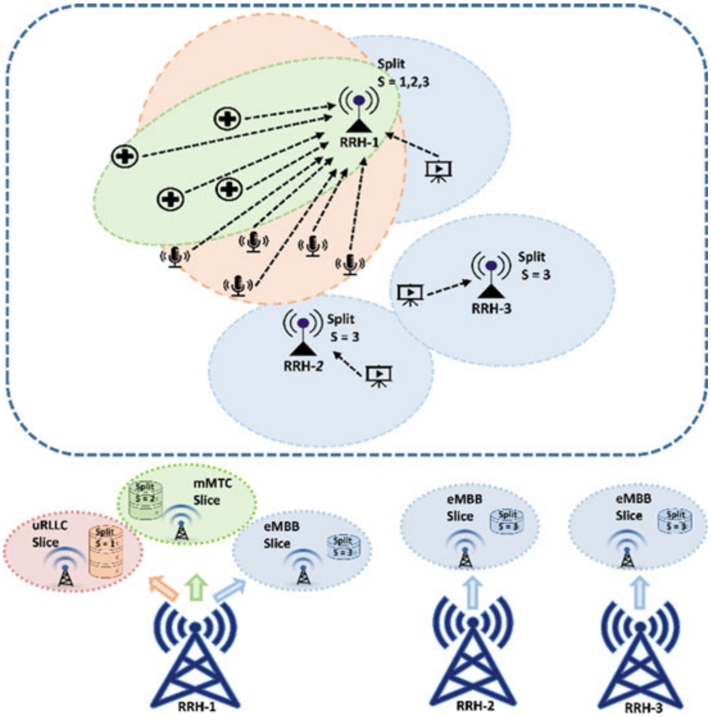
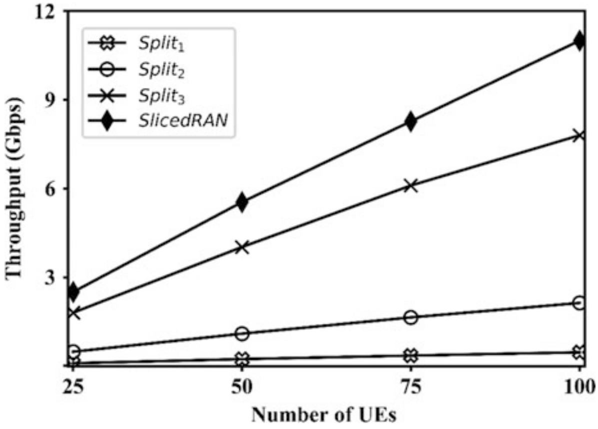


Fig. 11.6 Scheme of UE association in SlicedRAN, thanks to slicing, which allows a multiple virtual functional split to be placed on each eNB

Fig. 11.7 Throughput analysis as a function of the number of UEs



11.3 SLA-Based Dynamic Network Slicing

11.3.1 Network Slicing Properties

11.3.1.1 Slice Isolation

The concept of isolation is not new, and its level of integration changes over time following the network infrastructure evolution. With 5G, isolation is embedded with the network itself, with a flexible management by the operator through network slicing methods.

Depending on the domain (RAN, transport, and core) capabilities, the instantiation of a slice demands specific isolation properties. The RAN slice isolation solution aims to isolate and guarantee resources for network slices on new radio (NR) RAN. Slice-specific QoS assurance, air interface dynamic RB (Resource Block) resource sharing, and static RB resource reservation are available, differentiating in service latency, reliability, and isolation requirements [22].

The scalability and flexibility principles of 5G are fundamental for the establishment of end-to-end network slicing solutions on top of different scenarios using a shared physical network infrastructure. For public and industrial networks built on 5G shared network infrastructure, the RAN slicing isolation focuses on RB resource sharing and slice-specific QoS, with basic security functionalities. On the other side, special RAN slicing isolation solutions are built on 5G industry private network infrastructure with tailored requirements, where the resources are completely independent, customizable, and with high-level-security functionalities.

In the RAN domain, the type of slicing technique defined for a specific scenario (public or private) has an impact in the air interface. In particular, when the scenario requires specific customization, static or dynamic RB resource-sharing solutions can be applied. For a static approach, high requirements are posed on service isolation and bandwidth. Examples of industries using this method are power grid systems, government, and public security private networks. On the other side, resources are dynamically shared when basic service assurance is required, like for smart grid inspection and media live broadcast industries.

11.3.1.2 Resources Prioritization

The task relevant to radio resource allocation becomes more challenging with network slicing, as it introduces a two-tier priority in the system. The first tier refers to the priority of different slices, i.e., inter-slice priority, as each slice has its own priority defined according to the agreements between the slice owner and the network provider. The second tier refers to the priority among the users of the same slice, i.e., intra-slice priority [23].

The priority of slices is determined by various ways in terms of the traffic flow type such as multimedia traffic and elastic traffic, SLAs between a service provider and a client, and the amount of traffic volume of the slices. For example, if the traffic volume of a specific slice is higher than others, the highest priority can be given to the slice for enhancing the QoS of the slice. With this solution, as illustrated in [24], the slice prioritization approach can be combined with routing forwarding schemes to reduce the interference imposed on each slice through the application of different routing policies to each flow.

11.3.1.3 Service Level Agreement (SLA)

The SLA is an official agreement between service provider and tenant or between service providers, based on which the level of rendered service is precisely defined. SLAs contain QoS properties that must be maintained by a provider. These are generally defined as a set of SLOs.

In slice-based 5G networks, every slice needs an individual SLA, which would have unique elements, metrics and structure in comparison to the SLAs of other slices within same network. These metrics are used to describe the level and volume of communication services and to measure the performance characteristics of the service objects [25]. The operator provides and maintains services to the tenant through one or more multiple slices, which is acknowledged by the tenant. A set of QoS metrics of the slices service, such as security, power, throughput, latency, etc., are real-time monitored to verify the compliance with the terms among the parties.

In case of SLA violation, both the service provider and tenant predefine an appropriate penalty value in the SLA. Depending on the importance of the violated SLO and/or the consequences of the violation, the provider in breach may avoid dispatch or obtain a diminished monetary sanction from the client [26]. As both the service provider and the client are ultimately businesses, they are free to decide what kind of sanctions they will associate to the various types of SLA breaches, in accordance with the importance of the SLO that was not fulfilled.

11.3.2 Real-Time 5G Radio Access Network Slicing Management

While significant effort has already been achieved at 3GPP specifications level regarding the network slicing architecture in 5G networks, management solutions for the exploitation of this feature in the next generation radio access network (NG-RAN) still present multiple open challenges.

Figure 11.8 illustrates a general framework for network slicing with focus on the RAN, based on a layered architectural approach, and it is well aligned with most proposals from the literature [27, 28].

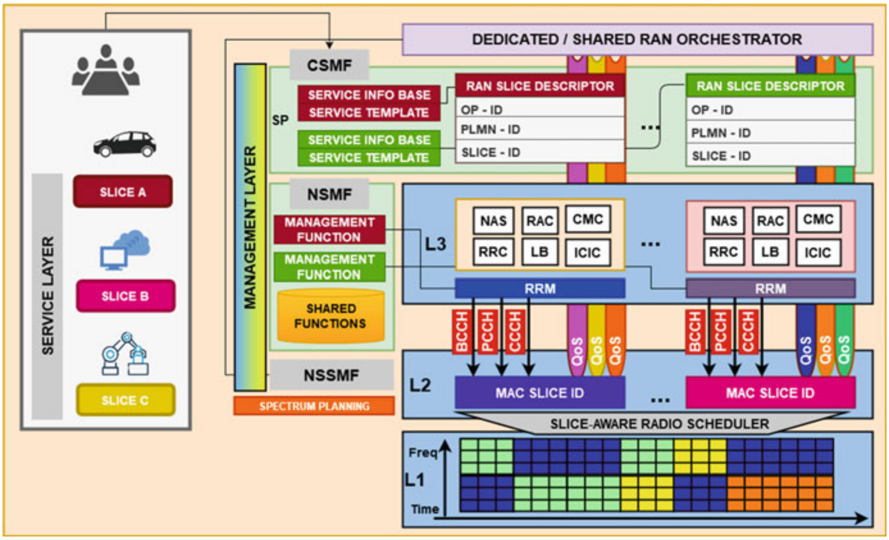


Fig. 11.8 Architectural management framework for 5G-NR network slicing

As initial step, the service layer (SL) handles the negotiation between the slice tenant and the service provider (SP) of the SLA requirements and key performance indicators (KPIs) to be monitored.

The description of the service requirements is done using a slice template, which may dynamically update to maintain and optimize the performance. The updated service template is then forwarded to the management layer, where the communication service management function (CSMF) translates the service requirements in network slice requirements. These requirements are acquired by the network slice management function (NSMF) for the definition of the E2E management and orchestration of the network slice instances (NSIs). In parallel to the NSMF, the network slice subnet management function (NSSMF) sorts the NSIs according to the specific sub-domain, e.g., RAN-NSSMF and CORE-NSSMF. The RAN NSSMF may directly communicate with the RAN orchestrator for the mapping of each slice requirement across the radio stack layers. Moreover, it supports a wide range of functionalities and performance/fault management, such as spectrum planning, admission control, SLA conformance monitoring, and traffic forecasting. For the management of the resources prioritization and isolation policies, the RAN-NSSMF enables the spectrum planning block, which manages the long-term allocation of spectrum chunks for each slice given its capacity requirement and the desired level of resource isolation.

To conclude, given the sharing capabilities of 5G-NR, the RAN resources are instantiated in strategic Points of Presence, e.g., the cell sites for SCs and the central offices for edge data centers.

11.3.3 Static and Dynamic Network Slicing Approaches

The major element underlying network slicing is the mechanism deployed for resource allocation among slices. One of the earlier approaches considered in 3GPP suggests the base station resources are statically partitioned among the slices based on fixed network shares [29]. Multiple providers share the same infrastructure, while the resources are allocated according to QoS objectives. With this method, resource overprovisioning is employed to contrast SLA violation, introducing as side effect the reduction of the system performance due to the possible allocation of unusable resources. Moreover, the stochastic behavior of the medium introduces complexity when it comes to allocating the resources of a new slice instance.

While traditionally such network resource allocation problems have been solved using analytical queueing theoretic and optimization methods, innovative system models for the rapid instauration and definition of network slice instances and dynamic allocation of the resources must be investigated. With the introduction of dynamic network slicing, multiple tenants adjust their network capacities during different time periods of the day or the week. Machine learning (ML) and/or traffic forecasting techniques can be deployed to assist the slice provider to handle unexpected network situations and traffic pattern fluctuation which would involve SLAs violation. Concurrently, reinforcement learning (RL) solution such as Q-Learning solver can efficiently approximate the optimal slice admission policy that maximizes the mobile network operators' (MNOs) revenue. RL is a computational approach for goal-directed learning and decision-making, with the goal being to select actions to maximize future rewards [30]. An agent learns the future state through direct interaction with its environment, without the need of extra supervision or complete models of the environment.

Even though RL techniques are capable of executing in an online learning fashion with a much more reasonable computation cost, their processing time and reward mechanism may not be rapid enough to correctly respond to multiple-user RAN real-time traffic variation, implying congestion and SLA's violation. To overcome this issue, without the support of machine learning (ML) techniques, this work presents a novel dynamic RAN network slicing solution capable of providing customized network services able to guarantee the service provider's requirements while effectively utilizing network resources, as illustrated in the next subsection.

11.3.4 Dynamic RAN Slicing Performance: A 5G Case Study

In the next subsection, the innovative proposed approach is presented. This solution represents a 3GPP backward compatible framework able to dynamically perform the optimal radio resources partitioning among the served slices through the simultaneous analysis of the tenant SLA and real-time service requirements of the served users connected to each slice.

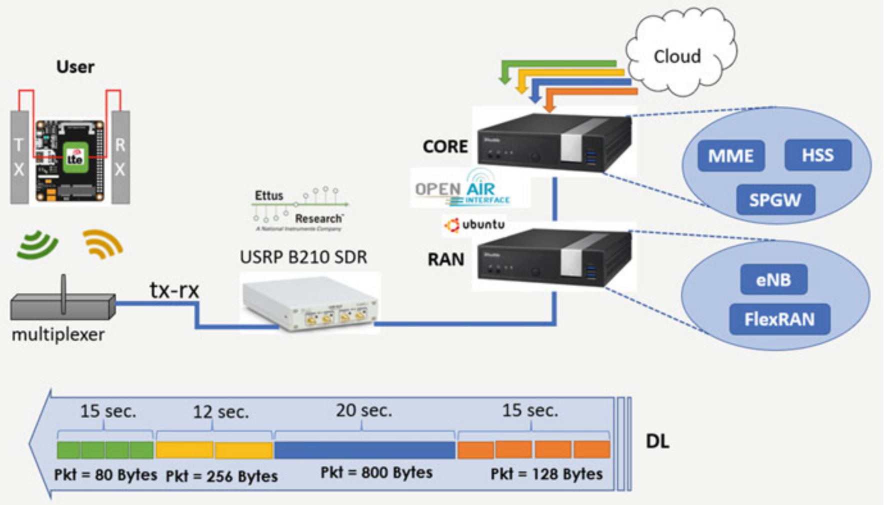


Fig. 11.9 Testbed design and architecture

As a comparison metric, the proposed dynamic network slicing approach is tested against the classical static slicing, using a real testbed and UE. At the time of writing this section, no open-source standalone 5G testbed solutions are available. Nevertheless, many 4G-based testbeds equipped with 5G functionalities are free and online available. Figure 11.9 shows the design and architecture of the testbed.

For the technological model view, OpenAirInterface (OAI) [31] platform has been chosen to conduct the experiments. OAI is an open experimentation and prototyping platform created by the Mobile Communications Department at EURECOM to enable innovation in the area of mobile/wireless networking and communications. For the RF part, universal peripheral radio software (USRP) B210, from Ettus Research [32], is utilized. This equipment provides a fully integrated software-defined radio (SDR) specifically designed for low-cost experimentation.

Multiple services are simulated using a Raspberry Pi platform as UE. This device is equipped with an LTE module for the communication with the 4G RAN.

Deployed on a separate machine from the CN, FlexRAN controller is a SDN controller belonging to the Mosaic5G project [33] and represents a flexible and programmable platform for software-defined RAN [34].

The tested scenario consists of a variable downlink UDP traffic injected from the CN toward the user using a single slice. For the same network environment and input traffic, the experiment is repeated for both the static and dynamic network slicing approaches, as it is illustrated in Fig. 11.10.

Since for the static approach the amount of reserved radio resources cannot vary during the service session, an overprovisioning of RBs is necessary to avoid violation of SLA requirements. As highlighted by the blue line, 27–28 RBs are provisioned in order to handle all the possible variation of input traffic, even if

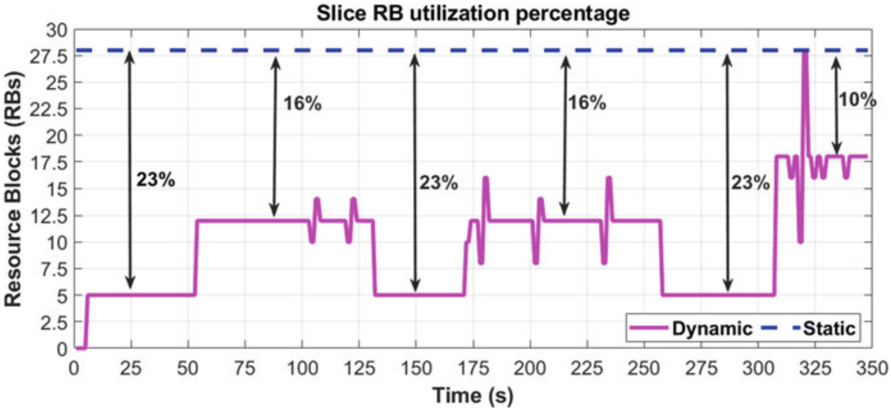


Fig. 11.10 Dynamic and static slicing experimental results

the average utilization is 13 RBs during the entire experimental. Dynamic slicing overcomes this waste of resources through a mapping of the slice RBs aligned with the input traffic pattern, as illustrated by the pink solid line. For this scenario, dynamic slicing reduces up to 23% the allocation of radio resources which can be utilized for other services or assigned to other slices.

The input burst traffic is subjected to multiple data peaks due to queue adjustments in the scheduler or retransmission of a large number of packets. These unexpected variations do not damage the traffic of other users since the resources are appropriately divided and isolated among the slices.

11.4 Continual AI-Driven Zero Touch Network

11.4.1 AI-Enabled Cloud-RAN

Algorithmic innovation can unleash the potential of the beyond 5G (B5G) communication system and underpins the progressive changes across all entities. In this regard, artificial intelligence (AI)-driven networks are envisioned to be the cornerstone for supporting a plethora of network services to harness the full potential of network slicing and enabling the automation of demand-aware management and orchestration (MANO). The zero-touch network is conceived as a next generation of network management that leverages the principles of NFV and SDN to fully automated operations. Zero-touch and automation of network slicing enable an on-demand configuration without the need for fixed contractual agreements and manual intervention as well as reduce capital expenses (CAPEX) and operating expenses (OPEX) in C-RAN. Studies [35, 36] show that mobile operators spend 60–80% of CAPEX on RAN technologies. On the other hand, China Mobile Research Institute

(CMRI) claims that by adopting C-RAN, a 15% reduction in CAPEX and a 50% reduction in OPEX can be achieved. The tendency toward zero-touch and fully automated approach has spurred intensive research interest to the application of ML in network slicing. ML can be used for solving NP-hard problems in C-RAN and has extensive application to automate various functions of MANO, such as resource management, dynamic network configuration, service creation, anomaly and fault detection, security, and reliability. Due to the nonstationary nature of the B5G network, the statistical model is unable to provide a well-aimed description for complex problems. In a challenging situation where prediction modeling faces several limitations, RL can provide a promising technique to be incorporated in the B5G network. Machine learning control (MLC) mechanisms solve optimal control problems for complex control tasks where it might be difficult or impossible to model the network [37]. In particular, continual lifelong RL can be considered as a very attractive approach in telecommunication networks where it is hard to collect a large number of training samples in such interactive environments [38].

11.4.1.1 Network and Service Management

Due to the exponential growth in the demand for service provisioning and immense challenges in B5G, it is necessary to pursue network slicing technology as a key enabler. A slice instance is self-contained in terms of traffic flow and operation which has specific characteristics where it can support several use cases, as indicated in Sect. 11.2. In this regard, network slicing can provide better performance, flexibility, and scalability compared to one-size-fit-all networks. For the transformation of the legacy NFs from hardware-based to softwarization, we need a new network management system. The European Telecommunications Institute (ETSI) NFs virtualization (NFV) has developed an NFV-MANO framework [39]. MANO plays a significant role in managing the correct operation of the NFV infrastructure (NFVI) as well as VNFs. Moreover, it provides the functionality required for the provisioning of VNFs, and the related operations, such as the configuration of the VNFs and the infrastructure that these functions run on. It includes the orchestration and lifecycle management of physical and/or virtual resources that support the VNFs [40 41]. The zero-touch framework and its underlying AI as a service (AIaaS) solutions enable MNOs to transform the corresponding network and service operations through the application of Machine Learning and flexible cloud scalability and also pave the way to assist with slice creation and guarantee the committed SLAs.

11.4.1.2 Zero-Touch Network Slicing

The zero-touch network and service management (ZSM) framework reference architecture [42] is designed to support fully automated network and service management. To this end, the ZSM architecture supports a set of architectural design principles including (i) modularity for creating self-contained and loosely coupled services to prevent monoliths and tight coupling, (ii) extensibility that enables the network to extend new services and service capabilities, (iii) scalability that fulfills the increasing or decreasing demands to deploy managed entities and thereby modules can be independently scaled, (iv) resiliency to cope with the degradation of the infrastructure and other management services, and (v) simplicity to ensure minimal complexity while still meeting the functional and nonfunctional requirements. The modular characteristic is paired with the use of intent-based interfaces, closed-loop operation, and AI/ML techniques to empower the full automation of the management operations [43].

As shown in Fig. 11.11, the architectural building blocks consist of management services, management functions, management domains, integration fabric, and data service. The management services are the pivotal entities in ZSM framework reference architecture that federate together management domains including a set of capabilities for communication purposes, automation orchestration, and managing one or more entities such as infrastructure resources that can be physical (e.g., physical network functions (PNFs)), virtual (e.g., VNFs or software-based services), and/or cloud-based (e.g., "X-as-a-service" (XaaS) resources). Each management domain splits into sub-domain to consider different management concerns. The integration fabric is a special management function that enables the communication between management functions within management domains while offering a set of communication capabilities such as synchronous and asynchronous communication. The data service allows us to separate data storage from data processing and support different types of storage mechanisms and database technologies to provide current management data such as configuration data, performance/fault alarm events, and topology data for AI-based closed-loop automation procedure. To achieve closed-loop operation, a management framework needs to provide means for the ordered invocation of the steps or phases of the closed-loop (e.g., observe and decide) [42]. Therefore, we should boost and tune the current dominant AI paradigm to improve learning in complex telecommunications environments for intelligent control and decision.

11.4.1.3 Data-Driven C-RAN Management

The data-driven approach enables the 5G/B5G network to automatically reinforce itself by using ML and data training without prior knowledge. However, it is difficult to obtain the user-related data, and we face some limitations to understand the demand behavior due to privacy issue. Unlike Deep Learning (DL) methods that need a large amount of data to understand and solve the problem, RL is a branch of ML which solves challenging decision-making and control issues in telecommunication environment where the agent generates the corresponding dataset on the fly through interaction with the environment (network).

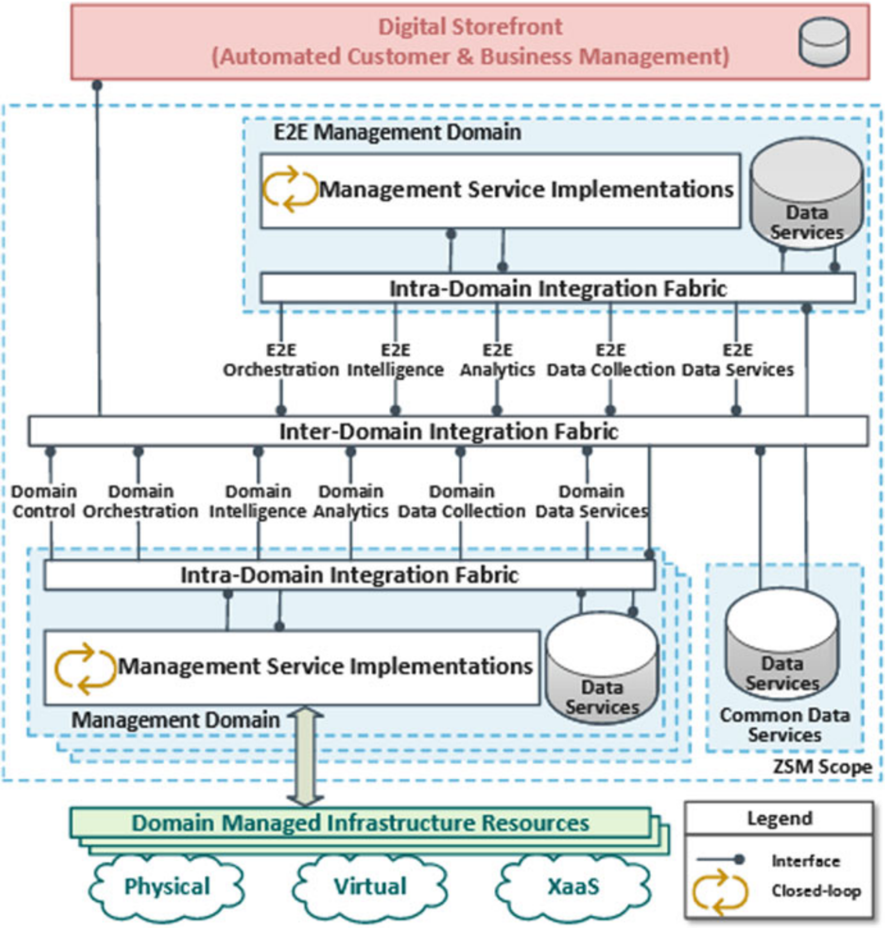


Fig. 11.11 ZSM architecture diagram [8]

In RL, the agent takes actions at a certain system state and learns the optimal action in a given situation and observes/perceives the corresponding responses from the environment. This is achieved through exploration and exploitation. A reward (or penalty) refers to feedback signals from the environment that implies the agent's performance. A challenge for the agent is to attain trade-offs between exploration and exploitation. The information on the environment will be provided by a state that helps an agent to select its action. The agent receives a reward while taking a good action, and it receives a penalty when taking a bad action so it pursues a trial-and-error strategy for possible optimal state-action pairs as a policy. The principal constituents of an agent comprise the policy and value function. A policy defines how an agent acts from a specific state. It maps states to actions, the actions that assure the highest long-term reward, and determines the next action based on the current state. The value function represents how good is a state for an agent to be in and attempts to find a policy that maximizes the returns by maintaining a set of estimated expected returns for various policies.

A model-based agent can handle partially observable environments, and the agent tries to build a model of how the environment works; here, the model refers to knowledge about how things happen in the environment, so the agent plans to get the best possible behavior. In contrast, a model-free agent (e.g., Deep Q-learning, Policy Gradient, Actor-Critic) just considers experiences and tries to figure out a policy of how to behave optimally to achieve accumulated long-term reward, and then it can become proficient without increasing either sample complexity or time complexity. For network slicing when a new slice is instantiated, the intelligent agent should configure the network parameters of the C-RAN based on network states while guaranteeing the maximum performance for other slices operations. We consider Actor-Critic methods as the state-of-the-art approach.

11.4.1.4 Lifelong Machine Learning in C-RAN

Lifelong machine learning is an advanced ML paradigm that learns continuously, accumulates the knowledge learned in the past, and uses/adapts it to help future learning and problem-solving. Lifelong learning has several key characteristics: continuous learning process, explicit knowledge retention and accumulation, and the use of previously learned knowledge to help learn new tasks [38]. It considers systems that can learn many tasks over a lifetime from one or more domains. They efficiently and effectively retain the knowledge they have learned and use that knowledge to learn more efficiently and effectively new tasks [39]. A lifelong approach should be computationally efficient when storing knowledge in long-term memory and needs the ability to choose relevant prior knowledge for solving new tasks while casting aside irrelevant or obsolete information. Indeed, it ensures the effective and efficient interaction of the retention and transfer elements [44]. The main criterion is the sequential nature of the learning process where only a small portion of input data from one or a few tasks is available at once. The main challenge is to learn without catastrophic forgetting [45]. It is well-known that neural networks (NNs) suffer from catastrophic forgetting, which refers to the phenomenon that when learning a sequence of tasks, the learning of each new task may cause the NNs to forget the models learned from the previous tasks. Without solving this problem, a NN is hard to adapt to lifelong or continual learning, which is important for AI [46].

11.4.2 Zero-Touch Resource Allocation Management

The overview efficiency and effectiveness of a network depend on how it utilizes and manages the corresponding resources. The traditional resource allocation models take a lot of time and computing capacity as well computational complexity. For machine learning algorithms, we need only build a learning model, and the machine can complete a lot of work through self-learning. Therefore, the application of AI in the field of resource allocation is an inevitable future trend [47]. In this case, without a training set, RL can be considered as a powerful method where the action is resource allocation and the objective is to find the best policy to maximize the return and obtain better performance in terms of network metrics.

Although in dynamic RL-based allocation of multiple resource types, the state space can increase exponentially with the number of resources, and also the approximation for off-policy control methods may not exhibit good convergence. To solve this issue, we use deep reinforcement learning (DRL). DRL is a combination of RL and DL in that it is leveraged to extract knowledge based on gained experience by interacting with the environment (network). In order to optimize the sliced resources, the agent(s) of network slicing should consider and update the Q-function for the optimal actions. Deep Q-learning benefits from a NN to approximate the Q-value function. The state can be considered as input, and the Q-value of all possible actions is generated as the output. An agent (neural network) in DRL continuously interacts with a network and well-suited to problems that have numerous possible states with high dimension. We use a function approximation like a NN to calculate the Q-values. In this regard, DRL has great potential for handling the extensive uncertainty and dynamic nature of network slicing based on modelling around deep neural networks (DNNs).

The deep Q-network (DQN) uses a critic to estimate the return or future rewards to accelerate the learning process as well as reducing the memory required to store the parameters of the model. Thus, how to deal with the uncertain dynamics of network nature is a crucial issue. Unlike the Q-learning, the NN, in contrast to learning some new values through several iterations, acquires these by predicting Q-values as close as possible to the target. The DNNs are very powerful, but NNs can be unstable or limited; therefore, to use DNNs as a function approximation in DRL, we use an experience replay buffer to store transitions. Initially, we should store random experiences in the buffer until the buffer is flipped.

The DQN-based method only works for control problems with a low-dimensional discrete action space, so it is not possible to apply Q-learning to continuous action spaces in a straightforward manner. Instead, we can use policy-based Actor-Critic methods. The Actor-Critic agents refer to how good the action is taken (value-based agent) and controls how our agent behaves (policy-based agent). Consequently, continuously learning and accumulating knowledge would yield the necessary experiences to adapt a zero-touch resource allocation agent to a new telecommunication environment quickly and to perform efficiently.

11.4.3 Challenges and Open Research Directions

While AI-driven MANO plays a pivotal role in empowering fully automated operations in ZSM, we believe that there are still many challenges and open research directions that need to be addressed by the community.

A. ***Functional Split***

The dynamic and flexible functional split in C-RAN is envisaged as a promising solution to overcome the capacity and latency challenges in the FH. The high-capacity and very-low-latency FH requirements can be considered as the main drawback of C-RANs. To cope with this challenge, all efforts are to relax and mitigate the FH requirements by a dynamic functional split approach between local radio remote heads and the centralized baseband unit pool. A functional split determines the number of functions left locally at the antenna site and the number of functions centralized at a high processing powered data center. The zero-touch network should enable us take on this challenge by promoting the dynamic placement of functional split options through a unified approach with other resource allocation issues. For example, the AI-driven mechanisms can dynamically select the optimal functional split concerning mobile traffic demand and daily traffic patterns/variations and thereby fine-tune FH bandwidth and C-RAN resources efficiently.

B. ***Co-location of multi-access edge computing (MEC) and C-RAN***

In contrast to a traditional network, in modern service-driven networks, some sites and functions are becoming surplus to requirements. MEC is a relatively novel paradigm that in synergy with C-RAN can be highly complementary technologies, especially for emerging services that need high bandwidth and low latency. The MEC architecture consists of functional entities and a MEC-MANO approach that follows the infrastructure-as-a-service (IaaS) model. To jointly enable MANO on C-RAN and MEC, we need posterior intervention and innovation concerning the harmonization of the C-RAN and MEC-MANO parts. The idea is to pursue a unified approach where NFV and MEC share the same MANO and NFV infrastructure (NFVI), because like C-RAN that benefits from the NFV paradigm, MEC also uses this approach to virtualize the elements and run applications in a virtualization platform. The MEC should interact with C-RAN and application concurrently. Indeed, many applications are running on the cloud while these cloud environments have different virtualization stacks compared to C-RAN. As a result, the efficient co-location of MEC and C-RAN is the key challenge, and intelligent zero-touch approaches can be candidate solutions.

C. ***Security***

The integration of SDN and NFV in network slicing raises multifaceted security and privacy issues that have to be addressed through the automation of security management functions. Moreover, the different levels of interaction on inter-domain and inter-slice can cause new security challenges, and thereby we should consider different levels of security for slices. In this context, it can be viewed in the following perspectives: (i) security aspects in the life cycle of a slice, (ii) security challenges concerning other slices (inter-slice security), and (iii) security issues inside a slice (intra-slice security). The ETSI has introduced the NFV MANO architecture with a security life cycle management module, namely, the NFV security manager (NSM).

This module guarantees secure network services operation in NFV environments by reacting to internal and external events. It is important to define how an AI-driven MANO and modules can solve security issues by adding new properties or elements to the slices. The AI-driven approach can guarantee security in inter-slice and intra-slice isolation. At the inter-slice level, the AI solutions can prevent attack-triggering events from one slice to other network slices when control functions are shared, such as reconfiguring the network parameters to increase component isolation in the slice. To solve the intra-slice security issues and control the components inside the slices, AI methods can be run as virtual software-based solutions such as autonomous traffic detection and modeling. The ability of AI methods to learn from the environment and to perform the proper configuration to cope with vulnerabilities will self-automate the slicing approach leading to new opportunities for their adoption.

11.5 Conclusion

This chapter presented a detailed overview of the C-RAN network architecture and discussed the advantages and challenges that need to be solved before its benefits can be fully exploited. Combining various management techniques based on the concept of network slicing, the scalability and flexibility principles advantages of this architecture enable multiple scenarios characterized by different service requirements.

Three C-RAN management approaches are presented. Using the functional splits principle, the first method presented a novel solution able to provide isolated and tailored slices for different services with customized functional splits per slice when CU and RRHs are connected through a FH/BH network. Through the sharing of the network infrastructure, this approach showed how multiple operators may instantiate their own slice requirements and satisfy different QoS.

Network slicing will only be as flexible and capable as the operators' network and spectrum assets allow. Novel orchestration methodologies of the physical radio resources should be investigated, where we present a novel framework for dynamic allocation of the radio resources. Through the dynamic real-time evaluation of the service KPIs, the proposed method defines the physical capacity of each slice, showing an increment of the QoS compared to classical static slicing methodologies.

Despite the business success of network slicing, MNOs consider equally important the speed at which slices are delivered and managed. The answer to this challenge lies in automating the lifecycle management of network slices, as investigated in the last part, where the zero-touch slice lifecycle ensures performance guarantees for existing slices with minimal human, and often error-prone, intervention. The benefits of AI-based methods on top of zero-touch networks are illustrated, ensuring a seamless service delivering for each network slicing.

The aforementioned analysis demonstrated many implementation choices and configurations settings using C-RAN, where MNOs could operate together and establish services according to their needs, inside different sections of the network infrastructure. This principle is aligned with the 5G specifications and represents a further step toward a global homogeneous network.

Acknowledgments This research work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreements No 722429 (5GSTEPFWD) and No. 722788 (SPOTLIGHT).

References

1. China Mobile. (2011). C-RAN: The road green RAN white paper, *China Mobile,* vol. V3.0.
2. 3GPP. TR 38.801, Technical specification group radio access network; Study on new radio access technology: Radio access architecture and interfaces, 3GPP, 2017.
3. ITU, IMT Vision- framework and overall objectives of the future development of IMT for 2020 and Beyond, ITU-R M.2083-0, 2015.
4. Lin, Y., Shao, L., Zhu, Z., Wang, Q., & Sabhikhi, R. K. (2010). Wireless network cloud: Architecture and system requirements. *IBM Journal of Research and Development, 54*(1), 4–1.
5. Suryaprakash, V., Rost, P., & Fettweis, G. (2015). Are heterogeneous cloud-based radio access networks cost effective? *IEEE Journal on Selected Areas in Communications, 33*(10), 2239–2251.
6. Alliance, N. (2015). *NGMN 5G white paper*. Technical report.
7. Tang, J., Wen, R., Quek, T. Q., & Peng, M. (2017). Fully exploiting cloud computing to achieve a green and flexible C-RAN. *IEEE Communications Magazine, 55*(11), 40–46.
8. Costa-Perez, X., Garcia-Saavedra, A., Li, X., Deiss, T., Oliva, A. D. L., Giglio, A. D., Iovanna, P., & Moored, A. (2017). 5G-crosshaul: An sdn/nfv integrated fronthaul/backhaul transport network architecture. *IEEE Wireless Communications, 24*(1), 38.
9. García Saavedra, A., Oliva Delgado, A. D. L., Costa Pérez, X., Gazda, R., Mourad, A., et al. (2019). Integrating fronthaul and backhaul networks: Transport challenges and feasibility results. *IEEE Transactions on Mobile Computing*.
10. ICPR, eCPRI Interface Specification, eCPRI specification v1.0, 2017.
11. I. W. Group. Next generation fronthaul interface. *I. W. Group.*
12. Kitindi, E. J., Fu, S., Jia, Y., Kabir, A., & Wang, Y. (2017). Wireless network virtualization with sdn and c-ran for 5g networks: Requirements, opportunities, and challenges. *IEEE Access, 5*, 19099–19115.
13. Wang, X., Cavdar, C., Wang, L., Tornatore, M., Chung, H. S., Lee, H. H., Park, S. M., & Mukherjee, B. (2017). Virtualized cloud radio access network for 5G transport. *IEEE Communications Magazine, 55*(9), 202–209.
14. Ferrus, R., Sallent, O., Perez-Romero, J., & Agusti, R. (2018). On 5G radio access network slicing: Radio interface protocol features and configuration. *IEEE Communications Magazine, 56*(5), 184–192.
15. Ranaweera, C., Wong, E., Nirmalathas, A., Jayasundara, C. & Lim, C. (2017). 5g c-ran architecture: A comparison ofmultiple optical fronthaul networks. In *2017 international conference on Optical Network Design and Modeling (ONDM)*, 2017.
16. Zaidi, Z., Friderikos, V., Yousaf, Z., Fletcher, S., Dohler, M., & Aghvami, H. (2018). Will sdn be part of 5g? *IEEE Communications Surveys and Tutorials, 20*(4), 3220–3258.
17. Ojaghi, B., Adelantado, F., Kartsakli, E., Antonopoulos, A., & Verikoukis, C. (2019). Sliced-ran: Joint slicing and functional split in future 5g radio access networks. In *IEEE International Conference on Communications (ICC)*, Shanghai, 2019.
18. Small Cell Forum, Small cell virtualization functional splits and use cases, Small Cell Forum, 2016.

19. Foukas, X., Marina, M. K., & Kontovasilis, K. (2017). Orion: RAN slicing for a flexible and cost-effective multi-service mobile network architecture. In *Proceedings of the 23rd annual international conference on Mobile Computing and Networking, MobiCom*, New York, 2017.
20. Garcia-Saavedra, A., Costa-Perez, X., Leith, D., & Iosifidis, G. (2018). FluidRAN: Optimized vRAN/MEC Orchestration. In *IEEE INFOCOM*, 2018.
21. Oliva, A. d. l., Hernandez, J., Larrabeiti, D., & Azcorra, A. (2016). An overview of the CPRI specification and its application to C-RAN- based LTE scenarios. *IEEE Communications Magazine, 54*(2).
22. Categories and Service Levels of Network slicing White paper.
23. Network slicing management & prioritization in 5Gmobile systems.
24. Slice Management for Quality of Service Differentiation in wireless network slicing.
25. The Structure of Service Level Agreement of Slice-based 5G Network.
26. *Managing violations in service level agreements.*
27. Chang, C.-Y., et al. (2018). Slice orchestration for multi-service disaggregated ultra-dense RANs. *IEEE Communications Magazine, 56*(8), 70–77.
28. Navarro-Ortiz, J., Sallent, O., & Pérez-Romero, J. (2020). Radio access network slicing strategies at spectrum planning level in 5G and beyond. *IEEE Access, 8*, 79604–79618.
29. 3GPP TS 22.101 V15.1.0, Study on Radio Access Network (RAN) sharing enhancements, June. 201.
30. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.
31. Nikaein, N., et al. (2014). OpenAirInterface: A flexible platform for 5G research. *ACM SIGCOMM Computer Communication Review, 44*(5), 33–38.
32. https://www.ettus.com/all-products/UB200-KIT/
33. Nikaein, N., et al. (2018). Mosaic5G: Agile and flexible service platforms for 5G research. *ACM SIGCOMM Computer Communication Review, 48*(3), 29–34.
34. Foukas, X., et al. (2016). FlexRAN: A flexible and programmable platform for software-defined radio access networks. In *Proceedings of the 12th international on conference on emerging networking experiments and technologies*. ACM.
35. Habibi, M. A., Nasimi, M., Han, B., & Schotten, H. D. (2019). A comprehensive survey of RAN architectures toward 5G mobile communication system. *IEEE Access, 7*, 70371–70421.
36. Bouras, C., Ntarzanos, P., & Papazois, A. (2016). Cost modeling for SDN/NFV based mobile 5G networks. In *Proceeding of the 8th International Congr. Ultra Mod- ern Telecommun. Control Syst. Workshops*, Lisbon, Portugal, 2016, pp. 56-61.
37. Moe, S., et al. (2018). Machine learning in control systems: An overview of the state of the art. *Artificial Intelligence, XXXV*, 250–265.
38. Chen, Z., & Liu, B. (2018). *Lifelong machine learning* (2nd ed.). Morgan& Claypool Publishers.
39. ETSI NFV ISG, "Network Function Virtualization (NFV) Management and Orchestration," GS NFV-MAN 001 v0.8.1,Nov. 2014.
40. ETSI GS NFV-MAN 001, "Network Functions Virtualisation (NFV); Management and Orchestration," http://www.etsi.org/, December 2014, ETSI Industry Specification Group (ISG).
41. Mijumbi, R., et al. (Jan. 2016). Management and orchestration challenges innetwork functions virtualization. *IEEE Communications Magazine, 54*(1), 98–105.
42. ETSI GS ZSM 002, "Zero-touch Network and Service Management (ZSM); Reference Architecture," Aug. 2019.
43. Benzaid, C., & Taleb, T. (2020). AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions. *IEEE Network*, 1–9.
44. Tessler, C., Givony, S., Zahavy, T., Mankowitz, D. J., & Mannor, S. 2016. A deep hierarchical approach to lifelong learning in Minecraft. ArXiv e-prints.
45. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, .A, Slabaugh, G., Tuytelaars, T. Continual learning: A comparative study on how todefy forgetting in classification tasks, arXiv preprintarXiv:1909.08383.

46. Wenpeng Hu, Zhou Lin, Bing Liu, Chongyang Tao, Zhengwei Tao, Jinwen Ma, Dongyan Zhao, & Rui Yan. (2019). Overcoming catastrophic forgetting for continual learning via model adaptation. In *International conference on learning representations*.
47. Zhang, H., Feng, M., Long, K., Karagiannidis, G. K., & Nallanathan, A. (2019). Artificial intelligence-based resource allocation: Applications in ultradense networks. *IEEE Vehicular Technology Magazine.*

Chapter 12
Resource Management for Cost-Effective Cloud Services

Armin Okic, Ioannis Sarrigiannis, Umberto Fattore, Bin Xiang, Alessandro E. C. Redondi, Elisabetta Di Nitto, Angelos Antonopoulos, Kostas Ramantas, Christos Verikoukis, Luis M. Contreras, and Marco Liebsch

Abstract In line with the virtual network function (VNF) paradigm, network functions are abstracted and relocated from dedicated appliances to generic servers, thus providing the enabler for savings in terms of total cost of ownership (TCO). In fact, hardware overprovisioning induced costs can be saved due to the on-demand capability of scaling up and down the server capacity via a software setting. Moreover, when the user migrates between networking ecosystems, cloud management services needs to accommodate the migration of virtual resources between networks. Therefore, we address how resources are pooled, forecasted and migrated between abstract servers to have computing resources on-demand. This is demonstrated within a fog-enabled C-V2X architecture and UDN deployment. Furthermore, the pooling of computational resources for implementing RAN func-

A. Okic (✉) · B. Xiang · A. E. C. Redondi · E. Di Nitto
Politecnico di Milano, Milano, Italy
e-mail: armin.okic@polimi.it; bin.xiang@polimi.it; alessandroenrico.redondi@polimi.it; elisabetta.dinitto@polimi.it

I. Sarrigiannis · K. Ramantas
Iquadrat, Barcelona, Spain
e-mail: isarrigiannis@iquadrat.com; kramantas@iquadrat.com

U. Fattore · M. Liebsch
NEC Laboratories Europe GmbH, Heidelberg, Germany
e-mail: umberto.fattore@neclab.eu; marco.liebsch@neclab.eu

A. Antonopoulos
Nearby Computing, Barcelona, Spain
e-mail: aantonopoulos@nearbycomputing.com

C. Verikoukis
Centre Tecnològic de Telecomunicacions de Catalunyal, Parc Mediterrani de la Tecnologia (PMT), Castelldefels, Spain
e-mail: cveri@cttc.com

L. M. Contreras
Telefónica I+D/Global CTIO Unit, Madrid, Spain
e-mail: luismiguel.contrerasmurillo@telefonica.com

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_12

tions according to cell load requirements in 5G RAN can provide cost-effective centralized detection at the MEC (mobile edge computing) node. The optimization framework for jointly optimizing and managing MEC resources for baseband processing is considered.

12.1 Introduction

With the massive increase in mobile users and generated cellular data traffic, which according to Cisco is foreseen to reach sevenfold on 2016 values in 2021, the telco community is revisiting the traditional mobile network architecture completely and aiming to introduce novel hardware and software technologies to support and optimize the delivery of different application services. The 5G and beyond networks aim to cope with this demand, to maximize the quality of service for end users, while ensuring that mobile operators are managing these complex networks in an efficient way by reducing the total cost of ownership.

One of the approaches on which 5G and beyond networks will rely on are virtualization technologies and slicing of networks. The virtualization is envisioned to enable full flexibility in mobile networks by partitioning network functions into virtual instances, which can be deployed on general-purpose hardware and moved dynamically in the network wherever it is needed. On the other hand, slicing relies on virtualization technologies and provides a basis for network operators to "slice" network resources, in both the RAN and Core parts of the network, that can be allocated for specific types of network traffic, mobile users, services or applications. Once the network functions or services are abstracted into virtual functions, they can be reallocated to different parts of the mobile network, where generic processing hardware is deployed, i.e. cloud or edge of the network. Resource management of virtual instances provides optimal allocation of network resources in response to the underlying mobile network demands. Such an approach leads to improvements in the utilization of network infrastructure and in the reduction of total costs.

In this chapter, the main focus is on enabling technologies targeting resource management for cloud infrastructures and supported services. The chapter is organized as follows: Section 12.2 provides an overview of general 5G-enabled MEC architecture based on the VNF paradigm, showcasing a C-V2X deployment as a possible use-case scenario. Section 12.3 introduces 5G system enhancements for flow-optimal routing of user traffic based on collaboration between MECs for UDNs. Section 12.4 targets virtualization of radio access network functions and estimation of RAN computational load requirements in order to enable cost-effective centralization of RAN resources; and finally, Sect. 12.5 provides an optimization framework for the reallocation of network resources within a MEC architecture.

12.2 E2E 5G-Enabled MEC Architecture

Multi-access edge computing (MEC) [1], formerly known as mobile edge computing, is a term given by the European Telecommunications Standards Institute (ETSI), when referring to the cloud computing capabilities offered at the edge of the network. The term was changed in September 2017, in order "to embrace the challenges in the second phase of work and better reflect non-cellular operators' requirements" [2]. MEC is responsible for delivering computing, storage and networking resources to the end user, similar to the cloud computing paradigm. User equipment (UE) can benefit from ultra-low latency (e.g. 1 ms) and high bandwidth (e.g. 10Gbps) but also from increased reliability and security, since the "cloud" is deployed at the edge of the network, in the end user proximity.

We consider an E2E MEC-enabled architecture [3], depicted in Fig. 12.1, where a heterogeneous radio access network (RAN) topology is considered. In particular, we consider a network that includes standalone 5G base stations (gNBs) and a cloud RAN deployment, where baseband units (BBUs) are connected with remote radio head (RRH) units. This architecture fully supports network function virtualization (NFV) by enabling the virtualization of computing, network and storage resources at the MEC and cloud hypervisors, located at the edge and core nodes, respectively. The 5G core control functions reside in the core node and the user plane function (UPF) that steers the traffic to desired applications and network functions reside in the edge node. The gNBs communicate among them using the NG interface and with the UPF using the N3 interface.

12.2.1 Virtual Network Functions

Software-defined networks (SDN) [4] replace the specific network equipment that is currently widely used with software that can be executed in generic purpose hardware, while NFV [5] enables the virtualization of this networking software. The restrictions applied by the legacy LTE networks, e.g. the fixed placement of the network functionalities, are eliminated with the aid of SDN and NFV technologies. Hence, application and network functionalities are handled as virtual network functions (VNFs) and are managed by an NFV Orchestrator (NFVO) [6] that is able to manage the various locations of the distributed system. Finally, the virtualized infrastructure manager (VIM) is responsible for the management and control of the computing, storage and network resources of the NFV infrastructure, while the NFVO performs the computing and network resource orchestration. They both reside in the core node (Fig. 12.1).

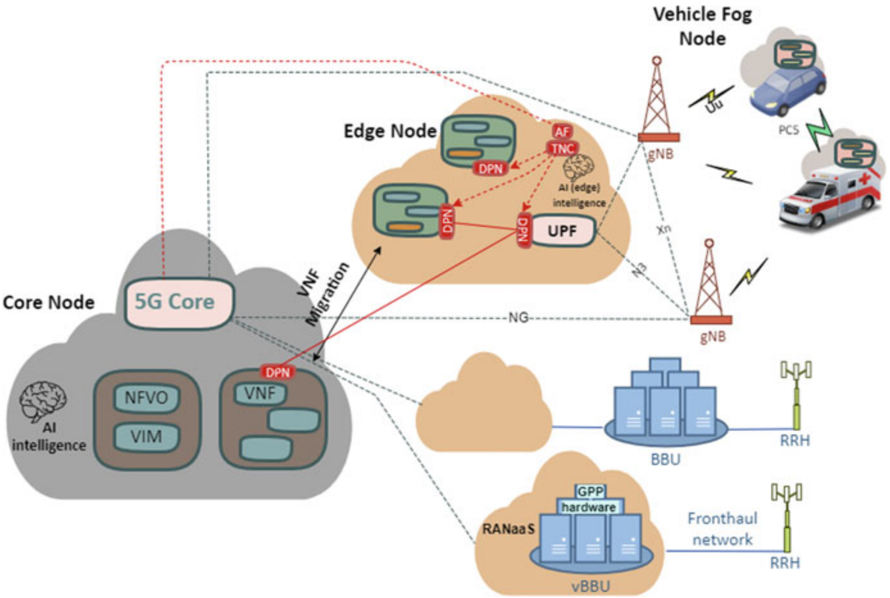


Fig. 12.1 E2E MEC-enabled architecture

12.2.2 Life Cycle Management of Virtual Network Functions

The VNFs of our architecture are described by a life cycle, which is governed by the NFVO. The NFVO is considered as the central controller of the system and executes periodic checks in order to monitor the availability of the resources, while ensuring that the NFV infrastructure adapts to the various traffic variations. In addition, the NFVO acts as an admission controller, in terms of filtering the incoming requests and (re)allocating the physical resources. The VNF life cycle in our system supports the live migration and scaling functionalities:

- **Migration** is the process that involves moving a VNF to a different hypervisor for resource optimization purposes. In the non-live migration, the instance shuts down, is moved to a new hypervisor and gets restarted. When no service interruption occurs, the migration is considered as live; the instances at the old and new hypervisor are running simultaneously while service migration is performed, and as a final step, the contents of the RAM are migrated. Our architecture supports both migration types, but only live migration is used.
- **Scaling** is the ability of the VNFs to scale out upon increased resource utilization and scale-in upon resource underutilization. This can be achieved either by horizontal scaling, i.e. the instantiation of more VNFs, or vertical scaling, i.e. the increase of the resources allocated to a VNF. While vertical scaling is used by traditional applications that require larger hardware to scale, cloud-based or distributed applications need more discrete hardware. Additionally, horizontal

scaling provides elasticity, redundancy and continuous availability, while vertical scaling can suffer from downtime when the host server is down or when the instance gets scaled. Thus, our architecture supports horizontal scaling only.

12.2.3 Fog-Enabled C-V2X Architecture for Distributed 5G Applications

Vehicle-to-everything (V2X) communications is a major area of Internet of things (IoT) that will enable communication between vehicles and between vehicles and infrastructure. Cellular-V2X (C-V2X) was introduced by the Third-Generation Partnership Project (3GPP), under the Release 14 [7], using LTE-based RAN for V2X communications.

In order to address some of the core and edge limitations, fog computing “distributes computing, storage, control and networking functions closer to the users, along a cloud-to-thing continuum” [8]. Routers, switches, servers, vehicles and smartphones are some of the underlying infrastructures that were previously working autonomously but can now participate in a connected environment, aiding in the fog computing realization. Therefore, expanding the E2E MEC-enabled architecture that was described in the previous section, the vehicles can also provide resources, such as computing, network and storage, creating a fog-enabled C-V2X architecture [9]. Apart from the core and edge nodes, resource virtualization is also supported at the vehicle fog nodes. The gNBs communicate with the vehicles using the Uu interface. Finally, the vehicle fog nodes can also communicate between them directly using the PC5 interface (Fig. 12.1**).**

12.2.4 Autonomous Driving Application Case Study

12.2.4.1 Applications System Definition for C-V2X-Based Deployments

In accordance with the NFV paradigm, a distributed application model can be envisaged, where applications consist of VNFs. Each VNF can run as:

- Virtual machine (VM) - environment that fully virtualizes a physical computer. VMs can host resource demanding applications, but their instantiation time could reach a few minutes.
- Container - lightweight virtual environment. Containers are used for lighter applications, and their instantiation time could be up to few seconds.
- Unikernel - even more shrunken virtualization environment. Unikernels can be instantiated almost instantly, but they can only run a single process, and the development of unikernel-based applications is more challenging.

The combination of the aforementioned virtual environments, distributed across the three different processing layers, creates an application-as-a-service function chain (AaaSFC).

12.2.4.2 AssSFC for Driving Applications

An autonomous driving application can be described as a collection of many applications that cooperate. In a simplified version, it could involve HD (high-definition) sensor data, such as lidar that measures the distances by illuminating targets with laser light and HD cameras, and the navigation and traffic jam avoidance data, where map data are combined with real-time traffic data. The autonomous driving application deployed as an AaaSFC could involve the following modules (Fig. 12.2):

At the core:

- One VM module that hosts the latest map data (MD) of a geographical area.

At the edge:

- One VM module that monitors the real-time traffic (RTD) data of an urban area.

At the vehicle:

- Two unikernel modules where the data of the lidar and the HD camera (LHD) are collected and the route is determined (RT).
- One container module that monitors the data of the unikernels and notifies the involved parties (MON).

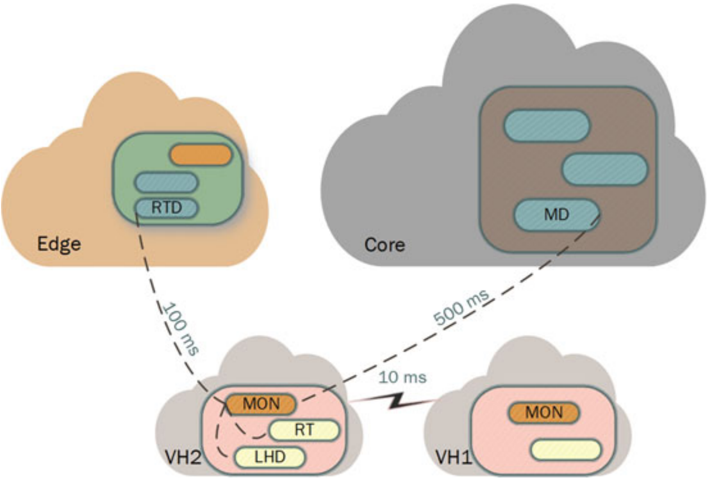


Fig. 12.2 AaaSFC for autonomous driving

In case there is a request for emergency service, the best route is requested by the RTD from the MON, based on real-traffic data, and the RT is updated with the route. As soon as the vehicle is moving, the data from the LHD will be used by the MON to notify all nearby vehicles using the direct interface in order to adjust their speed. When a traffic jam is detected, the RTD forwards an alternative route to the MON. In case the RT map data are not updated, the MON requests the additional map from the MD and forwards it to the RT. In the last part, the edge acts as the intermediate in vehicle-to-core communication.

12.2.4.3 AaaSFC Life Cycle Management

Based on the application, each VNF has specific virtual resources allocated to it, and each virtual link that connects two VNFs can tolerate a maximum amount of latency. In order to prevent possible service level agreement (SLA) violations that can occur under real data traffic conditions, live migration or scaling decisions might occur.

- Core-to-edge live migration decision: In the case of increased traffic, the required latency of a service hosted at the core could be violated. Thus, we propose a live migration action from the core to the edge, in order for the service to be closer to the vehicles. In the case of insufficient edge resources, a live migration action from the edge to the core of other VNFs could take place in order to free up the needed edge resources for the core service accommodation.
- Edge-to-edge migration decision: Another action that could result in migration is due to the vehicles' mobility. Since the vehicles move from one gNB's proximity to another's, a live migration action of the service might be needed from the old to the new edge node, following the user's handover process between the gNBs.
- Scaling decision: In the case of one or more of the VNFs that are part of the AaaSFC reach their maximum utilization capacity, the NFVO will decide a scale-out operation by creating a new copy of the overutilized VNF. This process can be repeated until the traffic can be adequately handled. The NFVO is also responsible for the reverse process, the scale-in, when the traffic is restored to normal.

12.2.4.4 E2E Performance Evaluation

In order to validate the described architecture and life cycle management actions, a set of experiments have been conducted, with the aid of a 5G experimental platform. This platform consists of one control server for the management and orchestration of the physical and virtual resources and one core and two edge servers for the core and edge processing needs, respectively. Common hardware was used (seventh and eighth generation Intel i5 processors, 32 GB RAM, 1 Gbps Ethernet network interfaces). With respect to the software installation, the selected VIM is OpenStack

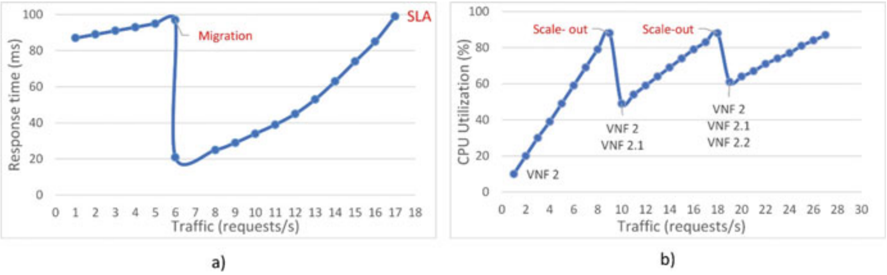


Fig. 12.3 (**a**) live migration and (**b**) scaling demonstration

[10], while the NFVO software that was used is Open Source MANO [11]. In what follows, the live migration and scaling functionalities are demonstrated.

Live Migration Experiment

The first experiment shows a vehicle that leaves the proximity of a gNB and enters another's. As depicted in Fig. 12.3a, the period leading up to the first six requests/s, the vehicle is still connected to the serving gNB and server where the response time reaches the maximum SLA (100 ms). As the traffic requests increase, the SLA will be violated. Thus, we implement a live migration action of the service from the original serving edge server closer to the new edge server associated to the handover gNB, and that will logically be closer to the vehicle. Hence, the latency is reduced, as the service is hosted closer to the vehicle's proximity and the edge-to-edge communication is eliminated. Compared with monolithic environments that do not support migration features, using the live migration action, we can support more requests without SLA violation.

Scaling Experiment

The second experiment demonstrates a high service demand (Fig. 12.3b). When stressed (over 90% CPU utilization), the VNF that serves the current service (VNF 2) becomes unstable. Thus, a scale-out process takes place prior to the 90% threshold, and VNF 2.1 is instantiated. Then, the traffic is equally distributed to the two VNFs. As the traffic keeps increasing, the average CPU utilization increases as well, and a third VNF (i.e. VNF 2.2) is instantiated to handle the additional traffic. The reverse process (i.e. the scale-in) takes place when the traffic is restored back to normal and the extra VNFs are destroyed. This way, we can efficiently manage the applications and serve more users on-demand, compared with the traditional systems that do not support scaling features.

12.2.5 Conclusions

In this section, we provided an E2E MEC-enabled architecture, able to exploit the interplay between the core and the edge. Additionally, we introduced VNF life cycle management functionalities that are able to manage a multilayer architecture. Furthermore, we introduced the vehicle fog layer to the infrastructure, resulting in a fog-enabled C-V2X architecture for distributed 5G applications, while we proposed the definition of an AaaSFC for C-V2X-based applications. Finally, we introduced an experimental 5G platform, based on which we demonstrated the live migration and scaling features of our architecture.

12.3 Optimal MEC Integration and Collaboration in UDNs

The 5G mobile networks are expected to deal with an enormous number of users, which led to the notion of ultradense network (UDN) [12] to support hotspot concentrations of users. Furthermore, a different group of users will have heterogeneous requirements, e.g. with regard to some automotive applications, low latency will be a strict requirement, and high mobility of users will have to be considered. To address low latency, the European Telecommunications Standards Institute (ETSI) [21] is working on integrating its specified MEC system with the 5G system, which is being defined by the Third-Generation Partnership Project (3GPP) [13, 14]. Whereas the deployment of services in MEC units (allocated at the edge of the network) allows reducing the topological distance between the mobile user and the service (i.e. reducing the latency of the connection), there is also a need to consider strong collaboration mechanisms between MEC units to deal with the high mobility of vehicles that also includes techniques for proactive allocation of resources anticipating the needs of the network in the near future.

An initial prerequisite to enable efficient collaboration and integration of MEC units in the 3GPP's 5G system is to detect all the main features of the next-generation mobile network which may be leveraged. Of particular interest is the infrastructure beyond the (radio) access network, referred to as 5G core (5GC) [13]. This aspect is addressed in Sect. 12.3.1. Furthermore, Sect. 12.3.2, briefly addresses the motivations and key issues beyond MEC unit integration and collaboration in UDN. Finally, Sect. 12.3.3 addresses motivations and benefits beyond proactive allocation of resources on MEC units and introduces enabling technologies for such kind of allocation, in particular discussing different machine learning approaches for accurate user mobility prediction.

12.3.1 Enhancing the 5G System for Flow-Optimal Routing of User Traffic

In the transition from the fourth to the fifth generation, many enhancements have been introduced in order to increase the overall flexibility and the programmability of the core network, in particular in the direction of enhanced virtualization of functions and components. In the control plane, basic control functions previously collected in single entities (e.g. MME, HSS) have been split into smaller and lighter functions, each deployable separately and all communicating the one with the other through a subscription/notification mechanism based on HTTP/2. This mechanism of control plane functions interworking is named service-based architecture (SBA). In the user plane, the complexity of two different entities (i.e. serving and PDN gateway (SGW and PGW)) has been simplified to a unique user plane function (UPF) element, depending on the specific need being able to perform either as a PDU Session Anchor (PSA) or as an Uplink Classifier (UP CL). Whereas the first UPF configuration still acts similarly to a classic fourth-generation PGW (as an anchor point for user traffic towards services), the introduction of an UPF configured as UP CL allows to enable an anticipated breakout for user traffic when this is needed, i.e. when services are allocated in an edge location and not beyond the PSA UPF. Such configuration was not possible with the previous fourth-generation mobile core, where traffic was forced to traverse the PGW element unless ad hoc non-3GPP fully adopted solutions were used [16, 17]. Therefore, in the 5GC for each user plane path, there must be at least one UPF (PSA) and optionally other UPFs acting as UL CL. Figure 12.4 depicts an example of the 3GPP configuration of function in the case of two possible data networks (DN), where two different PSA UPFs are used for each of the user paths. In an edge deployment scenario, DN-2 could, for instance, represent local services deployed at the edge. The aforementioned figure also depicts 3GPP's terminology for interfaces (e.g. N2, N4 for control plane; N3, N9 and N6 for user plane), together with the Access and Mobility Management Function (AMF) and Session Management Function (SMF) functions.

Whereas the current 3GPP solution indeed enables enhanced access to edge services, it still implies for the need of multiple user plane elements, bringing into account various drawbacks: first, each UPF encapsulates packets through the *GRE Tunnelling Protocol for User Plane* (*GTP-U*), which is known to reduce overall flexibility of packet steering and adds additional encapsulation overhead for each UPF [16]; second, each UPF element is under the control of the 3GPP control plane, which is centrally deployed, and therefore an eventual traffic steering rule which needs to be applied at the edge is anyway forced through the centralized control plane, resulting in a poorly optimized mechanism. A solution to the cited drawbacks is to deploy a unique UPF at the edge (i.e. near the access network) and then apply a more flexible traffic steering of plain-IP packets beyond the UPF (i.e. in the N6 interface according to 3GPP's terminology). This solution basically applies beyond 3GPP's domain and, therefore, can be implemented in compliance

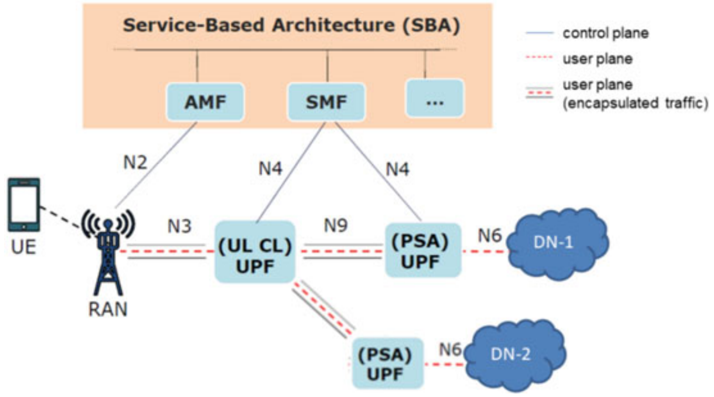


Fig. 12.4 5G core (5GC) 3GPP's specified architecture [13]

with current specifications, allowing at the same time for enhanced performances in the edge computing scenarios [18]. The solution is currently under discussion in various standardization bodies, such as the Internet Engineering Task Force (IETF) Distributed Mobility Management (DMM) Working Group (WG) [15]. Going beyond, some works further investigate the described solution adapted to constrained edge devices (e.g. drones) by deploying at the edge a low-power/low-consumption UPF implementation (e.g. using open-source data plane solutions [22]).

Moving the UPF PSA towards the access network requests for some alternative mechanism to treat traffic steering beyond the UPF (and, consequently, beyond 3GPP's domain). This may be obtained through data plane node (DPN) additional elements, properly instructed from some controlling entity, enforcing traffic steering rules on user traffic both in the uplink (directing towards the proper data network) and in the downlink (addressing the proper associated UPF PSA). These DPNs may implement a different kind of user plain protocols and methodologies in order to do this (e.g. segment routing, Identifier/Locator Separation Protocols) but, in any case, need to be instructed by a control entity taking proper decisions in a coherent way with the 3GPP control plane functions. A candidate element to allow this coordination between 3GPP and not 3GPP control entities is the application function (AF), defined by 3GPP at the scope of interworking with external functions [14]. The AF can be used either to receive information about modification occurring in the 3GPP control plane and related to some group of users (e.g. a handover is occurring) or to "suggest" decisions that are normally taken by 3GPP control functions (e.g. relocating a UPF responsible for some user traffic). From an implementation point of view, the AF behaves as any SBA function, subscribing to the control functions of interest (i.e. for user traffic updates, the AF should subscribe to the Session Management Function (SMF)).

The deployment described can be fully integrated with multi-MEC units, with the UPF PSA deployed at the edge platform and then DPNs located both at the edges and core to steer plain-IP user traffic properly. The 3GPP control plane functions are centralized, while the AF may be deployed locally with each MEC platform together with a Transport Network Controller (TNC) which instantiates policies on DPNs: in this way, whenever a policy should be instantiated locally on a MEC unit, there will be no need to pass through the centralized 3GPP control entities (this, of course, is true only for some specific scenarios). Figure 12.1 depicts the described components as well.

12.3.2 MEC Platforms Collaboration and Integration in UDNs

Next-generation mobile networks expect in the very near future an enormous increase in mobile users and devices [12], resulting in ultradense networks (UDNs) needing to deal with a very high number of users and high heterogeneity of requirements. Among all requirements, one which 5G networks promise to address is to enable ultra-low latency and reliable communication (URLLC), and one of the main enablers for this is given by multi-access edge computing (MEC) platforms. Such edge platforms deployed near the mobile user (i.e. near the access networks) allow having various required services instantiated in the very proximity of the mobile user, hence reducing the topological distance and, therefore, in many cases, user traffic latency as well as offloading the exchange of traffic from the core of the network [21].

Dealing with UDNs, therefore, requests the integration of MEC platforms in mobile networks. Whereas this is already addressed both in ETSI MEC [21] and some 3GPP study items and examples of such integration are addressed in Sect. 12.2, more efforts can be made to have a smoother integration of MEC platforms in the 5G system, starting from the 5GC user traffic plain-IP optimization described in Sect. 12.3.1. Whereas the aforementioned considerations stand for more architectural-related matters, other important issues to be considered also include taking a decision on the optimized placement of MEC platforms and/or of services and functions deployment on such platforms, i.e. on which platform each service should be deployed in order to maximize the gain for connected mobile users.

Apart from high density and strong requirement heterogeneity of users, 5G networks are also expected to deal with high mobility: connected vehicles moving across highways and cities, people using public transportation and connected flying devices (e.g. drones). This implies an additional challenge for MEC platforms to deal with because it will not be sufficient to deploy services in one specific platform for some users, but it will also be needed to have smooth mechanisms to move such services in order to always assure demanded requirements, e.g. following the user during its mobility [26]. From the MEC platform point of view, this implies being able to allow a smooth collaboration between MEC platforms to exchange services users' states and data both from an operational aspect (i.e. leveraging at the scope interfaces such as Mp3 interface [21]) and technological aspect (i.e. allowing for a fast migration of functions and user status leveraging container deployments [30], unikernels [31], etc.).

Being able to have an optimized collaboration of MEC platforms and fast migration of services between platforms may not be sufficient to offer service continuity and user requirements. To fill this gap, a proactive rather than a reactive approach for service allocation may be used, hence anticipating the need of the network (and of mobile users) and proactively act accordingly [28]. This is discussed in the following section.

12.3.3 Proactive Allocation of Services in Distributed MEC Architecture

Proactively allocating services in a distributed MEC architecture means being able to decide in advance which service to allocate in which edge location (i.e. MEC platforms) based on some information from the network.

Motivations behind the proactive allocation of resources and advances of such an approach mainly consist of the possibility to anticipate the need of the network and therefore potentially avoid long downtimes when migrating/allocating a service [32], which could cause an overall degradation in respecting requirements for the said service.

12.3.3.1 Enabling Technologies for Proactive Allocation

In order to enable proactive allocation of services on MEC platforms, it is necessary to know in advance *what* is needed (i.e. which services will be requested by mobile users) and *where* it is needed (i.e. in which edge location deploying these services will result in higher gain). Some kind of intelligence is therefore needed to answer these questions: such intelligence can be centralized among different MEC platforms (and therefore have a general view of each platform resources and location) or may be distributed; in the latter case, an additional level of collaboration between MEC platforms in order to exchange information at the scope is needed. Figure 12.1 depicts such intelligence as allocated either in the core or in the edge network.

A further point regards based on what such intelligence should make allocation decisions. For instance, in [23] such a decision is based on a plurality of different parameters, e.g. the resource availability on each MEC unit, the behaviour of network links and, among all, the expected demand from mobile users in the near future; only being able to know in advance what mobile users will need will in fact give the possibility to proactively configure the network to serve such users' demands adequately. [23] assumes the scope is to always keep the topological distance between services and users as low as resource availability allows (i.e.

reducing latency) and therefore investigates and proposes methodologies to predict user locations in the future instants, hence predicting user mobility.

In general, state of the art suggests many technologies allowing for user mobility prediction, ranging from modified Mobility Markov Chains (MMC) [24] to machine learning (ML)-based techniques; the latter is trying to detect historical patterns on the previous positioning of users to predict future location.

12.3.3.2 ML-Based Approach for Predicting User Mobility

Many different works exist in state of the art focusing on user mobility prediction, based on different methodologies; of these, some of the latest focus on machine learning approaches. In general, all these prediction works can be collected in two main categories: (1) works in which user locations are classified based on points of interest (PoI), e.g. home, work, restaurant, etc., and (2) works in which a grid-based approach is used; hence the user location is bound to a specific location in a grid. For instance, previously cited MMC-based [24] belongs to the first category.

Still belonging to the first PoI category but focusing on ML approaches are [29, 27]. Whereas the first leverages neural networks (NN) and the latter spatial-temporal extended recurrent neural networks (RNNs), both use a high number of previous positions in registered PoIs in order to predict those which will be occupied in the future.

The utilization of RNN for predicting user mobility has increased in the latest years since an RNN configuration results very useful in cases in which we want to predict the future value of a time series. Therefore, assuming all previous positions of a user building a time series, RNNs allow predicting with good accuracy the next position value for the user. Also, enhanced RNN configurations may be used, as for instance, long short-term memory (LSTM) RNNs [25] in order to consider the long history of previous data and not only the very recent patterns. Both [32, 33] use LSTM to perform user mobility prediction at different scopes ([32] targets on dual-connectivity towards a base station for an overall increase in performances), and results show that they achieve significant accuracy levels.

However, previously aforementioned works perform a single user next-position prediction, whereas for proactive allocation of services, it is more useful to predict the density of users at each MEC unit. In this context, the authors proposed the AutoMEC framework [23], based on the LSTM configuration, that is focused on leveraging density predictions to take decisions towards allocation, migration and scaling of MEC services. To do so, the designed solution is composed of two phases: a prediction phase and a decision phase.

In the prediction phase, LSTM and gated recurrent unit (GRU) RNNs are compared in order to find the best configuration in terms of loss, accuracy and model training time performances. A performance evaluation is carried out on simulated vehicular data generated through the SUMO[1] simulation tool on a real map, including a 20 km highway segment and some urban and rural areas in Germany. In the work, the LSTM configuration with four layers resulted in the best performance, allowing for above 85% accuracy after ten epochs.

In the decision phase, an experimental decision algorithm is designed, exploiting the predicted data in order to take decisions on the allocation, migration or scaling of MEC services. The algorithm also considers resource- (e.g. storage consumption) and service-related (e.g. minimum requirements for the service) parameters.

In order to prove the value of the framework, the authors compared the RNN-based decisions with an equivalent decision algorithm without prediction and with the ideal case of 100% accurate prediction. As depicted in Fig. 12.5, the case with RNN prediction (blue line) allows for a lower delay (upper-left figure) while still having similar storage utilization (lower-left figure) than the no prediction case (red line): this results in a higher efficiency value, defined as the ratio between available storage and overall delay (upper-right figure). Furthermore, the case with RNN prediction also performs very similarly to the ideal case (green line): this happens because of the significant prediction accuracy achieved thanks to the LSTM configuration used in the cited work.

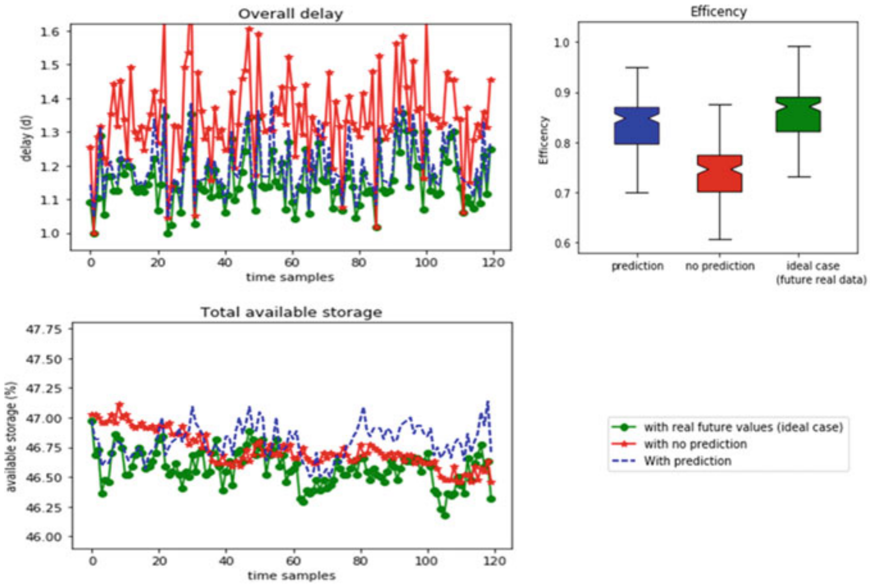


Fig. 12.5 With and without predicted mobility service allocation comparison [23], in terms of overall delay (upper-left), total available storage (lower-left) and an efficiency value (upper-right)

[1] An open-source, highly portable, microscopic and continuous multimodal traffic simulator, available at: https://www.eclipse.org/sumo/

12.3.4 Conclusions

In these sections, the optimal integration and collaboration of MEC units in UDNs have been discussed. This focused both on how to leverage the main features offered by standardization bodies (i.e. 3GPP 5GC, ETSI MEC) to enable a smooth integration of MEC units in 5G and how to harness technologies for proactive allocation of services in distributed MEC resources, e.g. using machine learning techniques to anticipate the future need of the network. All the aforementioned aspects of related works, both in research and standards, demonstrate that effective efforts are still ongoing on this topic and there is still significant space available for improvements and enhancements. In particular, presented works harnessing on ML-based prediction approaches show the value of using mobility prediction for real-time allocation and migration of edge resources in a multi-MEC environment.

12.4 Forecasting DSP Resources for Dynamic 5G C-RANs

For the first time, centralized or cloud radio access network (C-RAN) architecture was introduced in 2011 by China Mobile [31, 34]. They proposed C-RAN as a network architecture in which the conventional LTE base station, known as eNodeB, is disaggregated into two parts: (i) Radio Remote Head (RRH), located at remote site, primarily responsible for analog-to-digital signal conversion, and (ii) baseband unit (BBU), separated from a remote site, dedicated for the processing of baseband signals (see Fig. 12.1). This separation enables that several base stations combine their baseband processing into one shared BBU; in other words, numerous RRHs can utilize the resources of one centralized BBU. Sharing of common BBU among several RRHs leads to an increase in hardware utilization, reduction of deployment and maintenance costs, improvements in energy efficiency and enablement of certain technologies (e.g. Cooperative Multi-Point) [35].

The C-RAN architecture is also defined as one of the enablers of 5G RAN, in which RRH units are considered as distributed units (DUs) and BBUs are centralized units (CUs) [36, 37]. The separation point between DU and CU is identified by a functional split of base station protocol stack, which specifies the processing functions of both units. Further details on cost-effective deployments of baseband processing within C-RAN architecture are given in Sect. 12.4.1.

Since the beginning of C-RAN, it was envisioned that improvements in the processing power of multi-CPU architectures and maturing of virtualization technologies would lead to the development of baseband signal processing within virtualized instances deployed within general-purpose processing (GPP) hardware of cloud infrastructures [34]. Recent enhancements in those technologies enable virtualization of RAN functions into VNFs within NFV architecture, leading to the flexibility of virtual RAN (vRAN) deployments in different parts of the mobile network. Such a paradigm supports dynamical reallocation of vRAN resources, which is further motivated through several reasons: (i) computational loads of vRAN functions can be estimated [38, 39]; (ii) computational requirements are related to network utilization and repetitive patterns of users behaviour, which can be forecasted [40–42]; and (iii) reduction in CAPEX costs of the network [38, 43, 45]; Sect. 12.4.2 provides further insights about dynamic vRAN resources allocation within C-RAN architecture.

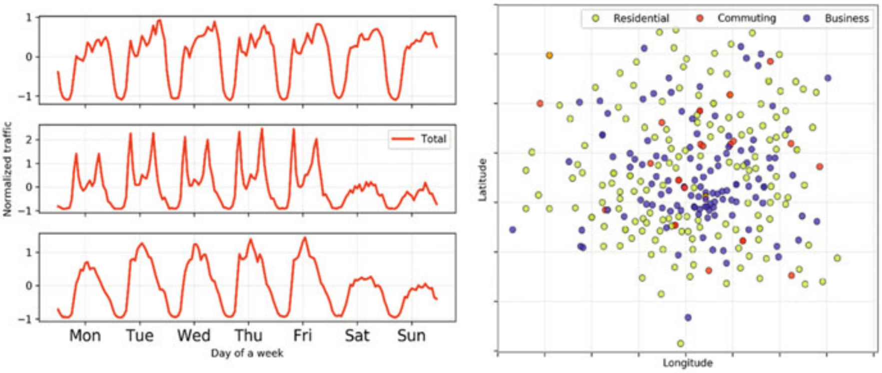
12.4.1 Cost-Effective Sharing of Computational Resources Among RRHs

In general, for network cost estimation, network operators are calculating total cost of ownership (TCO), which includes capital expenditure (CAPEX), relevant to the construction of the mobile network and operating expenditure (OPEX), relevant to costs for network operation. A high increase in the number of users in mobile networks has a significant impact on network resources that requires further investment by operators to purchase and deploy new networking infrastructure that typically can reduce the average revenue per user. This motivates network operators to expand existing network infrastructure in a more efficient way. The deployment of new sites, in general, is one of the highest CAPEX costs, while China Mobile estimates that 72% of the total power consumption comes from the cell sites (OPEX) [34]. Fortunately for the network operators, both TOC costs can be reduced with the implementation of C-RAN architecture.

The centralization of RAN computational resources is part of the 5G standard, which supports different disaggregation options (or functional splits) of RAN, leading to higher flexibility for network operators in adjustments of existing infrastructure towards 5G networks. Moreover, a flexible functional split is introduced enabling optimal utilization of RAN resources over time [35, 44].

Apart from reducing network deployment and maintenance costs, C-RAN directly supports the 5G paradigm as far as enabling higher densification of base stations, joint coordination of adjacent next-generation base stations (gNodeBs) with control of intercell interference, optimal utilization of RAN resources and introduction of sharable RAN resources between network operators within RAN as a Service (RANaaS) concept.

In order to satisfy the quality of service (QoS) of users, network operators need to provide network services and network connectivity to users no matter how big is the total demand on the network. This often imposes that network operators provision network resources to support peak hour traffic demands. Due to high user mobility, users' patterns are highly dynamic over more extended time periods and depend on the geographical locations of the deployed base stations. To demonstrate this attribute, Fig. 12.6 was plotted that represents different patterns of users' behaviour over time in different areas of one typical European middle-sized city. The left-hand side graph shows three dominant behaviour patterns related to residential,



(a) (b)

Fig. 12.6 Different behaviour patterns of users: (**a**) three main clusters of users' behaviour over a week corresponding to residential, commuting and business areas, respectively; (**b**) spatial distribution of base stations in a metropolitan city, with indicated dominant base station behaviour

commuting and business areas, respectively, from top to bottom. The right-hand side graph represents a map of base stations locations with indicated types of behavioural patterns.

Overlapping this map on top of a geographical map,[2] it was confirmed that the underlying geographical areas are matching behaviour patterns. As types of covered areas are not varying over time, together with users' moving patterns, the traffic demands at the base station level are repetitive over time and can be expected. Several works are exploiting such a feature and provide a framework for sharing BBU resources among different RRHs [45, 46]. However, their solutions are based on static RRH-BBU associations (static C-RAN), mainly relying on complementary traffic patterns (e.g. between residential and business areas) to increase hardware utilization and, therefore, with limited improvements. Another approach is known as dynamic C-RAN, which utilize virtualization technologies and placement of BBU functions into virtual BBUs (vBBUs). In such a way, vRAN resources can be dynamically scaled up and down, depending on the actual or predicted traffic demands [38] (more details are provided in Sect. 12.4.2). This paradigm can be further improved by introducing flexible C-RAN, which enables the movement of RAN computational resources in a dynamic manner between DU and CU, and vice versa [44].

[2] Exact locations of base stations are hidden in the geographical map, not to disclose any information about the mobile operator's deployment of the network.

12.4.1.1 RAN as a Service in 5G and Beyond

RAN as a Service (RANaaS) is part of a XaaS (anything as a service) paradigm in which any function of a system can be separated from the context, virtualized and offered as a service on-demand through cloud platforms. In RANaaS, network operators or network providers can share RAN computational resources with other network operators and, in this way, increase revenues of already deployed network infrastructure during time periods when it is underutilized. In order to make RANaaS possible, it is necessary that flexible and dynamic virtualization of RAN resources is supported within the C-RAN architecture. RANaaS is cloudified radio access network delivered as an on-demand service in an elastic and pay-as-you-go manner [47, 48]. A recent overview on state of the art in radio virtualization topics is given in [49], where the authors provide an evaluation of currently available solutions towards RANaaS.

12.4.1.2 Enabling Technologies for Computational Resource Allocation

Two main enabling technologies for disaggregation of the RAN network are VNF and SDN. Those two technologies often go hand by hand and are enabling the virtualization of network functions and orchestration of virtual instances along the network. The virtualization of RAN functions is very complex since it tackles computationally intensive processes that require computational parallelization in cloud/edge environment. For these reasons, it is critical to enable the processing of RAN functions on the general-purpose hardware. Thereby, RAN virtual instances need to be deployed in cloud platforms that support multi-CPU architectures [50].

From the other side, the complexity of processing centralized virtualized RAN functions depends on the disaggregation point, known as functional split, which defines the splitting point between DU and CU. In the 5G standard, different options are considered and supported, while the decision of the splitting point is left for the network operator, and it is essential since it (i) defines the complexity of the DU and CU, together with reallocation flexibility of virtual resources; (ii) characterizes links between DU and CU with the required amount of transportation data and delays, requirements on fronthaul network; and (iii) limits functionalities gained by grouping together several base stations [51].

To overcome all limitations imposed by making a hard decision of functional split placement, a flexible functional split solution is proposed [44], which requires the presence of software-defined radio (SDR) at the remote sites. However, due to enormous amounts of transportation data between DUs and CUs, especially in cases when DUs are simplified as much as possible, it is necessary to drive improvements in fast fronthaul networks to support such traffic demands with ultra-low delays. This brings the fronthaul network as one of the key enablers for 5G C-RAN deployment [51].

12.4.2 Forecasting DSP Computational Resources Using ML

Mobile users are creating specific spatiotemporal patterns in the utilization of network resources, as it is shown in Fig. 12.6. Furthermore, those patterns are repetitive over time, since underlying purposes of areas where base stations are deployed are not rapidly varying over time, even tending to stay unchanged, while from the other side, social behaviours of mobile users are as well relatively stable over time. This means that traffic loads will impact the mobile network in a repetitive manner, which motivates network operators to manage network resources in such a way to leverage characteristics of users' data traffic demands. For those reasons, network operators are trying to observe network behaviour and to predict it over time. In this section, the focus is on how forecasting RAN functions' computational loads can impact the overall performance of the network within the dynamic 5G C-RAN architecture.

12.4.2.1 Dynamic 5G C-RAN Scenario

As described in Sect. 12.4.1.1 to support dynamic reallocation of RAN resources and to enable network operators to share computational resources, it is necessary to virtualize RAN functions. Once those functions are abstracted and placed into virtual instances, i.e. virtual machines or virtual containers [52], they can be deployed in a dynamic manner in different places of network with possibilities to scale up/down virtual instances depending on the requested traffic (see Fig. 12.1). The flexibility of such action is limited by the following:

- Different functional splits are requiring different functions of a protocol stack to be virtualized, imposing different limitations (e.g. with full PHY functional split, HARQ function in the uplink requires that processing of each frame is finished within 8 ms together with transportation delays) [52]; apart from the protocol point of view, certain functional split options are limiting the flexibility of C-RAN as network infrastructure needs to support high amounts of transportation data, as aforementioned in Sect. 12.4.1.2;
- Depending on the virtualization technique, the dynamicity of reallocation is different, e.g. virtual machines need up to 5 minutes for up/down scaling of resources, while for virtual containers, this is reduced for up to 1 min [53];
- Dynamicity of underlying dataset – a high variation of data traffic at observed base stations means that predictability of such traffic is limited, which further limits performances of reallocation algorithms that rely on data traffic forecasting, while for approaches with reactive resource allocation, it imposes higher margin of reaction – leading to poorer savings.

Static C-RAN approaches are accounting for combing base stations that are behaving in a complementary manner, as described in Sect. 12.4.1, which means that DUs are sharing physical hardware specifically dedicated to baseband processing.

Conversely, with the dynamical sharing of computational resources, the grouping of DUs into CUs can be changed over time and, for that reason, provide higher savings.

An additional possibility for dynamic 5G C-RAN, as mentioned in Sect. 12.4.1.2, is the deployment of SDR RRHs that can also support dynamical functional splits functionalities. As the focus of this discussion is on the reallocation of RAN computational resources within cloud platforms, for further information, the reader is referred to [44, 51].

12.4.2.2 ML Approach for Forecasting DSP Resources

Hereafter, a model to estimate computational resources of RAN functions in a cloud environment is proposed. The model is based on the realistic base station-level dataset and C-RAN architecture, with a fully physical (PHY) functional split. In order to estimate computational loads of RAN functions, due to full PHY functional split related to digital signal processing (DSP) functions, the OpenAirInterface[3] software within a virtual environment with controllable computational resources is employed. In order to simplify the analysis, the focus is only on the computational load of the decoding function. Also, it is the most intensive DSP function in the PHY layer, with the most stringent delay requirements. From the dataset of one European metropolitan city aggregated at 15-min intervals for the set of deployed base stations, modulation and coding schemes (MCS) and physical resource block (PRB) traces are extracted. An example of those traces is given by Fig. 12.7a, from which it can be observed that both time series experience repetitive patterns over a week period, which can be predicted. In Fig. 12.7b, the output performance of OAI software for different pairs (MCS, PRB) is shown, which provides an estimation of the computational loads due to the DSP decoding function based on the number of CPU operations. More details about the experiment setup and results are provided in [38].

Finally, MCS and PRB traces are combined with the OAI results in order to obtain computational load traces, which are further used in the use case for forecasting purposes.

Several simple forecasting algorithms are proposed:

- Last Value (LV) – it is assumed that the next interval value remains the same as the current value.
- Last Day (LD) – similar to LV, but it maps the value of a previous day to the prediction.
- Multiple Linear Regression (MLR) – a simple statistical approach technique that is fitted on the training dataset.
- Multilayer Perceptron Regressor (MLPR) – a simple neural network that approximates function relating input to the output and it is fitted on the training dataset.

[3] OpenAirInterface is an open-source software that provides a full implementation of LTE network, both in the core network and radio access network. https://www.openairinterface.org/

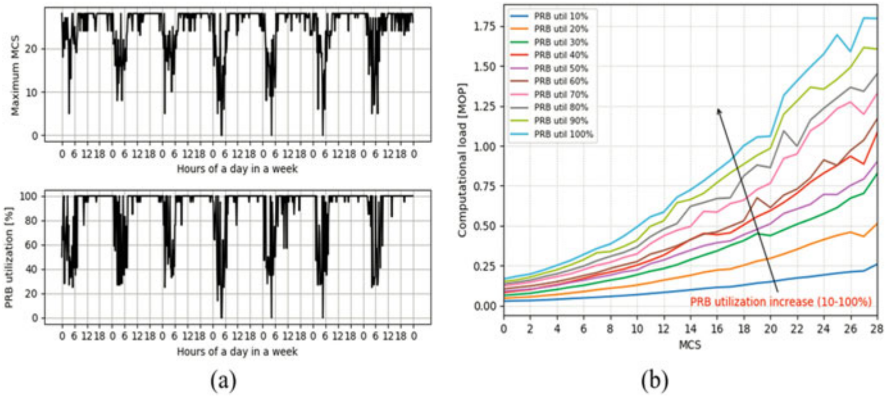


Fig. 12.7 Estimation of computational loads of decoding DSP function: (**a**) an example of MCS and PRB traces for one base station over a full week; (**b**) output from OAI in estimated computational loads of decoding function for each input pair (MCS, PRB)

The proposed algorithms are used for forecasting DSP decoding computational load for a given look-ahead interval, which corresponds to the specific time sample in the future. For instance, if the look-ahead interval is equal to 1 h, the algorithm directly predicts the computational load value at that time interval. Obviously, a higher look-ahead interval leads to greater prediction errors in forecasting, but such information does still provide useful insights of required computational loads for long-term usage. For these reasons, the forecasting algorithms are evaluated in terms of how accurately they can predict future computational loads for different look-ahead intervals, keeping the average and maximum error as low as possible.

Figure 12.8 depicts how the proposed forecasting algorithms are performing for different look-ahead intervals by showing distribution boxplots created over a set of base stations from a realistic dataset. As an error metric, the Root Mean Squared Error (RMSE) is used, while employing six different look-ahead intervals. It is observed that for the lower look-ahead intervals, MLPR and MLR methods are outperforming simpler estimation techniques, while for higher look-ahead intervals, LD value provides more stable results in terms of maximum RMSE. In general, MLPR provides the most stable results in terms of average errors, while maximum RMSE drastically changes with further forecasts. Conversely, the LV method completely fails in predictions of future computational loads for high look-ahead intervals.

12.4.2.3 Performance Evaluation

To demonstrate the effects of computational load forecasting on the allocation of virtualized RAN resources, three benchmarks are introduced:

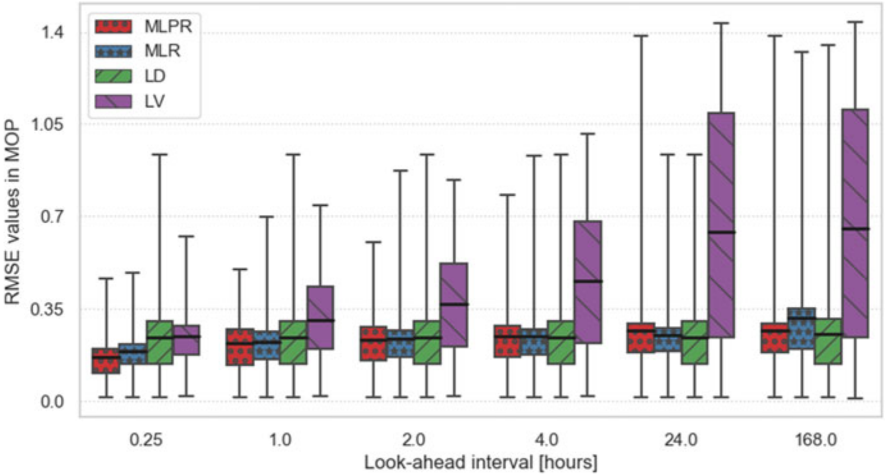


Fig. 12.8 RMSE values of forecastingcomputational loads for different forecasting methods over different look-ahead intervals

- Dynamic C-RAN architecture with Oracle predictor – represents an adoption of Oracle predictor which is able to predict actual traffic perfectly without introducing any additional errors
- Static C-RAN approach – provides allocation of resources within C-RAN architecture to support peak loads at a base station level
- Without C-RAN architecture – considers an implementation in which distributed units are not able to share centralized resources and need to support peak loads

Among the benchmark results, it is noted that dynamic C-RAN with Oracle predictor can save up to 25% of resources compared to static C-RAN and up to 2.5 times more resources compared to the scenario without a deployed centralized architecture. In the light of these results, to evaluate the different forecasting approaches and the maximum possible savings the dynamic C-RAN approach is used as the benchmark; which means that, within the same dynamic resource allocation scenario, the proposed approaches are compared with actual traffic at the base stations or Oracle predictor. To do so, the worst-case scenario of prediction is considered, which means that on predicted traffic value the maximum error of the prediction algorithm is added.

It is intuitive that the forecasting method that provides the lowest maximum errors is outperforming other approaches as it is behaving as close as possible to the Oracle method. The maximum savings in this scenario are 25%, while MLPR is reaching values between 15% and 20% of savings depending on the look-ahead interval. This means that future works should be focused on the provision of ideal predictors in order to increase savings of dynamic reallocation of virtual RAN computational resources. Detailed discussion on results with in-depth insights on this problem is provided by [38].

12.4.2.4 Conclusion

Dynamic allocation of RAN computational resources within C-RAN architecture is one of the promising mechanisms of 5G RAN. With virtualized RAN functions and forecasted computational loads, network resources can be scaled up/down dynamically to follow actual variations in network traffic loads; this is able to promote more efficient utilization of networking infrastructure for network operators and thus promote savings from multiple perspectives, including resource and energy savings. Furthermore, the RANaaS concept provides a platform for network operators for obtaining an optimal deployment of networking resources within a fully virtualized and sharable C-RAN architecture. However, there are still several challenges to be tackled such as improvements of GPP hardware to accommodate computations of virtualized RAN functions in the cloud, further enhancements towards RAN computational load estimation and forecasting and finally in terms of a dynamic functional split within the C-RAN architecture.

12.5 Optimization Framework for MEC Resource Management

In this section, we study resource allocation problems in the context of MEC and try to build an optimization framework for MEC resource management. Specifically, in Sect. 12.5.1, we introduce the background of MEC services for 5G and beyond networks and highlight the context of multiple-edge clouds. Then, in Sect. 12.5.2, we study the optimization of traffic offloading in this context, with the focus on reducing the total latency of multiple types of user traffic. In Sect. 12.5.3, we extend the previous optimization towards a resource calendaring problem for user requests in a more complex environment by considering time domains of requests, network routing, and storage provisioning.

12.5.1 MEC Services for 5G Networks and Beyond

5G networks aim to meet different users' quality of service (QoS) requirements in several different application scenarios and use cases; among others, latency is certainly one of the key QoS requirements that mobile operators have to deal with. In fact, the classification devised by the International Telecommunications Union-Radio Communication Sector (ITU-R), shows that mission-critical services depend on strong latency constraints. For example, in some use cases (e.g. autonomous driving), the tolerable latency is expected to reach less than 1 ms [54]. Multi-access edge computing (MEC) is one of the key enablers for 5G networks, which provides an IT service environment and cloud-computing capabilities at the edge of the mobile network, within the radio access network and in close proximity to mobile subscribers. This approach is foreseen to take a step towards addressing the stringent latency requirements of critical services that still require highly efficient network operation and service delivery. Many works study the resource management problem in the context of MEC [55, 56]. Most of them mainly consider a single MEC system or flat MEC nodes. Notice that the computation power that can be offered by an edge cloud is limited if compared to a remote cloud. Considering that 5G networks will likely be built in an ultradense manner, the edge clouds attached to 5G base stations will also be massively deployed and connected to each other in a specific topology. Thus, exploiting the cooperation among multiple-edge clouds and carefully allocating edge resources to each connection, we can provide a solution to the limitations of a single MEC unit.

12.5.2 MEC for Optimal Traffic Offloading

In this section, we study the case of a complex network organized in multiple-edge clouds, each of which is connected to the radio access network of a certain location. All such edge clouds are connected through various topologies, typically organized in some kind of hierarchy. This way, each edge cloud can serve end user traffic by relying not only on its own resources but also offloading some traffic to its neighbours when needed. We specifically consider multiple classes of traffic and their tolerable latency requirements.

Our main objective is to let the infrastructure serve user traffic within the boundaries of the QoS requirements and the available resources. Specifically, we aim at minimizing the total traffic latency of transmitting, outsourcing and processing user traffic, under a constraint of user tolerable latency for each class of traffic [57]. This optimization turns out to be a mixed-integer nonlinear programming (MINLP) one which is an NP-hard problem [58]. To tackle this challenge, we propose an effective heuristic, named sequential fixing that permits obtaining near-optimal solutions in a very short computing time, even for large-scale scenarios. We also propose a simple greedy approach that obtains slightly suboptimal solutions w.r.t. the sequential fixing approach and still very fast. Finally, we evaluate the impact of the parameters (viz. tolerable latency, computation and link capacity) on the optimal and approximate solutions obtained from our proposed model and heuristics.

12.5.2.1 System Definition

We consider a multiple-edge network composed of *edge points*, denoted by set $\mathcal{E}$. Each edge point $i \in \mathcal{E}$ includes both an *edge mobile network* of a specific capacity C_i and a co-located *edge cloud* of computation capacity S_i. We assume that different users' traffics in the network are aggregated according to their *service types*, denoted by set $\mathcal{N}$. The service requirement for each type of traffic $n \in \mathcal{N}$ is defined by the tolerable latency τ_n for serving the total traffic rate λ_n. Different slices of the mobile network capacity C_i and edge cloud computation capacity S_i can be allocated to serve different types of user traffic based on the requirement. Each link $l \in \mathcal{L}$ (set of links) between different edge points has a fixed bandwidth, denoted by B_l.

The model is defined from the perspective of one *user ingress point* that is receiving user traffic and, when needed, offloads it to the other edge clouds. The same way can be applied to any edge points when they receive traffic. We assume that all types of traffic can be split and processed on all edge clouds. The available network and computation capacity of each edge network are known, for example, through broadcast messages exchanged in the network. The bandwidth of each link in the network can also be estimated through periodic measurements. However, in real scenarios, these network parameters may change dynamically. We cope with this problem by considering a time-slotted system, where time is divided into equal-length slots (short periods where network parameters can be considered as fixed), and we study the optimal allocation of network and computation resources inside such slots to minimize the total latency of all traffic under tolerable latency constraints for each type of traffic. The same optimization process can be repeated for consecutive slots. This latency is the sum of the *wireless network latency* in the user ingress point and the *outsourcing latency* which, in turn, is composed of the *processing latency* in the edge clouds and the *link latency* between edge clouds.

12.5.2.2 Optimization Versus Heuristics

Wireless Network Latency We model the transmission of traffic in user ingress point as an *M*/*M*/1 processing queue. The *wireless network latency* for transmitting the user traffic of type n, denoted by t_n^W, can therefore be computed as:

$$t_n^W = \frac{1}{c_n - \lambda_n}, \forall n \in \mathcal{N}, \tag{12.1}$$

where c_n is the capacity of the network slice for traffic type n in the ingress edge network and $c_n > \lambda_n, \forall n \in \mathcal{N}$. To ensure that the capacity of all slices does not exceed the total capacity C_i of the ingress edge network, we have the constraint $\sum_{n \in \mathcal{N}} c_n \leq C_i$.

Processing Latency We denote with $\alpha_{n,\,i}$ the percentage of type n traffic processed on edge node i and with $\beta_{n,\,i}$ the percentage of computation capacity S_i sliced for traffic n. The processing of user traffic is described by an *M*/*M*/1 model. Let $t_{n,i}^P$ denote the processing latency of edge cloud i for user traffic n and $\forall\,(n, i) \in \mathcal{N} \times \mathcal{E}$; it can be expressed as:

$$t_{n,i}^P = \begin{cases} \frac{1}{\beta_{n,i} S_i - \alpha_{n,i} \lambda_n} & if\ \alpha_{n,i} > 0, \\ 0, & otherwise. \end{cases} \tag{12.2}$$

In the above equation, if $\alpha_{n,\,i} > 0$, we should have $\alpha_{n,\,i}\lambda_n < \beta_{n,\,i}S_i$, otherwise $\alpha_{n,\,i} = \beta_{n,\,i} = 0$, which means that when traffic n is not processed on edge cloud i (i.e. $\alpha_{n,\,i} = 0$), the latency is set to 0 and no computation resource of i should be sliced to n (i.e. $\beta_{n,\,i} = 0$). At the same time, the consistency constraints are $\sum_{i\in\mathcal{E}}\alpha_{n,i} = 1, \forall n \in \mathcal{N}$ and $\sum_{n\in\mathcal{N}}\beta_{n,i} \le 1, \forall i \in \mathcal{E}$.

Link Latency We denote with $\mathcal{L}_i$ the set of links in the shortest path from the user to edge cloud i. When each type of traffic is outsourced from user to i, the transmission latency depends on $B_l, \forall l \in \mathcal{L}_i$, as well as on the total traffic passing through these links, denoted by F_l. To compute F_l, we first write the volume of traffic which flows to an edge network j as $\sum_{n\in\mathcal{N}}\alpha_{n,j}\lambda_n$. Then, $F_l = \sum_{j\in\mathcal{E}, if\ l\in\mathcal{L}_j}\sum_{n\in\mathcal{N}}\alpha_{n,j}\lambda_n, \forall l \in \mathcal{L} = \bigcup_{i\in\mathcal{E}}\mathcal{L}_i$. Let $t^L_{n,i}$ denote the link latency. When i is the user ingress point, denoted by *ingress*, $\mathcal{L}_i =$ and ${}^{L}_{n,i} = 0$. $\forall\,(n, i) \in \mathcal{N} \times (\mathcal{E} - \{ingress\})$, $t^L_{n,i}$ is defined as:

$$t^L_{n,i} = \begin{cases} \sum_{l\in\mathcal{L}_i} \dfrac{1}{B_l - \sum_{j\in\mathcal{E}, if\ l\in\mathcal{L}_j}\sum_{n'\in\mathcal{N}}\alpha_{n',j}\lambda_{n'}}, & if\ \alpha_{n,i} > 0, \\ 0, & otherwise. \end{cases} \tag{12.3}$$

The link latency is counted only if a certain traffic segment is processed on i. The constraint for F_l is written as $F_l < B_l, \forall l \in \mathcal{L}$.

To define the optimization problem, we define the longest serving time among edge clouds as the *outsourcing latency* for traffic n, i.e. $t^{PL}_n = \max\limits_{i\in\mathcal{E}}\left\{t^P_{n,i} + t^L_{n,i}\right\}, \forall n \in \mathcal{N}$. Then, the constraint for tolerable latency requirement is $t^W_n + t^{PL}_n \le \tau_n, \forall n \in \mathcal{N}$. Finally, the objective function is:

$$\mathcal{P}0: \min_{c_n,\alpha_{n,i},\beta_{n,i}} \sum\nolimits_{n\in\mathcal{N}}\left\{t^W_n + t^{PL}_n\right\}. \tag{12.4}$$

Problem $\mathcal{P}0$ contains both nonlinear and indicator constraints; therefore, it is a mixed-integer nonlinear programming (MINLP) problem, which is hard to be solved directly [58]. Hereafter, we propose two effective heuristics to solve this problem.

- ***Sequential Fixing (SF):*** This heuristic, detailed in Algorithm 12.1, is based on the following rationale: first, it solves a relaxation of the original problem (see line 1, where $b_{n,\,i}$ is a binary variable that indicates whether traffic n is processed on node i). Then, based on the relaxed solution $\tilde{b}_n, i^*$, which can be seen as the probability of node selection, edge nodes are ranked and selected (see line 2). We also rank the traffic types according to the rate and corresponding tolerable latency and allocate computing nodes to the different types of traffic via a specific scheme (see lines 3–6). Finally, we eliminate all unfixed $b_{n,\,i}$ to further prune the problem.

Algorithm 12.1 Sequential Fixing

```
1. Relax $b_{n,i}$ to continuous $\tilde{b}_{n,i}$ in $\mathcal{P}1$, then solve the relaxed problem to obtain optimal relaxed values $\tilde{b}^*_{n,i}$;
2. Rank nodes in $\mathcal{E}$ by descending values of $\sum_{n\in\mathcal{N}} \tilde{b}_{n,i}$, and keep top $\mathcal{K} \subset \mathcal{E}$;
3. Rank traffic types in $\mathcal{N}$ in descending order ($\lambda_n$ first, $\tau_n$ second);
4. Allocate nodes in $\mathcal{K}$ to ordered traffic types (i.e., setting $b_{n,i} = 1$) till either $\mathcal{K}$ or $\mathcal{N}$ is completely scanned;
5. If $|\mathcal{K}| < |\mathcal{N}|$ then Rank $\mathcal{K}$ in descending order ($S_i$ first, $|\mathcal{L}_i|$ second), allocate $\mathcal{K}$ to (remaining $\mathcal{N}$) repeatedly;
6. Else Allocate (remaining $\mathcal{K}$) to $\mathcal{N}$ repeatedly;
7. Set the remaining variables $b_{n,i} = 0$.
```

- ***Greedy approach:*** The greedy approach, detailed in Algorithm 12.2, first ranks edge nodes according to the link and computation capacities (see line 1). Then, it ranks the traffic types and allocates edge nodes (see lines 2–4). Finally, it branches the problem like *SF*.

Algorithm 12.2 Greedy approach

```
1. Generate priority order of edge nodes $\mathcal{E}$ ($|\mathcal{L}_i|$ first, $S_i$ second, by rotating the topology graph around the user
   ingress point and doing level traversal of the graph), keeping the top $\mathcal{K}$ nodes;
2. Rank traffic types $\mathcal{N}$ in descending order ($\lambda_n$ first, $\tau_n$ second);
3. Allocate $\mathcal{K}$ to $\mathcal{N}$ ($\forall n \in \mathcal{N},\ amount = \left\lceil \frac{\lambda_n|\mathcal{K}|}{\sum_{n'\in\mathcal{N}}\lambda_{n'}} \right\rceil$) in order, till either $\mathcal{K}$ or $\mathcal{N}$ is completely scanned;
4. While $\mathcal{N}$ has un-allocated elements do Re-allocate $\mathcal{K}$ to (rest of $\mathcal{N}$) one by one repeatedly;
5. Set remaining variables $b_{n,i} = 0$.
```

12.5.2.3 Performance Evaluation and Conclusion

We evaluate and compare the proposed models identified as *Optimal*, *SF*, *Greedy*, and *Random* in terms of different parameters. This last one is used as a reference and is based on randomly selecting nodes (for each instance, we select the best in 1000 iterations). We set up a scenario with high traffic load and low tolerable latency w.r.t. the limited resources. We adopt hierarchical topologies for the network, which are similar to [59]. We underline that our model and heuristics are general and can be applied in any topology with a predetermined routing, which can be performed in a preprocessing phase.

The performance on traffic delay (latency) is observed from different perspectives:

- **Effect of the tolerable latency $\boldsymbol{\tau_n}$:** Fig. 12.9.a shows the total latency variation w.r.t. the tolerable latency $\tau_n, \forall n \in \mathcal{N}$, with scaling factors (from 0.70 to 0.86 w.r.t. the initial value in the scenario). The vertical lines represent the thresholds, showing the minimum τ_n that can be requested. When τ_n increases, the total

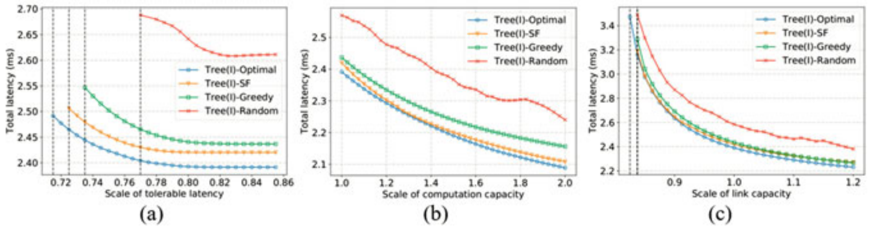


Fig. 12.9 Total latency versus: (**a**) tolerable latency, (**b**) computation capacity, and (**c**) link capacity

latency decreases in all cases and converges to almost the same point around 0.8. τ_n serves in our model as an upper bound and limits the solution space. With a low τ_n, the feasible solution set is smaller, and the total latency increases and vice versa. We observe that the gap between *SF* and *Optimal* is indeed small, around 1.43%.

- **Effect of the computation capacity $\mathbf{S_i}$:** Fig. 12.9.b shows the total latency w.r.t. the computation capacity $S_i, \forall i \in \mathcal{E}$, which is scaled from 1 to 2. The total latency decreases when S_i increases, e.g. as S_i increases by a factor of two, the latency decreases of about 0.3 ms. *Random* performs the worst, *SF* is very close to *Optimal* and practically overlapping in range (1.2, 1.5). *Greedy* also performs well, with a gap that reaches 1.76% w.r.t. *Optimal.*
- **Effect of the link capacity $\mathbf{B_l}$:** Fig. 12.9.c illustrates the total latency variation w.r.t. the link capacity $B_l, \forall l \in \mathcal{L}$, which is scaled from 0.82 to 1.2. As B_l increases, the total latency decreases and converges to different values (around 2.27 for *Optimal*, *SF*, *Greedy*, and around 2.38 for *Random*). *SF* performs very close to *Optimal* and overlaps with it in range (0.838, 0.92). *Greedy* also exhibits a good performance, with a gap of 1.44% w.r.t. *Optimal.*
- **Computing time:** Obtaining the *Optimal* results takes a very long time (around 1000 seconds for the smallest instances with $\mid \mathcal{N} \mid= 3$, while it was impossible for larger $\mid \mathcal{N} \mid$ values). Our approaches are fast in the considered network instances. With $\mid \mathcal{N} \mid= 3$, the solving time of *SF* is around 0.1 s and *Greedy* 0.2 s. When $|\mathcal{N}| = 15$, the solving time is still less than 1 s.

We proposed a novel mathematical model to perform a joint allocation of mobile network and edge computational resources to minimize the total latency of multiple classes of user traffic with their tolerable delay requirements in hierarchical multiple-edge networks. Two heuristics, sequential fixing and greedy, are proposed to solve this model. We evaluated the effects of the parameters on the optimal and approximate solutions. Results show that SF can obtain near-optimal solutions in a very short computing time, even for large-scale scenarios.

12.5.3 Towards Resource Calendaring for MEC

In Sect. 12.5.2, we studied the resource allocation approaches aimed at reducing the total latency experienced by mobile users. Here we study the resource calendaring problem for MEC by considering the time domain in the network. We provide an optimization framework that considers several key aspects of the resource allocation problem in the context of MEC [60]. We further extend the previous sequential fixing heuristic (Sect. 12.5.2.2) to solve the problem. Specifically, our proposed model and heuristics jointly optimize (1) *admission decision* (where connections are admitted and served by the network, based on the profit they can potentially generate with respect to the required resources for serving demands), (2) s*cheduling* of admitted connections, also called *calendaring* (taking into account the flexibility that some users exhibit in terms of starting and ending time tolerated for the required services), (3) *routing* of these flows, (4) the decision of which nodes will serve such connections as well as (5) the amount of processing and storage capacity reserved on the chosen nodes that serve such connections, with the objective of maximizing the operator's profit.

To the best of our knowledge, our work is the first one that considers all these five aspects together. Other works focus, instead, on specific aspects. For instance, the work in [61] considers task assignment, computing and transmission resources allocation to minimize system latency in a multi-layer MEC. They do not consider the resource scheduling problem. [62] studies online deadline-aware task dispatching and scheduling to maximize the number of completed tasks. They do not explicitly consider the routing problem that arises.

12.5.3.1 The Resource Calendaring Problem

We consider an edge cloud network represented by an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $v \in \mathcal{V}$ represents an edge *computing node* having D_v and S_v as *computation* and *storage capacity*, respectively. Parameters θ_v and ϕ_v denote, respectively, the *cost* of computation and storage capacity of v. Each *edge* $e \in \mathcal{E}$ corresponds to a network link characterized by its *bandwidth* B_e and its *cost per unit of flow* ψ_e. A set of *requests* (each requiring bandwidth, storage and computation resources), denoted by $\mathcal{K}$, is offered to the network and has to be accommodated. We assume that the arriving time and duration of the requests for the upcoming period are known. This can be achieved assuming that customers have announced their requirements in advance or that some history-based prediction tools [63] are used.

We discretize the time horizon into a set $\mathcal{T}$ of equal duration time slots, where the *slot length* is τ. Each request $k \in \mathcal{K}$ is defined as a tuple $(s^k, \alpha^k, \beta^k, d^k, \lambda^k, \mu^k)$. The parameter s^k is the *source node* of request k; α^k, β^k and d^k define the *arrival time*, the *latest ending time* (deadline) and the *duration* of request k, respectively. Finally, we consider a Poisson process for each request k arrival with an average rate λ^k during the period d^k, and μ^k is the revenue gained from serving request k. A request k has processing density [64] η^k and also requires a fixed amount of storage resource m^k on a node if k is to be processed on it.

A request k could be processed immediately (for delay-sensitive tasks) after its arrival or scheduled later (for delay-tolerant tasks). Also, it could be entirely processed on the local edge computing node or also on other nodes. In any case, it must be completed before the deadline β^k. Given a *calendar of requests* $\mathcal{T}$ over a time horizon, the proposed optimization approach (a) schedules the starting time of each request, (b) decides where to compute the requests, and (c) route some fractions of requests when it is necessary to process them on other edge nodes.

We formulate this problem from the following perspectives: *the life cycle of a request, network routing, link latency, processing latency and storage provisioning.* The modelling of link latency and processing latency is similar to Sect. 12.5.2.2. To model the *life cycle of a request*, we first formulate the admission controller and compute the starting time of a request, while the ending time depends on the link and processing latency. To model the *network routing*, we assume that a request can be split into multiple pieces only at its source node. Each piece can be offloaded to another edge computing node independently of the other pieces, but it cannot be further split. Finally, modelling the *storage provisioning* is close to the way formulating the processing capacity constraint. Due to the space limit, we do not present the formulations of all these components; instead we refer the interested readers to [60] for further details.

Finally, our goal is to maximize the profit represented by the total revenue obtained from serving the users' requests minus the network operation costs in terms of computation, storage and bandwidth resources under the constraints of requests and resources:

$$\mathcal{P}0 : \max \sum_{t\in\mathcal{T}} \sum_{k\in\mathcal{K}} \left\{ \mu^k z^{kt} - \sum_{v\in\mathcal{V}} \left\{ r^{kvt} D_v \theta_v + \rho^{kvt} m^k \phi_v + \sum_{e\in\mathcal{E}} p_e^{kvt} B_e \psi_e \right\} \right\} \tag{12.5}$$

where variable z^{kt} decides whether request k is scheduled at time point t, r^{kvt} is the percentage of node v 's computation capacity allocated for k at t, ρ^{kvt} is whether node v is processing k at t and p_e^{kvt} is the percentage of link e 's bandwidth allocated for a piece of request k at t. Problem $\mathcal{P}0$ contains both nonlinear and indicator constraints; therefore, it is a mixed-integer nonlinear programming (MINLP) problem, which is hard to be directly solved [58]. Moreover, we also face the following difficulties: (a) routing and request fraction variables are "intertwined"; (b) indicator functions and constraints cannot be directly processed by most solvers. To deal with the above issues, we propose an equivalent reformulation of $\mathcal{P}0$, which we call $\mathcal{P}1$ that we can efficiently solve with the branch and bound method. For space reasons, we do not include it here. The interested readers can refer to our technical report.[4]

[4] http://xiang.faculty.polimi.it/files/TechnicalReport.pdf

12.5.3.2 Optimization Framework and Performance Evaluation

We propose a heuristic named *sequential fixing and scheduling (SFS)*, which realizes a good trade-off between admitting "valuable" connections (that provide high return to the service provider) and the resources they request w.r.t. transmission rate, storage and computation. SFS is detailed in Algorithm 12.3.

Algorithm 12.3 Sequential Fixing and Scheduling

```
1.  Sort $\mathcal{K}$ in descending order by the ratio $\mu^k/(d^k\lambda^k m^k\eta^k)$;
2.  For $k \in \mathcal{K}$ do
3.    Find candidates $\mathcal{Q}^k$ for computing request $k$;
4.    If $\mathcal{Q}^k \neq \emptyset$ then
5.      Update residual bandwidth $B'_e$ based on the link status; create graph $\mathcal{G}'$ weighted by $B'^{-1}_e$;
6.      For $\mathcal{V}_i \in \mathcal{Q}^k$ do
7.        Set $b^{kv} = 1,\ \forall v \in \mathcal{V}_i$; fix route using Dijkstra; optimize $\mathcal{P}1$ to get profit $\mathcal{O}$ and solution $\mathcal{S}$;
8.        If $\mathcal{O} \geq 0$ then break;
9.    If $\mathcal{O} \geq \mathcal{O}^\star$ and $\mathcal{Q}^k \neq \emptyset$ then
10.     Update $\mathcal{O}^\star \leftarrow \mathcal{O},\ \mathcal{S}^\star \leftarrow \mathcal{S}$; admit $k$ and allocate resource based on $\mathcal{S}^\star$; set $\mathcal{P}1$'s lower bound $LB = \mathcal{O}^\star$;
11.   Else Reject $k$;
```

We start by sorting all requests in descending order by $\mu^k/(d^k\lambda^k m^k\eta^k)$ to give a higher weight to requests that generate more revenue and less cost. Then, we try to find candidates $\mathcal{Q}^k$ (ordered set) which contains sets $\mathcal{V}_i$ of best candidate nodes to process request k. If $\mathcal{Q}^k \neq \varnothing$, we further update the residual bandwidth B'_e and create a weighted graph $\mathcal{G}'$ with $B\prime^{-1}_e$. Then (Algorithm 12.3: lines 6–8), we select the first $\mathcal{V}_i$ in $\mathcal{Q}^k$ that allows us to find a profitable solution ($\mathcal{O} \geq 0$) according to the following criteria: we outsource k to the nodes in $\mathcal{V}_i$ by setting b^{kv} (whether k is to be processed on v). Based on $\mathcal{G}'$, we route each piece of request using the *Dijkstra* algorithm. Then, we optimize $\mathcal{P}1$ to get the profit and the solution denoted, respectively, by $\mathcal{O}$ and $\mathcal{S}$. If $\mathcal{P}1$ results infeasible ($\mathcal{O} < 0$), we reiterate on the other elements of $\mathcal{Q}^k$. If the new optimization improves, we update the best profit $\mathcal{O}^\star$ and solution $\mathcal{S}^\star$; we hence admit request k and allocate resources to it (including time slots, computation, bandwidth and storage). We also update the lower bound of $\mathcal{P}1$ to $LB = \mathcal{O}^\star$ to accelerate the optimization. Finally, if the result does not improve or no candidate is found, we reject k.

The network topologies are generated based on Erdös-Rényi random graph. We obtain a representative large topology (*30N50E30R*) having 30 nodes and 50 edges with 30 requests and a small one (*5N5E3R*) having 5 nodes and 5 edges with 3 requests. Requests with different settings (starting and ending time, revenue, etc.) are randomly generated. Note that our model and heuristics are general and can be applied to all network scenarios with any parameters setting. The computing times for all approaches versus different topologies are shown in Table 12.1. $\mathcal{P}1$ could be solved in a reasonable time only in the small topology (*5N5E3R*), which has an average time 146 s, while *SFS* took just 5 s and *greedy* 4 s. In a larger scenario

Table 12.1 Computing time for all approaches versus different topologies

	5N5E3R	30N50E30R
Optimal	146 s	–
SFS	5 s	1096 s
Greedy	4 s	822 s

(30N50E30R), *SFS* exhibited a computing time inferior to 1096s. The *greedy* needs less time, on average 822 s, at the cost of higher performance gaps with *SFS*.

The effects of request rate and revenue as well as computation capacity on the solutions obtained by different approaches are evaluated. In all cases, *SFS* exhibits better performance compared to *greedy* with clear gaps, and due to space limit, the results are available in our technical report.[4]

12.5.3.3 Challenges for Future Work and Conclusion

We formulated and solved the resource calendaring problem in mobile networks equipped with multi-access edge computing (MEC) capabilities. Specifically, we proposed both an exact optimization model as well as an effective heuristic able to obtain near-optimal solution in all the considered, real-size network scenarios. The decisions we optimized include admission control for the requests, calendaring (scheduling) and bandwidth-constrained routing, as well as the determination of which nodes provide the required computation and storage capacity. Calendaring, in particular, permits to exploit the intrinsic flexibility in the services demanded by different users, whose starting time can be shifted without penalizing the utility perceived by the user while, at the same time, permitting a better resource utilization. For large network scenarios, the resource calendaring problem becomes very hard to be solved in a reasonable time. In the future, we plan to decompose the problem and solve it in a distributed manner.

12.6 Conclusion

In this chapter, we summarize key aspects of network resource virtualization and its placement within edge/cloud infrastructure, in order to provide an overview of cost-effective resource management of virtualized instances within 5G-enabled MEC architectures. Moreover, through the applications of C-V2X and UDNs, we evaluate the efficiency of such an architecture in the interplay between core and edge of the network.

We demonstrated the live migration and scaling features of our architecture using a 5G experimental platform, and it was observed that significantly more requests could be supported without SLA violation by providing savings in CPU utilization in contrast to more rigid baseline approaches. In this context, AaaSFC appears to be a good candidate for cloud resource management in legacy and future emerging 6G mobile networks that will be predominantly cloud-based. The optimal integration and collaboration of MEC units in UDNs is introduced as a basis towards the proactive allocation of services harnessing on a distributed MEC ecosystem and the inherent features of present 5G standards. In the example of RNNs, we demonstrated how the gain improvements in available storage by applying predictions of mobility services. Moreover, we highlight how feedback from big data analytics can drive the network deployment and, in general, improve utilization of network infrastructure, especially for 5G RAN centralized architecture. By usage of realistic city-wide mobile network datasets, we investigated influences of users' behavioural patterns on the mobile network utilization and deployment of C-RAN architecture. We also provided a novel approach for forecasting computational requirements of virtualized RAN functions, which is able to ensure up to 25% of cost savings. Furthermore, we introduce an optimization framework for MEC resource management. We provided three algorithms to solve the resource management problem in MEC scenario by investigating in details the influence of each algorithm on performances. The created heuristic approaches were compared to the optimal solutions and showed a near-optimal behaviour. Also, we formulated and solved a resource calendaring problem for the allocation of computing resources in MEC scenario. Finally, we point out the challenges in the investigated areas, which can motivate future work and further improvements on these topics.

We can conclude that cloudification of mobile network services is one of the key points in 5G and beyond networks, and for this reason, increasingly more attention will be directed towards enabling the high dynamical reallocation of resources within cloudified network and on the migration of virtualized network resources and services within the network, while ensuring that QoS of users is not decreased. It is expected the approaches will provide a basis for network operators to obtain a cost-effective mobile networking infrastructure, while taking a step towards a green economy.

Acknowledgements This research work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 722788 (SPOTLIGHT).

References

1. ETSI White Paper No. 28, "MEC in 5G networks", 2018.
2. ETSI, "The standard news from ETSI", Issue 2, 2017, https://www.etsi.org/images/files/ETSInewsletter/etsinewsletter-issue2-2017.pdf
3. Sarrigiannis, I., Ramantas, K., Kartsakli, E., Mekikis, P.-V., Antonopoulos, A., & Verikoukis, C. (2020). Online VNF lifecycle management in a MEC-enabled 5G IoT architecture. *IEEE Internet of Things Journal, 7*(5), 4183–4194.
4. Kreutz, D., Ramos, F. M. V., Veríssimo, P. E., Rothenberg, C. E., Azodolmolky, S., & Uhlig, S. (2015). Software-defined networking: A comprehensive survey. *Proceedings of the IEEE, 103*(1), 14–76.

5. ETSI NFV. (2014). Network functions virtualisation (NFV); management and orchestration. *GS NFV-MAN, 001*.
6. Gonzalez, A. J., Nencioni, G., Kamisiński, A., Helvik, B. E., & Heegaard, P. E. (2018). Dependability of the NFV orchestrator: State of the art and research challenges. *IEEE Communications Surveys & Tutorials*, 99.
7. "Release 14 Description", 3GPP TR 21.914, Tech. Rep., 2018.
8. Consortium Architecture Working Group, "OpenFog reference architecture for fog computing", Feb. 2017.
9. Sarrigiannis, I., Contreras, L-M., Ramantas, K., Antonopoulos, A., & Verikoukis, C. Fog-enabled Scalable C-V2X Architecture for Distributed 5G and Beyond Applications. *IEEE Network*, pending publication.
10. The Openstack foundation, https://openstack.org/. Accessed 10 Aug 2020.
11. Open Source Mano, https://osm.etsi.org/. Accessed 10 Aug 2020.
12. Cisco, Cisco Annual Internet Report (2018–2023) White Paper. Online: shorturl.at/cNQTZ. Last read 12/08/2020.
13. System Architecture for the 5G System (5GS). 3rd Generation Partnership Project (3GPP). TS 23.501. July 2020.
14. Procedures for the 5G System (5GS). 3rd Generation Partnership Project (3GPP). TS 23.502. July 2020.
15. IETF draft: https://tools.ietf.org/html/draft-fattore-dmm-n6-cpdp-trafficsteering-01. 2019. Internet Engineering Task Force (IETF) Distributed Mobility Management (DMM) WG.
16. Sakshi, C., & Sivalingam, K. M. (2015). SDN based evolved packet core architecture for efficient user mobility support. In *Proceedings of the 2015 1st IEEE conference on network softwarization (NetSoft)*. IEEE.
17. Xin, J. et al. (2013). Softcell: Scalable and flexible cellular core network architecture. In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*.
18. Fattore, U., Giust, F., & Liebsch, M. (2018). 5GC+: An experimental proof of a programmable mobile core for 5G. IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD).
19. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modelling. arXiv preprint arXiv:1412.3555 (2014).
20. De Vita, F., Nardini, G., Virdis, A., Bruneo, D., Puliafito, A., & Stea, G. (2019). Using deep reinforcement learning for application relocation in multi-access edge computing. *IEEE Communications Standards Magazine, 3*(3), 71–78.
21. ETSI. (2019). Multi-access Edge Computing (MEC); Framework and Reference Architecture. TS. European Telecommunication Standards Institute (ETSI).
22. Fattore, U., Liebsch, M., & Bernardos, C. J. (2018). UPFlight: An enabler for avionic MEC in a drone-extended 5G mobile network. IEEE 23rd Computer Aided Modeling and Design of Communication Links and Networks (CAMAD).
23. Fattore, U., Liebsch, M., Brik, B., & Ksentini, A. (2020). AutoMEC: LSTM-based user mobility prediction for service management in distributed MEC resources. *Proceedings of the 22nd international ACM conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*.
24. Gambs, S., Killijian, M-O., & del Prado Cortez, M. N. (2012). Next place prediction using mobility markov chains. In *Proceedings of the first workshop on measurement, privacy, and mobility*, pp. 1–6.
25. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.
26. Taleb, T., & Ksentini, A. (2013). Follow me cloud: interworking federated clouds and distributed mobile networks. *IEEE Network, 27*(5), 12–19.
27. Liu, Q., Wu, S., Liang, W., & Tan, T. (2016). Predicting the next location: A recurrent model with spatial and temporal contexts. In *Thirtieth AAAI conference on artificial intelligence*.

28. Ntalampiras, S., & Fiore, M. (2018). Forecasting mobile service demands for anticipatory MEC. In *2018 IEEE 19th international symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)* (pp. 14–19). IEEE.
29. Tkačík, J., & Kordík, P. (2016). Neural turing machine for sequential learning of human mobility patterns. In *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 2790–2797). IEEE.
30. Andrew, M. et al. (2016). Migrating running applications across mobile edge clouds: Poster. In *Proceedings of the 22nd annual international conference on mobile computing and networking*.
31. Ventre, P. L., Lungaroni, P., Siracusano, G., Pisa, C., Schmidt, F., Lombardo, F., & Salsano, S. (2018). On the fly orchestration of unikernels: Tuning and performance evaluation of virtual infrastructure managers. *IEEE Transactions on Cloud Computing, 2018.*
32. Wang, C., Zhao, Z., Sun, Q., & Zhang, H. (2018). Deep learning based intelligent dual connectivity for mobility management in dense network. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)* (pp. 1–5). IEEE.
33. Wickramasuriya, D. S., Perumalla, C. A., Davaslioglu, K., & Gitlin, R. D. Base station prediction and proactive mobility management in virtual cells using recurrent neural networks. In *2017 IEEE 18th Wireless and Microwave Technology Conference (WAMICON)* (pp. 1–6). IEEE.
34. Chen, K., & Duan, R. (2011). *C-RAN the road towards green RAN*. China Mobile Research Institute, white paper 2.
35. Checko, A., Christiansen, H. L., Yan, Y., Scolari, L., Kardaras, G., Berger, M. S., & Dittmann, L. (2014). Cloud RAN for mobile networks—A technology overview. *IEEE Communications surveys & tutorials, 17*(1), 405–426.
36. Agiwal, M., Roy, A., & Saxena, N. (2016). Next generation 5G wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials, 18*(3), 1617–1655.
37. Sexton, C., Kaminski, N. J., Marquez-Barja, J. M., Marchetti, N., & DaSilva, L. A. (2017). 5G: Adaptable networks enabled by versatile radio access technologies. *IEEE Communications Surveys & Tutorials, 19*(2), 688–720.
38. Okic, A., & Redondi, A. E. C. (2020) Optimal resource allocation in C-RAN through DSP computational load forecasting. In *2019 IEEE 30th annual international symposium on Personal, Indoor and Mobile Radio Communications (PIMRC).*
39. Rost, P., Talarico, S., & Valenti, M. C. (2015). The complexity–rate tradeoff of centralized radio access networks. *IEEE Transactions on Wireless Communications, 14*(11), 6164–6176.
40. Okic, A., Redondi, A. E., Galimberti, I., Foglia, F., & Venturini, L. (2019). Analyzing different mobile applications in time and space: A city-wide scenario. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)* (pp. 1–6). IEEE.
41. Okic, A., & Redondi, A. E. (2019). Forecasting Mobile cellular traffic sampled at different frequencies. In *2019 12th IFIP Wireless and Mobile Networking Conference (WMNC)* (pp. 189–195). IEEE.
42. Furno, A., Fiore, M., Stanica, R., Ziemlicki, C., & Smoreda, Z. (2016). A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. *IEEE Transactions on Mobile Computing, 16*(10), 2682–2696.
43. Checko, A., Christiansen, H. L., & Berger, M. S. (2013). Evaluation of energy and cost savings in mobile Cloud RAN. In *OPNETWORK 2013*. OPNET.
44. Alba, A. M., Velásquez, J. H. G., & Kellerer, W. (2019). An adaptive functional split in 5G networks. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 410–416). IEEE.
45. Bhaumik, S., Chandrabose, S. P., Jataprolu, M. K., Kumar, G., Muralidhar, A., Polakos, P., ... & Woo, T. (2012). CloudIQ: A framework for processing base stations in a data center. In *Proceedings of the 18th annual international conference on Mobile computing and networking* (pp. 125–136).

46. Chen, L., et al. (2017). Complementary base station clustering for cost-effective and energy-efficient cloud-RAN. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE.
47. Sabella, D., Rost, P., Sheng, Y., Pateromichelakis, E., Salim, U., Guitton-Ouhamou, P., Di Girolamo, M., & Giuliani, G. (2013). RAN as a service: Challenges of designing a flexible RAN architecture in a cloud-based heterogeneous mobile network. In *2013 future network & mobile summit* (pp. 1–8). IEEE.
48. Sabella, D., De Domenico, A., Katranaras, E., Imran, M. A., Di Girolamo, M., Salim, U., Lalam, M., Samdanis, K., & Maeder, A. (2014). Energy efficiency benefits of RAN-as-a-service concept for a cloud-based 5G mobile network infrastructure. *IEEE Access, 2*, 1586–1597.
49. Santos, J. F., Kist, M., Rochol, J., & Da Silva, L. A. (2020). *Virtual radios, real services: Enabling RANaaS through radio virtualisation*. IEEE Transactions on Network and Service Management.
50. Wubben, D., Rost, P., Bartelt, J. S., Lalam, M., Savin, V., Gorgoglione, M., ... Fettweis, G. (2014). Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN. *IEEE Signal Processing Magazine, 31*(6), 35–44.
51. Larsen, L. M., Checko, A., & Christiansen, H. L. (2018). A survey of the functional splits proposed for 5G mobile crosshaul networks. *IEEE Communications Surveys & Tutorials, 21*(1), 146–172.
52. Nikaein, N. (2015). Processing radio access network functions in the cloud: Critical issues and modeling. In *Proceedings of the 6th international workshop on Mobile cloud computing and services* (pp. 36–43).
53. Jindal, A., Podolskiy, V., & Gerndt, M. (2017). Multilayered cloud applications autoscaling performance estimation. In *2017 IEEE 7th international symposium on cloud and service computing (SC2)* (pp. 24–31). IEEE.
54. Xiang, W., Zheng, K., & Shen, X. S. (2017). *5G mobile communications*. Springer.
55. Lyu, X., Tian, H., Ni, W., Zhang, Y., Zhang, P., & Liu, R. P. (2018). Energy-efficient admission of delay-sensitive tasks for Mobile edge computing. *IEEE Transactions on Communications, 66*(6), 2603–2616.
56. Cheng, K., Teng, Y., Sun, W., Liu, A., & Wang, X. (2018). *Energy-efficient joint offloading and wireless resource allocation strategy in multi-mec server systems*. IEEE ICC.
57. Xiang, B., Elias, J., Martignon, F., & Di Nitto, E. (2019). *Joint network slicing and mobile edge computing in 5G networks*. IEEE ICC.
58. Kannan, R., & Monma, C. L. (1978). On the computational complexity of integer programming problems. In *Optimization and operations research* (pp. 161–172). Springer.
59. Tong, L., Li, Y., & Gao, W. (2016). *A hierarchical edge cloud architecture for mobile computing*. IEEE INFOCOM.
60. Xiang, B., Elias, J., Martignon, F., & Di Nitto, E. (2020). *Resource calendaring for mobile edge computing in 5G networks*. Submitted to IEEE ICC.
61. Wang, P., Zheng, Z., Di, B., & Song, L. (2019). *Joint task assignment and resource allocation in the heterogeneous multi-layer mobile edge computing networks*. IEEE Globecom.
62. Meng, J., Tan, H., Li, X.-Y., Han, Z., & Li, B. (2020). Online deadline-aware task dispatching and scheduling in edge computing. *IEEE Transactions on Parallel and Distributed Systems, 31*(6), 1270–1286.
63. Hong, C.-Y., Kandula, S., Mahajan, R., Zhang, M., Gill, V., Nanduri, M., & Wattenhofer, R. (2013). Achieving high utilization with software-driven wan. *ACM SIGCOMM Computer Communication Review, 43*(4), 15–26.
64. Kwak, J., Kim, Y., Lee, J., & Chong, S. (2015). DREAM: Dynamic resource and task allocation for energy minimization in mobile cloud systems. *IEEE Journal on Selected Areas in Communications, 33*(12), 2510–2523.

Chapter 13
Demonstrating Cloud-Based Services for UDNs: Content Distribution Case Study

Roberto Torre, Sarah Irum, Riccardo Bassoli, Gerrit Schulte, and Frank H. P. Fitzek

Abstract This chapter will discuss a framework for testing SDN-based services for 5G and beyond and will showcase how content distribution can be effectively implemented within an SDN cooperative-based ecosystem as a case study. The framework is engineered for testing cloud-based services in a virtual MSC (mobile small cell) environment. We showcase the testbed with an eMBB use case, namely, massive content distribution over cellular networks. We not only aim to show capability of the testbed but also provide tangible evidence as to how content distribution can be effectively enhanced using network-coded cooperation (NCC) in synergy with virtual mobile small cells in terms of reliability and energy consumption in contrast to legacy unicast approaches.

13.1 Introduction

We are now knocking on the doors of the next generation of mobile communications referred to as 5G that is based on key enabling technologies such as massive MIMOs, mmWave backhauling, and densifications of nodes such as ultradense networks (UDNs) to cater for delivering enhanced broadband connectivity and massive scale communications based on the IoT (Internet of Things) paradigm. These were driven by the market stakeholders, where due diligence had forecasted that mobile data services would be subject to explosive growth coupled with an increase in connected devices, e.g., the *m*obile and wireless communication *e*nablers for *t*wenty-twenty (2020) *i*nformation *s*ociety (METIS) project predicted that by 2020, worldwide mobile traffic alone will increase by 33 times to that of the 2010 figures, and moreover by 2022, Cisco reported [1] that 82% of that traffic will

R. Torre (✉) · R. Bassoli · F. H. P. Fitzek
Technische Universität Dresden, Dresden, Germany
e-mail: roberto.torre@tu-dresden.de

S. Irum · G. Schulte
Acticom, Berlin, Germany
e-mail: sarah.irum@acticom.de

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_13

consist of video traffic and the number of connected devices will increase up to more than 18 billion devices. During this time, the Internet of Things (IoTs) will become dominated by massive wireless devices such as smartphones, tablets, machines, and sensors.

Going beyond 5G deployment, the market will continue to grow, and new delay stringent application will emerge that will require the 5G architecture to evolve. Even though several new technologies are emerging as part of the 6G story such as terahertz and visible light communications, it is clear that legacy technology enablers will continue to evolve such as softwarization in networks and small cell technology. Softwarization introduces massive flexibility into managing networks effectively, where the future envisages a complete virtualization of the network that includes mobile devices acting as an additional pool of networking resources. While small cell technology will evolve to become ultrafast hyper dense networks to support ultrahigh spends. The question arises as to how we can exploit UDNs and virtualization for enhancing the delivery of so-called enhanced broadband services, toward introducing further gains into the network in terms of reducing energy consumption, while reducing traffic on the macro cell network.

Under the umbrella of broadband services, disseminating popular data to subscribers groups is currently established using unicast connections every time a user requests data, which is particularly inefficient if many co-located users request the same content. Examples of these scenarios include live video streaming for social events (stadiums, concerts, or conferences), gaming (mobile gaming applications like Pokemon Go or in-site gaming championship where the competitors are ranging on microseconds latency). However, the legacy approach on establishing unicast links is clearly inefficient, in particularly in massive content delivery scenarios where users are close to each other.

In this context, this chapter will first review the latest developments in enabling technology tools and data dissemination in cellular networks as a basis for our novelty. Moreover, we develop the notion of energy-efficient content distribution for virtual UDNs, where we harness virtualization, and D2D (connectivity) through local area networking, to develop a new networking topology for cost-effective data distribution. Moreover, NCC is investigated as overlay technology for enhancing network resiliency. These topics are elaborated in Sect. 13.2. This concept is evaluated using a new experimental tool that was purposely designed for testing cloud-based services within a virtual MSC (mobile small cell) environment. In this context, we discuss a framework for testing SDN-based services for 5G and beyond in Sect. 13.3 and will showcase how content distribution can be effectively implemented within an SDN cooperative-based ecosystem as a case study in Sect. 13.4; and finally the conclusion is given by Sect. 13.5.

13.2 Enabling Technology Tools

13.2.1 Network Coding

The concept of network coding (NC) was introduced in [2]. Then, [3] demonstrated its use to leverage network throughput and resilience in wireless networks. RLNC was first introduced in [4], and it was introduced to be optimal for wireless network due to the randomness of the protocol. However, it was observed that there was no need to code all packets in the network unless the channel losses were unsustainable. Systematic RLNC [2] is a variant of RLNC in which the packets are first sent without coding; then coded packets are transmitted. It has been observed that systematic RLNC results in a higher probability of decoding the generation [36] and reduces the decoding complexity at the UEs when compared to full-vector RLNC [5]. Despite the clear benefits of this new approach, any loss in the channel would increase the latency of the decoding packets due to the need in RLNC to retrieve packets in order. [6] introduced a new protocol in which the redundancies are included inside the generation, which considerably reduced the system latency. However, the scenario that we want to address, where multiple nodes in a small cell that can be heterogeneous talk to each other, entails a new issue. The protocol presented in [6] responded poorly to networks where the packets could arrive from different paths. A more robust version of this protocol was presented in [7], which is the protocol used in this chapter.

13.2.2 Cooperative Relaying

In [8], the benefits of small cells or cloudlets to provide massive content delivery in cellular networks are introduced. Later on, in [9], he introduced the concept of mobile cloud (MC). He defined it as a group of nodes inside a small cell that share their resources opportunistically, cooperating with each other to obtain a common benefit. Laneman gave a detailed overview of the benefits of cooperative relaying in regards to the network performance [10]. Last but not least, there are as well many approached on opportunistic D2D communication that enables data dissemination by taking into account the geographical position of the users [11].

13.2.3 Virtualization

Virtualization offers separation of software from physical infrastructure, and it provides an abstraction layer to allow different operating systems, mobile operators, or users to share common underlying physical resources while isolating virtual resources. This approach supports an efficient, flexible, and dynamically re-

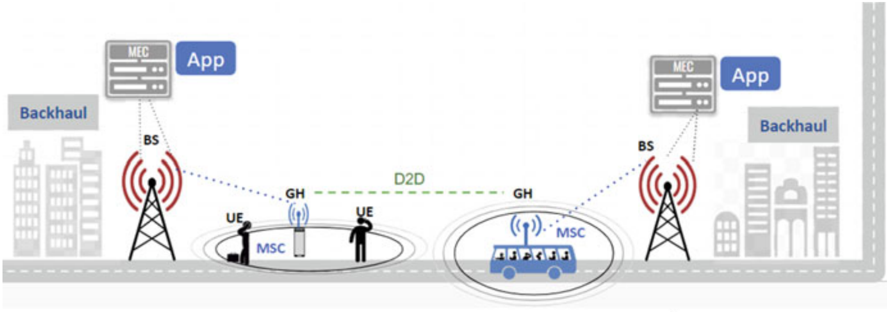


Fig. 13.1 Use case scenario of mobile small cell cooperation

configurable usage of physical resources and an efficient service creation process. Virtualization not only provides the dynamic setup of services but also supports the functionalities necessary for implementing multi-tenancy and autonomic management.

Virtualization provides an effective and exploitable deployment of computation capabilities at the mobile network edge, allowing managing and orchestrating network services from different services providers in dense small cell scenarios and different use cases. It can be applied to communication infrastructures, such as core/edge network elements and access points in small cells. The virtualization of network edge services offers several computational capabilities at the network edge. Network functions (NFs) can be virtualized using SDN-based approaches. The applications of mobile edge virtualization include Internet video streaming that demands more bandwidth and higher video quality.

13.2.4 On-Demand Mobile Small Cells

Future communication is expected to be a network of a large number of mobile users that will interact together with many applications such as VR/AR games or streaming videos requiring low response times. Therefore, mobile small cells can be deployed via a cluster of UEs using D2D technology [12]. This type of small cells can also be deployed to cover the urban landscape. These cells can be formed using mobile devices, i.e., (user equipment) or low-cost nodes such as remote radio units (RRHs) which are connected to aggregation nodes through wireless links. The base stations (BSs) of these cells mainly perform transceiver functionalities and are less involved than a fully functional BS. Mobile small cells are also called on-demand because the can be flexibly deployed in case of significant traffic variations in short time ranges, etc. Figure 13.1 shows the use case scenario of mobile small cell communication.

13.2.5 Software-Defined Networking (SDN)

SDN has emerged as the most promising candidate to improve network programmability and dynamic adjustment of the network resources. Unlike the traditional network paradigm that focuses on hardware-centric networking, where switches have their own data and control planes, SDN is defined as a control framework that supports the programmability of network functions and protocols by decoupling the data plane and the control plane, which are currently integrated vertically in most network equipment. SDN follows a logically centralized architecture through SDN controller that acts as a control entity. SDN controller provides an application programming interface (API) and is responsible for an abstraction of network resources through APIs. One of the main benefits of this architecture resides on the ability to perform control and management tasks of different wireless and wired network forwarding technologies (e.g., packet/flow switching or circuit switching) by means of the same network controller. To enable SDN, OpenFlow protocol is most commonly deployed. OpenFlow protocol offers control plane functionalities that include system configuration, management, and exchange of routing table information. It also provides logical switch abstraction, maps high-level instructions of the protocol to hide vendor-specific hardware, etc. This abstraction enables SDN to perform network virtualization, that is, to slice the physical infrastructure and create multiple coexisting network slices (virtual networks) independent of the underlying wireless or optical technology and network protocols. In a multi-tenant environment, these virtual networks can be independently controlled by their own instance of SDN control plane (e.g., virtual operators).

13.2.6 Related Work

Today, if a client requests a video stream from a server, the cellular base station will establish a unicast link to provide the user with that service. However, if multiple co-located users in the same cell request the same video, the base station will establish replicated links that send replicated information, which is inefficient. To tackle this problem, two different approaches can be used [13], shown in Fig. 13.2. The first approach multicasts over cellular networks similarly to Wi-Fi. The second approach uses MCs. The clients are organized in groups, and the base station grants the clients requests in a way that the information is sent only once to the whole MC. Table 13.1 shows a comparison of the state-of-the-art techniques for data dissemination in cellular networks.

eMBMS [14] is based on a protocol called FLUTE. FLUTE works as HTTP but unidirectional. eMBMS is a protocol designed by the 3GPP group that uses single frequency networks (SDN) to distribute the signal using the same frequency.

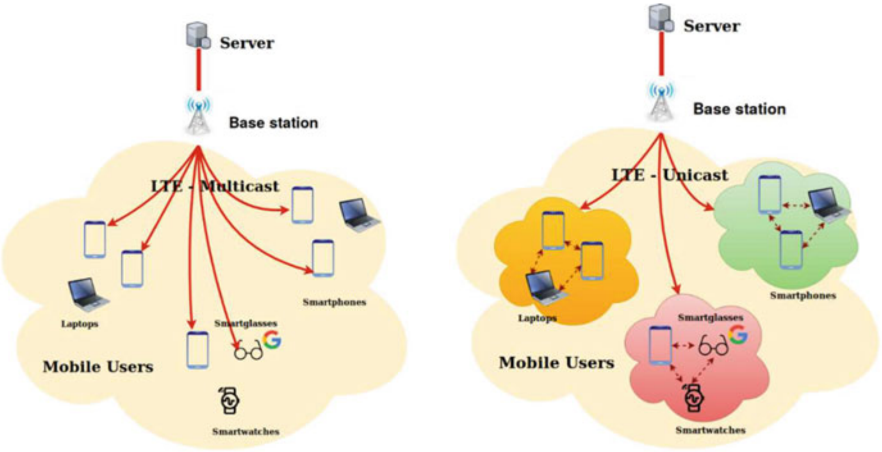


Fig. 13.2 The two most used architectures for data dissemination in LTE networks [13]

Table 13.1 State-of-the-art techniques in massive cellular dissemination

	LTE-A	NC	Short-range	FEC
eMBMS [14]	✓	✗	✗	Raptor codes
MicroCast [15]	✓	✓	Unicast	Network coding
NCMI [16]	✓	✓	Unicast	RLNC
NCVCS [17]	✗	✓	Multicast	Network coding
NCC system [18]	✓	✓	Multicast	RLNC

MicroCast [15] is a subgrouping scheme in which each user utilized two network interfaces, namely, a cellular interface and a short-range interface. They showed improvements in the network by offloading the cellular network into Wi-Fi.

NCMI [16] leverages D2D communications to improve the performance in cellular networks. In their example, the link capacity for the cellular network and the D2D link are the same. RLNC is used to recover missing packets.

The network coding-based video conference system (NCVCS) [17] uses Wi-Fi multicast along with RLNC to provide a solid and local content distribution inside a LAN. They only use one interface, since the cellular network is left out of the scope of their work.

NCC is the protocol we present in this chapter. The analytic model was introduced in [18] and extended in [19]. The implementation was done in [20], and in [21] a wireless demonstrator, where four clients shared a video file, was presented. NCC focuses on low energy consumption while maintaining high reliability of successful content distribution.

13.3 NCC for Energy-Efficient Content Distribution in MSCs

13.3.1 Motivation for NCC in MSCs

Cisco reported that the amount of traffic will increase from 7 eB in 2016 to 49 eB in 2021. The majority of this traffic (78%) will be video traffic, since AR/VR applications, video streaming, video games, and video surveillance are taking over the main interests of the people in the last decade. Moreover, URLL (ultra-reliable and low-latency) applications such as car-to-car communications or video games are topics of interest as well. Thus, a mechanism to reduce the usage of the network, to increase the throughput, or to reduce the end-to-end latency is needed [1]. The concept of tackling all fronts at the same time sounds challenging; therefore different solutions for each of the requirements should be developed.

Furthermore, the influence of network coding in mobile clouds has not been completely studied, leaving research unknowns such as how many nodes should cooperate in the cloud, how should they communicate, where to put the base station, and multiple others.

13.3.2 Architecture and Design

In our solution based on the notion of NCC networks, we extend the concept of small cells to a more specific one. We define a MC (mobile cloud) as a group of nodes inside a small cell that share their resources opportunistically, cooperating with each other to obtain a common benefit. This protocol uses small cells to offload the traffic from the cellular communication into the short-range communication network inside the small cell, like Wi-Fi. The protocol comprises of two different phases, the cellular phase and the cooperative phase, that can occur sequentially or in parallel.

We use RLNC in the video server to encode the video packets and in each UE of the MC to recode and decode the packets. In RLNC, the encoder gathers the packets in blocks of size g, namely, generation size. It linearly combines packets in the block multiplying them by random coefficients, creating new coded packets that are transmitted through the network [4]. At the destination, the decoder only needs enough linear-independent packets to fill a coding matrix of size g. Then, the decoder performs Gaussian elimination to that matrix to obtain the original packets. The main benefit of using RLNC in this setup is that every NC coefficient is random, so during the recoding phase, the UEs do not need to wait for any deterministic coefficient. They simply recode the incoming packet with a new random coefficient and change the RLNC header. This reduces the latency that is inherent to coding, unlike in other coding protocols like Reed Solomon or turbo codes. Furthermore, in the case of lossy scenarios, the source only needs to send extra redundant packets to the destination, avoiding feedback techniques that would delay the communication.

The variable n represents the maximum number of nodes that a MC can handle, hereafter referred to as the cloud size. We assume that all members in the mobile cloud can have cellular communication with the eNB and they are close enough to communicate with the rest of the nodes in the MC through a short-range technology, namely, Wi-Fi. They all request the same content, and nodes can eventually join or leave the cloud [9].

MCs are formed in periodic formation phases, where nodes can join or leave the cloud on the fly. Therefore, the structure of the MCs can be modified at each formation phase. However, the maximum cloud size is always n. That is, clusters of $n' < n$ UEs are only formed if the number of the remaining UEs that request the same video streaming is less than n. MC formation phases are centralized, so the eNB is in charge of the formation of the MCs as it possesses a global view of the network. The process of content distribution was divided into two different phases.

Cellular Phase The eNB distributes the g packets to the UEs connected via time-multiplexed unicast sessions in a round-robin fashion. Each of the n UEs is assigned an index, which defines the order in which they will receive the data packets from the eNB, i.e., as illustrated in Fig. 13.3, the first packet is sent to the first (black) UE, in the first time slot. In the second time slot, the eNB sends the second data packet to the second (blue) UE and so on. The eNB will send a second packet to first UE at timeslot $(n + 1)$ after sending a data packet to the n^{th} UE; this will be the $(n + 1)^{th}$ packet of the generation. In the specification of LTE-A [22], the data transmission takes place in a slotted channel, whose minimum scheduling unit is one subframe, with duration d_s = 1 ms. That is, the minimum unit for data transmission (downlink) in LTE-A is the physical resource block(PRB), which is defined as the number of consecutive OFDM symbols in the time domain and the number of consecutive subcarriers in the frequency domain [22]; in the time domain, two PRBs fit in one subframe. Therefore, in the frequency domain, only one PRB is utilized simultaneously in each cluster. At the end of this phase, all g packets will be distributed over the UEs, where each UE will have g/n or $g/n + 1$ packets depending on the order in which the connection between the eNB and the UE was established.

Cooperative Phase When a UE receives a packet from the server in the eNB, it will be in charge of redistributing the packet to the rest of the nodes in the MC. Since no feedback messages are transmitted, the eNB must inform the number of time slots allocated for the content distribution within the MC to the UEs. Each UE will be assigned an index i in order to create the TDMA schedule in the cellular phase. Every time slot, a UE sends a Wi-Fi multicast packet to the remaining UEs in the small cell. Each time slot the transmitting client is changed to distribute all resources in the small cell uniformly [15]. The time slot in this phase does not need to be the same as in the cellular phase since a different data rate can be used. Both cellular and collaborative phase can be observed in Fig. 13.3.

A timing diagram of our NCC protocol is depicted in Fig. 13.4. In this diagram, we show how the data dissemination protocol on three devices and five packets. In the cellular phase, the packets are first sent uncoded to the UEs. These uncoded

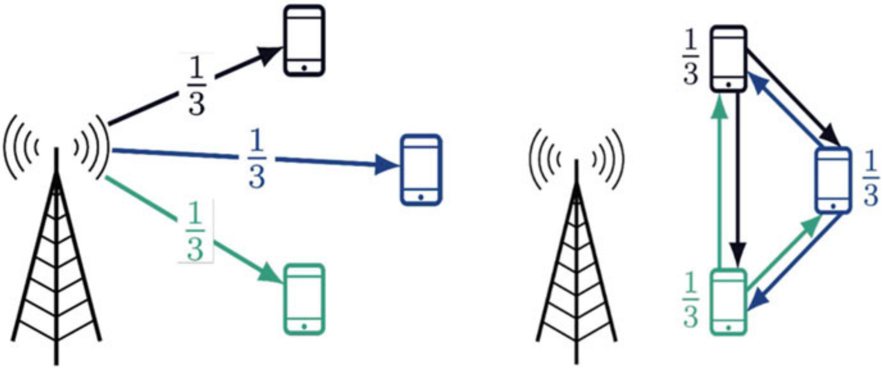
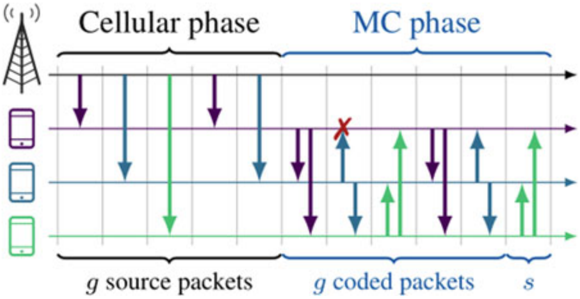


Fig. 13.3 Phases of NCC protocol: (**a**) cellular phase, and (**b**) collaborative phase [19]

Fig. 13.4 Timing diagram for the proposed NCC protocol [18]

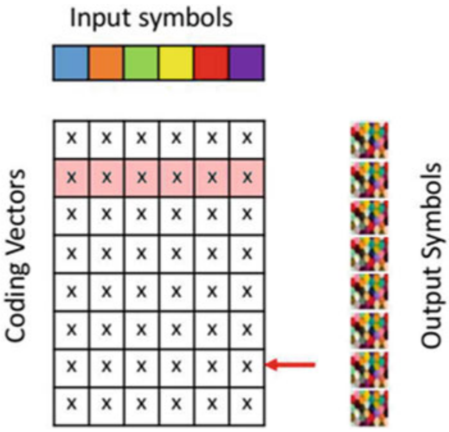


transmissions are called systematic transmissions [16]. Then, each UE sends the coded packets to the rest. The system recovers from an error that occurred in the second time slot, in the first coded transmission. This example comprises three nodes, a generation of five packets, and one coded transmission.

13.3.3 RLNC Protocol for MSC

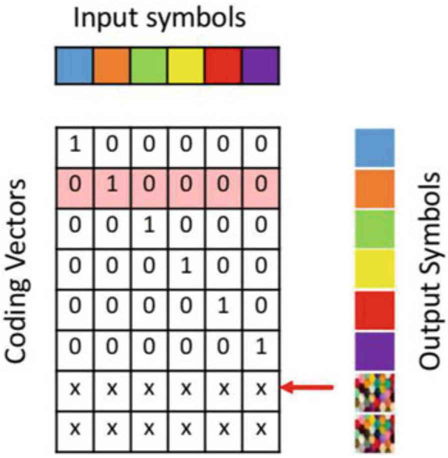
In this subsection, we are going to review the path toward the selection of the optimal NC protocol based on the family of network codes that is referred to as “random linear network coding” (RLNC) that was chosen due to the lack of complexity and good performance; this NC variant is based on linearly (linear) combining (coding) packets weighted by random coefficients [4].

Full-Vector RLNC In full-vector NC (see Fig. 13.5), all packets are encoded at the source and sent through the network. In the sink, all of them are decoded if the rank of the decoding matrix is higher than the NC generation or equal. However, if the rank is lower, none of the packets will be decoded, since the number of unknowns exceeds the number of equations in the receiver. This makes the error case



x	x	x	x	x	x
x	x	x	x	x	x
x	x	x	x	x	x
x	x	x	x	x	x
x	x	x	x	x	x
x	x	x	x	x	x
x	x	x	x	x	x
x	x	x	x	x	x

Fig. 13.5 Full-vector RLNC: all output symbols are coded and sent to the receiver [23]

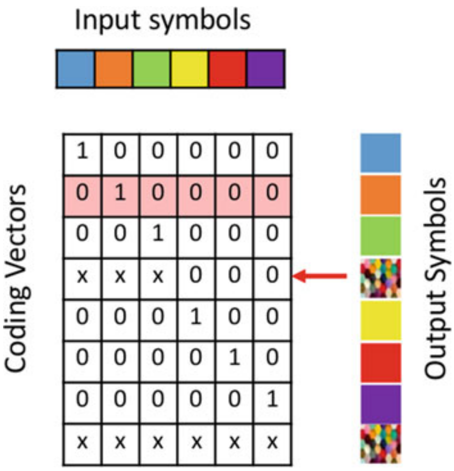


1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1
x	x	x	x	x	x
x	x	x	x	x	x

Fig. 13.6 Systematic RLNC: first, the original packets are sent uncoded. Then, the transmitter sends coded redundancies [23]

a catastrophic result for the system. Moreover, encoding each packet will increase the latency per packet and the complexity of the problem.

Systematic RLNC This approach leverages the low loss ratio in the channel. The source will send first uncoded packets (see Fig. 13.5) matching the generation size. These packets are called systematic packets. After those packets, coded ones often called redundancies are sent. This approach lowers the complexity of the protocol since the endpoints do not have to encode/decode every packet. Furthermore, if the erasure ratio in the channel is low, packets can be directly forwarded to the next layer, since they are not coded. The main drawback this approach has is that, in the event of a loss or corruption, the sink will have to wait until it received the first coded packet, that is, at the end of the generation. This increases the latency per packet (Fig. 13.6).



1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
x	x	x	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1
x	x	x	x	x	x

Fig. 13.7 PACE RLNC: the coded transmissions or redundancies are injected in the middle of the systematic or uncoded packets [23]

PACE RLNC This approach focuses on lowering the latency per packet issue of the previous protocol. Instead of sending all the redundancies at the end of the generation, it includes some of them inside the generation (see Fig. 13.7). Therefore, the receiver will not have to wait until the end of the generation to recover from a loss. This approach will stand out in networks where the erasure ratio is not that high so as to require the full vector approach and not that low as to require the systematic option. However, it was observed that in multipath communication where a sink is receiving packets from multiple sources, the jitter of the network is not negligible, and packets do not arrive in order.

This creates a problem from RLNC protocols since they are intended to deliver in-order packets. If a decoder receives a packet from the following generation, it will assume that the current generation is over and will try to retrieve all possible packets and start a new generation. In the event this packet arrives too early, multiple packets from the previous generation will be lost. To solve this issue, the PACE protocol was updated and renamed PACEMG or PACE Multigeneration [7], which is the protocol used in NCC.

PACE Multigeneration RLNC This new protocol is based on an infrastructure architecture of a decoder followed by multiple sub-decoders. A sub-decoder is an entity that is created when the first packet of a new generation arrives; it stores the data of a single generation and is destroyed when the generation is decoded. The decoder entity is now in charge of orchestrating the sub-decoders and redirects each new incoming packet to the correct sub-decoder [7, 6].

In Fig. 13.8, an example that showcases the advantages of PACEMG over PACE can be observed. The example is divided into the following four steps:

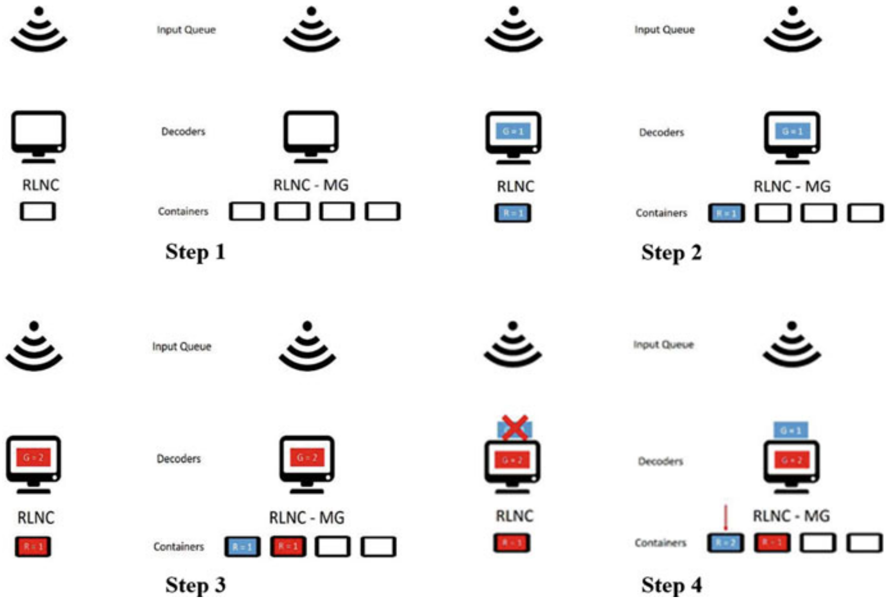


Fig. 13.8 Example: advantages of Pace MG over PACE in high-jitter networks [7]

- Step 1. As an example, an idle system with one decoder is presented. The decoder has no starting generation. We are assuming a wireless input where multiple sources can be delivering packets.
- Step 2. The first packet arrives. After receiving it, the systems obtain the generation of the packet and adjust their internal parameters (Generation, Rank, etc.).
- Step 3. A second packet arrives with a new generation number. In this case, the left system flushes its memory decoding all possible packets. The right system saves the second packet in a second sub-decoder, adjusts its internal parameters, and awaits new incoming packets.
- Step 4. A third packet arrives with an old-generation number. The system on the left considers this is an old-generation packet and discards it. The system on the right keeps it and updates the parameters of the corresponding sub-decoder. As a consequence, it observed a trade-off between the decoding probability and the average latency per packet. In a scenario where jitter was high, it was observed that the decoding probability increased as well as the average latency per packet.

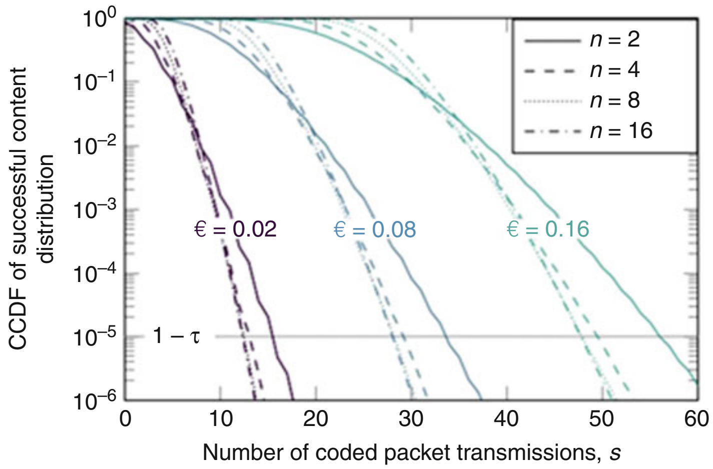


Fig. 13.9 Complementary CDF (CCDF) of successful content distribution for a Galois field [18]

13.3.4 *Analytical Model*

The analytical modeling of the NCC protocol used Markov chains, where the objective was to estimate the minimum amount of coded transmissions that are necessary for the receiver to perceive a certain amount of quality of experience (τ ≥ 1 − 10). The evaluation results consist of the estimation of a successful content distribution depending on the number of nodes and the error rates, the average throughput per node depending on the cloud size, and the average energy consumption per UE depending on the number of nodes in the cloud.

Figure 13.9 depicts the probability of getting successful retransmission depending on the number of coded transmissions.

The behavior observed is very similar to every error rate. In the first phases where the CCDF is not small, smaller clouds perform better. However, if we want to have high reliability (τ ≥ 1 − 10), larger sizes are needed.

In Fig. 13.10, the average throughput per user equipment can be observed. In the conventional mode, the maximum throughput that can be obtained for a single unicast LTE-A session with a UDP packet of 1470 bytes is 11.76 Mbps.

We observe here that the throughput is halved, depending on the error rate. This is due to timeslot allocation. The peak in throughput can be observed when the cloud sizes are in order of 20 UEs each. Figure 13.11 shows the energy savings in each UE when increasing the cloud size. While the energy consumed by the reception of LTE decreases, the energy in the Wi-Fi link increases. Despite the energy increase in the Wi-Fi link, the overall energy per user is decreased due to the fact that a Wi-Fi antenna in both transmission and reception consumes less than an LTE antenna. A small peak can be observed at the beginning of the graph, when the cloud size is around n = {2, 3, 4}. This energy increase occurs because the losses in the Wi-Fi

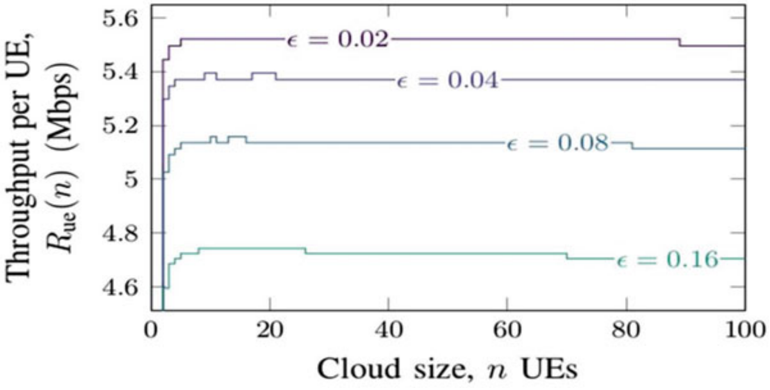


Fig. 13.10 Throughput per UE given a minimum packet distribution reliability [18]

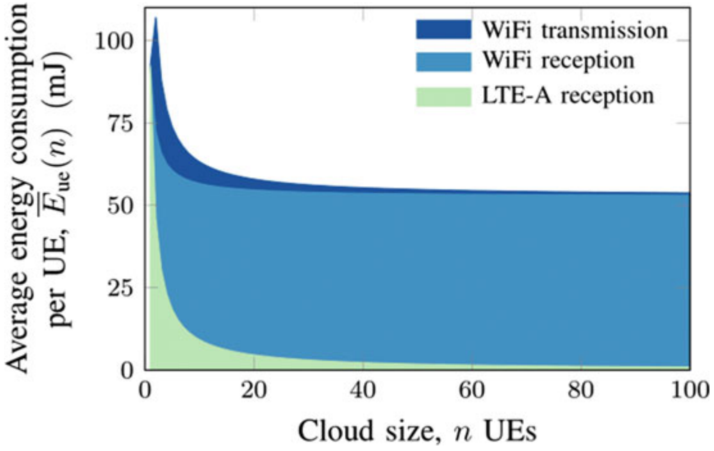


Fig. 13.11 Average energy consumption per UE given an error rate [18]

channel make the UEs need to send more redundancies, something that is not needed in the LTE communication. However, this effect is sharply reduced when the cloud size increases to a number of n > 5 units.

13.3.5 Physical Testbed

We built a testbed that consists of multiple clients in a MC that uses NCC to receive a stream from a server in the cloud. In order to fulfill the requirements, it is necessary for UEs to have two interfaces, one to request the video from the cellular communication in the cellular phase and one to share and receive packets from the

Table 13.2 Hardware specifications for testbed elements [20]

Element	Description	Specifications
User equipment	Acts as the client which requests the video stream to the server	Intel NUC6i5SYH 4 core 8 GB
Wi-Fi AP	Provides connectivity between the UEs in the cooperative phase	TP-link archer C9
Cloud service provider	Provides a location for the streaming server	Amazon Web Services
LTE-A dongle	Provides connectivity between the UEs and the server in the cellular phase	Huawei E3372
LCD screen	Display the video requested	Waveshare 5-inch LCD (B)

rest of UEs in the cooperative phase. Even though high-range mobile phones from the last generation can activate a feature to use the Wi-Fi link to help the cellular link reach faster download speeds, this feature is not very flexible to work with. Hence, we decided to use small portable computers with a Wi-Fi interface and attach them to an LTE-A dongle with a SIM card for the cellular interface [21].

The video streaming server is located in an Amazon Web Service cloud in Frankfurt, Germany, while all UEs are located in Dresden, Germany. The last device needed is a Wi-Fi access point, used to provide short-range communication in the cooperative phase. Finally, 5-inch LCD screens are attached to each UE to display the video. Table 13.2 lists the hardware specifications used for each element in the testbed.

The testbed works as follows. A server runs in the cloud, far away from the UEs. Each UE requests access to the server on its own. The server will grant the unicast connection and start sending the coded video stream. If there are more users connected, the server will send the packets in a round-robin manner to all the clients connected to it. The UEs will be connected in a multicast group, and a Wi-Fi access point will provide the connectivity. Each UE will recode and forward the packets received to the rest of the group. When the generation is complete, the UE will decode the coded packets and display them on the LCD screen.

In this subsection, we plot the results obtained through simulation and compare them to the ones obtained analytically. Testbed parameters are listed in Table 13.3.

The testbed parameters were selected so that the demonstrator emulates a scenario as close as possible to reality. Hence, we consider the number 100 in the analytical model since it is more intuitive to understand. However, due to the on/off nature of computers, we decided to use 32 symbols per generation. Regarding field size, we decided to use 2^8 because it reduces the chances of receiving linear-dependent packets. We used a base number of eight clients in the testbed. Nevertheless, this number was also changed to compare the results with different cloud sizes. Despite setting a fixed packet erasure rate to 10%, inherent losses on the Wi-Fi multicast channel make the end value deviate from the initial one. We define the coding ratio (CR) as the number of packets a UE sends per 100 packets

Table 13.3 Testbed environment parameters and values [20]

Parameter	Symbol	Settings
Generation size	G	32 packets
Field size	Q	2^8
Cloud size	N	2 . . . 16 UEs
Coding ratio	CR	{104 . . . 116}
Packet erasure rate (PER)	ϵ	0.1
Subframe duration	d_s	1 ms
Payload size	L	1400 bytes
Data rate at LTE-A and Wi-Fi link	R	11.76 mbps
Power consumption for LTE-A reception	$P_{cel,\ rx}$	924.57 mW
Power consumption for Wi-Fi transmission	$P_{wifi,tx}$	442.60 mW
Power consumption for Wi-Fi reception	$P_{wifi,rx}$	442.60 mW

received. The value of the coding ratio varies from 104 to 116 to compare them in the results. The packet length varied whether the transmissions were systematic or coded. We define the payload size as the number of bytes each packet can carry. On top of the payload, different layers of encapsulation are applied, always respecting the MTU to avoid segmentation.

Figures 13.12 and 13.13 show the CDF of the average packet loss in the UEs. Figure 13.12 shows the relation between the CR and the loss probability, which is complementary to the decoding probability. Figure 13.13 shows the optimal number of clients inside the MC. In both cases, we can observe similar behavior in the results. This occurs due to the nature of RLNC. If the coding ratio does not suffix the losses in the channel, none of the packets will be obtained (because they are coded). When the coding ratio starts matching the channel erasure rate, the number of packets decoded increases drastically to the maximum decoding probability. Another important observation is that this scheme does not match the "the more, the better" in both coding ratio and number of clients in the cloud regarding losses. The performance of using a coding ratio of 116 is worse than the optimal one (i.e., 112), and a larger number of clients (i.e., 16) does not reflect a better performance as in N = 8. If the coding ratio or the cloud size increases such that too many packets are on the fly, useful packets are delayed and stored in queues and by the time they arrive at the decoder is already too late. We observe a trade-off between the useful packets sent in the multicast link and the useless ones (due to linear dependence or network congestion).

Figure 13.14 shows the latency per packet of the NCC protocol. Since the upper part cannot be clearly observed, we decided to zoom in. Therefore, the projected colored box represents exactly the part of the graph (from 0.95 to 1.0). Since Fig. 13.14 includes several interesting insights, we divide it into five different parts. Hereunder a description of each part is described:

- The distance between the 0 ms mark and Line A represents the minimum latency the packets can have due to the transmission delay. The height of Line A

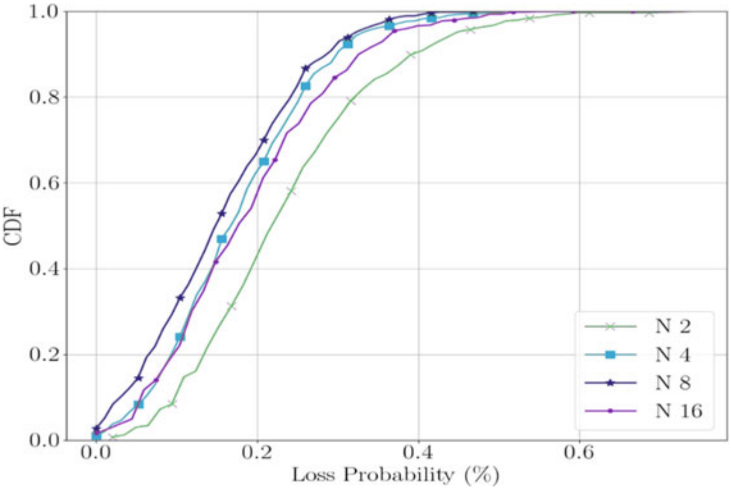


Fig. 13.12 CDF of the average packet loss probability for the testbed [20]

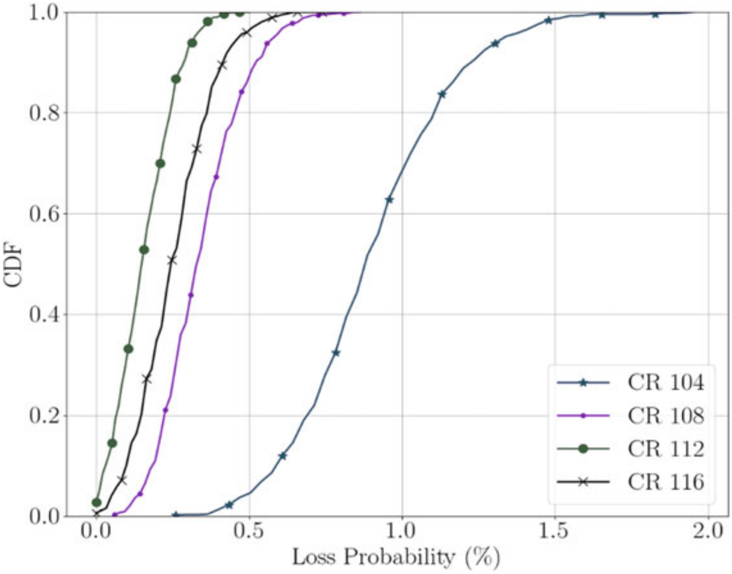


Fig. 13.13 CDF of average latency per packet observed in all eight UEs [20]

represents the percentage of packets that are decoded right after being received, without waiting in the coder of the queues.

- Line B represents the cases where a packet is lost or corrupted, but it is recovered within its generation. The protocol we use provides in-order delivery, which means that if a packet is lost, the rest of the packets that arrive later will wait

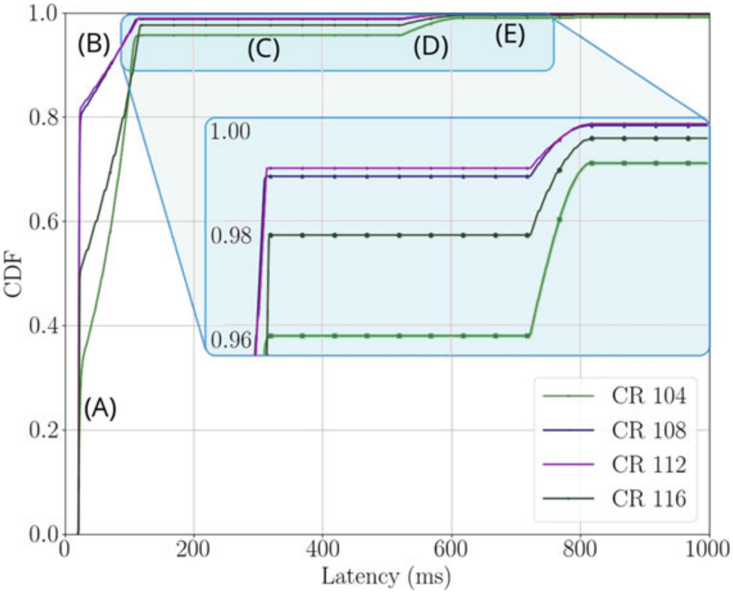


Fig. 13.14 Average latency per packet in the testbed evaluation [20]

until the lost packet is recovered. Hence, the first packet sent after the lost one will have a higher delay than the last packet sent before the error was corrected. This generates a linear slope, as observed in Fig. 13.14.

- Line C shows an internal timeout. This timeout is triggered in the case the protocol cannot recover the loss. The tuneable timeout starts when the first packet is decoded, and it is refreshed every time a new packet is decoded. In our protocol, we set a timeout of 500 ms, which can be observed in the Line C. No packets will arrive until the timeout (500 ms plus 20 ms of transmission delay) is reached. We are aware that the value of this timeout is not ideal and it depends on the losses of the channel and the architecture of the system.
- Line D shows the latency of the packets that arrive after an unrecovered error. If an unrecovered error occurs, the protocol will wait and store the packets that arrive during the waiting interval. The waiting interval is characterized by the timeout flag, which is our case is 500 ms. That means that, after an error, the protocol will wait for 500 ms to see if the error can be recovered. When the timeout is triggered, the protocol will forget.
- The lost packet and the stored packets will be decoded. In Fig. 13.13, it can be observed that Line D is not linear. The explanation of this shape lies in the in-order delivery nature of the protocol. A loss can occur at the beginning, at the middle, or at the end of the generation. When the timeout is triggered, the packets of the generation that arrived after the loss (which is a number that is smaller than the generation size) will be decoded, as well as all the packets of all the generations next. This means that the number of packets that wait for more

(the ones that have the same generation of the lost packet) will be less than the number of packets that wait for less.

- Finally, Line E shows an infinite latency for the rest of the packets. This gives us the percentage of unrecovered packets, i.e., the loss ratio.

13.4 Cooperation in SDN-Based Virtualized MSCs

13.4.1 Motivation and Related Work

5G envisions cooperation in future wireless networks as the collaboration of users accessing the common wireless medium, exchanging information regarding the network and channel status, forming small groups to apply cooperative operations with each other. The dense infrastructure networks combined with cooperative transmission between mobile devices have good potential for high efficiency. Cooperation, small cells, and network coding technologies are key enablers of 5G successful deployment [24]. Cooperation among user equipments (UEs) has several major advantages in terms of increasing energy efficiency [25], increasing network throughput, reducing communication costs, and increasing network coverage. Specifically, cooperation in short-range communication has led to significant performance gains such as increased energy efficiency and higher data rates.

Mobile edge virtualization with adaptive prefetching (MVP) has been proposed which enables content providers to embed their content intelligence as a virtual network function into the mobile network operator's (MNO) infrastructure edge [26]. In the proposed architecture, in order to achieve QoE-assured 4 K video on demand (VoD) delivery across the global Internet, the authors presented a context-aware adaptive video prefetching scheme.

SDN simplifies the structure of multitier networks. SDN can be applied to ultradense network cooperation where SDN controller implements cell coordination. In the literature, SDN has been used to offload users from macro cell through small cell cooperation in a software-defined two-tier network [27]. [28] modeled a heterogeneous network (HetNet) of two tiers (macro and small cells) with multiple-antenna BSs serving multiple users. The work focused on cell association policies by introducing the concept of association probability. SDN is used for the orchestration of the network. In the proposed work, the authors demonstrated that small base station (SBS) cooperation outperforms over uncoordinated BSs. [29] presented the distributed procedure of detouring the multicast traffic to WLAN in the current LTE and WLAN heterogeneous network architecture. Furthermore, the authors proposed an SDN-based centralized approach for multicast content dissemination. The presented work analyzed and showed that an SDN-based approach in heterogeneous cellular communication and multicasting brings significant radio resource savings.

13.4.2 SECRET Testbed: Enabling MSCs Technology

With the tremendous growth in wireless traffic and service, it is inevitable to extend virtualization to wireless networks. The introduction of new paradigms such as edge computing and radio access (RAN) functional split is evolving into new cloud computing services, where virtualization is coupled with cloud computing for supporting new multi-tenant business models, replacing the traditional design paradigms of the mobile network.

The framework envisaged by the ITN-SECRET [30] aims to narrow the gap between current networking technologies and the foreseen requirements of future 2020 networking and beyond to deliver higher networking capacity, ability to support more users, lower cost per bit, enhanced energy efficiency, and finally adaptability to the new nature of services and devices. The SECRET project builds on current technology trends to support beyond 5G, by aiming toward exploring the hyper dense deployment of mobile small cells.

The SECRET testbed is developed to enable SECRET project use case scenarios such as mobile small cells showing the benefits of network coding. Furthermore, SECRET testbed introduces the new architecture design with provisioning for multi-hop, cooperative clusters with multiple points of access to the overlay network. The testbed investigates multi-tenant sharing based on small cell virtualization in a bid to reduce the operators' OPEX and CAPEX figures.

This section discusses the SECRET testbed, a practical framework developed on OpenStack [31], and OpenDaylight [32] platform to provide integrated manageability for both resources in cloud infrastructure. The platform virtualizes SECRET mobile small cells and performs VM placement and migration, network flow scheduling and bandwidth allocation, and real-time monitoring of computing and networking resources. In the testbed, the cloud platform is developed to implement the mobile edge cloud capabilities, where virtual machines are spawned as virtual resources and forms the SECRET small cells as shown in Fig. 13.15.

The testbed analyses various performance metrics such as the capability to provide small cell interconnection services to multiple devices, authentication and authorization for the small cell uplink, application of network coding to communication links, and application of network coding to the SDN providing the mobile small cell services. The testbed includes the following components:

- OpenStack
- OpenDaylight
- MEC-network coding server
- Virtual mobile small cell
- Acticom plug-in

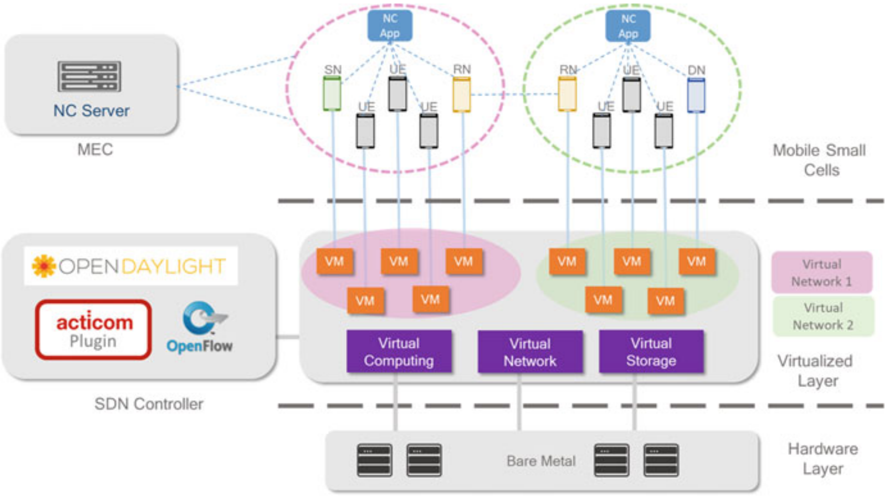


Fig. 13.15 Virtualized SECRET mobile small cell setup

13.4.2.1 OpenStack

OpenStack architecture includes several modules to provide a cloud platform for cloud infrastructures. This allows the user to choose from a variety of complementary services in order to meet different needs regarding computing, networking, and storage. OpenStack provides a modular solution in the form of computing, networking, and storage. The main characteristics of OpenStack are:

- Scalable: OpenStack is scalable up to 60 million virtual machines and billions of stored objects.
- Compatible and Flexible: OpenStack supports most virtualization solutions of the market such as Hyper-V, KVM, and QEMU, etc.
- Open: The entire code in OpenStack can be modified and adapted as it is open source technology.

OpenStack architecture consists of various modules including computing, networking, and storage as shown in Fig. 13.16.

13.4.2.2 OpenDaylight

The OpenDaylight controller supports the important technology trends for virtualized networking: software-defined networking (SDN), model-driven software engineering (MDSE), and model-driven network management. OpenDaylight is a modular open platform for customizing and automating networks of any size and scale. It manages the networks defined in OpenStack and controls the traffic between instances.

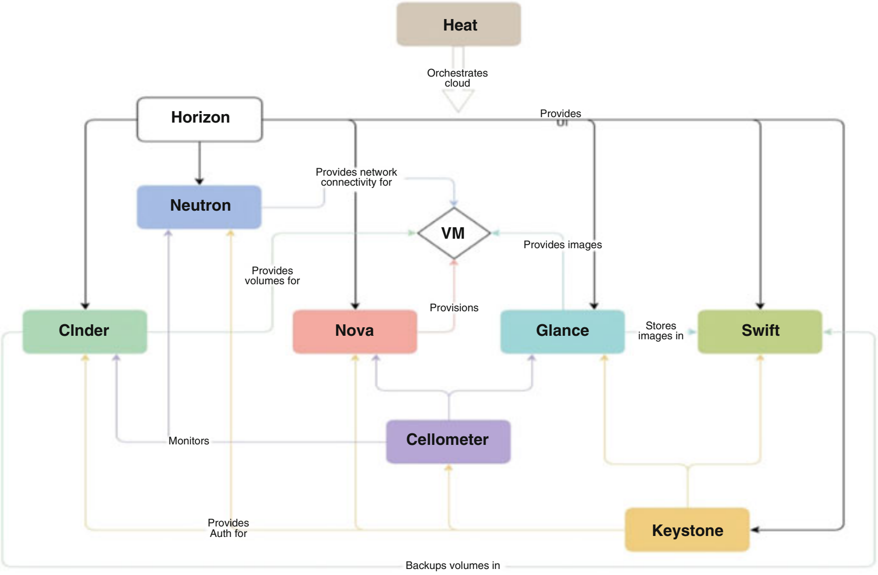


Fig. 13.16 OpenStack architecture [33]

OpenDaylight supports SDN platform protocols such as OpenFlow, OVSDB, NETCONF, and BGP to improve the programmability of networks. OVSDB is the Open-vSwitch database protocol which is used for management in software-defined networks. The network configuration protocol (NETCONF) provides mechanisms to install, manipulate, and delete the configuration of network devices. The border gateway protocol (BGP) is a complex protocol which focuses on the security and scalability of the network.

In the SECRET testbed, OpenDaylight is integrated with the OpenStack cloud platform to manage the traffic. To integrate the OpenDaylight controller with the testbed, the odl-netvirt-OpenStack plug-in is installed in the controller. The plug-in manages the flows on Open-vSwitches installed on compute and controller nodes.

13.4.2.3 MEC-Network Coding Server

Edge services virtualization offers computing capabilities to the network edge and brings services near to the network end nodes such as mobile subscribers. In the SECRET testbed, a streaming server is installed on the edge node, which streams the video to the mobile small cells. The streaming server provides network coding capabilities to the video stream, which encodes the data packets and forwards them to the mobile small cell.

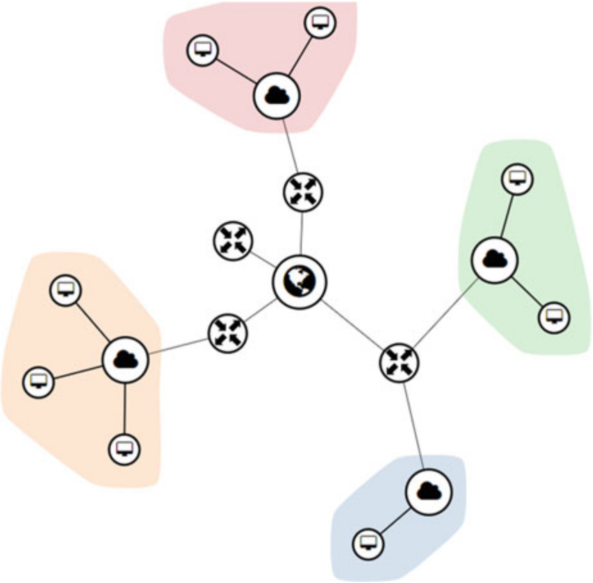


Fig. 13.17 Mobile small cell network topology

13.4.2.4 Virtual Mobile Small Cell

SECRET testbed contains several virtual networks representing virtual MSCs. In the OpenStack environment, self-service networks are considered as virtualized MSCs. The self-service network allows instances to communicate with each other in the same network. Multiple networks can be created in order to achieve a multi-tenant isolation setup in the testbed.

In Fig. 13.17, the example network topology is shown, where four self-service networks are created and attached to the provider network through routers. Routers are created to connect the self-service network to the provider network. Several virtual machines can be spawned on the network. In the setup, virtual machines spawned in the same MSC can communicate with each other. On the other hand, virtual machines in different MSC are isolated from each other.

13.4.2.5 Acticom Plug-in

The Acticom plug-in extends the SDN controller by using the 802.1x port control to provide authentication and authorization to the virtual machines in the testbed setup. When the supplicant requests authentication, the Acticom plug-in performs port control on the Open-vSwitch based on the supplicant credentials. For the access accept message, i.e., when the credentials of the requesting virtual machine could be successfully authorized according to the policies by the AAA backend, the plug-in allows the flows to the virtual machine on Open-vSwitch and enables access to the network. Of course, for the reject message, the Acticom plug-in denies all the flows

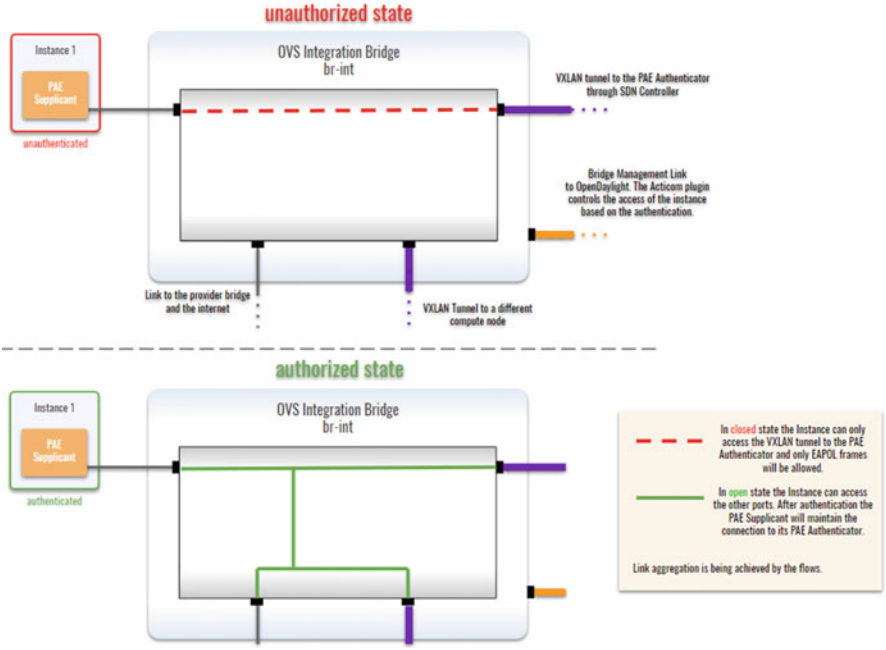

unauthorized state
Instance 1
PAE Supplicant
unauthenticated
OVS Integration Bridge
br-int
VXLAN tunnel to the PAE Authenticator through SDN Controller
Bridge Management Link to OpenDaylight. The Acticom plugin controls the access of the instance based on the authentication.
Link to the provider bridge and the internet
VXLAN Tunnel to a different compute node
authorized state
Instance 1
PAE Supplicant
authenticated
OVS Integration Bridge
br-int
In closed state the Instance can only access the VXLAN tunnel to the PAE Authenticator and only EAPOL frames will be allowed.
In open state the Instance can access the other ports. After authentication the PAE Supplicant will maintain the connection to its PAE Authenticator.
Link aggregation is being achieved by the flows.

Fig. 13.18 Port Control by Acticom plug-in

to the virtual machine and blocks its access to the network. Figure 13.18 shows the port control in the access control setup using the new plug-in.

When a virtual machine is spawned, all the traffic is blocked by the plug-in. After sending authentication messages and receiving the EAP success message by the authenticator, the plug-in modifies the flows in the flow table of OVS allowing incoming and outgoing traffic of the authorized virtual machine to the network. The modification is carried out by validating the MAC address and port number of the virtual machine against the authentication tuple (MAC, port).

The access control policies are defined for both incoming and outgoing traffic to the virtual machines according to their identities. The access control is fine-grained as it uses the credentials of the flow.

13.5 Demonstrating NCC for MSCs

13.5.1 NCC-Enabled Virtualized Testbed

MEC requires a flexible platform to provide SDN capabilities for different application scenarios. The MEC platform orchestrates network and processing resources and emulates the behavior and performance of the underlying network infrastructure. This section describes our proposed NCC communication in a virtualized MSC ecosystem.

The proposed approach is based on an overlay network that logically interconnects all the participating UEs in MSC in the physical network [35]. The SDN-based network allows isolation of overlay network through OpenFlow rules. The overlay network contains virtual objects (VOs), which represents a counterpart of UE in MSC, application components, and functionalities such as computing/storage. Each virtual UE is identified by a MAC address to enable communication among the virtual UEs in the overlay network.

The proposed approach demonstrates the NCC in a virtualized MSC environment, where the MEC server has been developed as illustrated in Fig. 13.15. The MEC server is assumed to be deployed on the eNB where UEs in the MSC are considered as the virtual machines running inside the MEC server. The network consists of a live video server, multiple video-receiving clients, and multicast communication between the clients. The NC server application is installed on the MEC server that transmits data packets to UEs through unicast sessions. Whereas, the UEs in the MSC are connected through the network provided by the MEC server. The UEs forward the data packets received by the server to each other using multicast. The coded packets are transferred afterwards in order to recover the errors that occurred in the previous transmissions.

13.5.2 Evaluation

13.5.2.1 Setup

Virtualized MSCs are deployed on the compute node. To consider zero errors in the first phase of data transmission, i.e., data transmission from server to UEs, the NC server application is installed on the compute node. The video server transmits the video stream and the UEs receive the video stream and displays it after decoding. The network between the UEs is the self-service network provided by the OpenStack API. In this network, the UEs transmit the data packets using multicast. The NC server is deployed on a different network. The network between the server and the UEs consists of multiple OpenFlow SDN switches. Open-vSwitch is used in the setup as an OpenFlow switch and is connected to the compute node to provide a computational resource to the network. The compute nodes host the virtual machines that act as UEs to provide client capabilities.

The testbed demonstrates the implementation of the NCC cooperation in MSCs. In the proposed architecture, virtualized MSCs are considered to handle NCC capability. The key components of the testbed are the OpenStack cloud platform to manage network functions and OpenDaylight SDN controller to manage the network flows. In the setup, the orchestrator utilizes OpenStack APIs to start virtualized MSC functions and instructs the SDN controller to manage the flows in NCC communication. The virtual machines are spawned on a compute node that

Table 13.4 Parameter settings [34]

Parameter	Settings
Cloud size, n	$n = \{1, 2, 3, 4\}$ UEs
Packet length	1400 bytes
Generation size	32 packets
Data rate of the LTE-A and Wi-Fi links	11.76 Mbps
Power consumption of LTE-A reception	924.57 mW
Power consumption of Wi-Fi transmission	442.60 mW
Power consumption of Wi-Fi reception	442.60 mW

acts as UEs (virtual objects). The SDN controller establishes the connection between the NC server and the UEs. Additionally, the SDN controller receives the network statistics and manages the network flows of the networks.

The compute node hosts virtual machines, where each VM runs a client and one VM runs a server. Kernel-based Virtual Machine (KVM) is used as a hypervisor on the compute node to host virtual machines. Moreover, hardware acceleration has been enabled to increase throughput. One MSC is deployed as an overlay network and the spawned virtual machines belong to the same network.

13.5.2.2 Configuration Parameters

Table 13.4 lists the configuration parameters. The coding ratio for two clients is considered to be 50%; however, with three and four clients, the coding ratio is considered 40% and 30%, respectively.

13.6 Results and Discussion

In this section, we present the energy consumption and the average transmission and reception throughput in each UE. We use a single unicast transmission as the reference point and we evaluate the setup for different settings of n ∈ {1,2,3,4}.

The average throughput analysis is shown in Fig. 13.19: user throughput per UE in MSC. We observe a reduction in the LTE reception traffic at a rate:

$$\mathrm{LTE}_{rx}^{n} = \frac{\mathrm{LTE}_{rx}^{1}}{n} \tag{13.1}$$

Regarding Wi-Fi, in the case of n = 2, the results state that the amount of data transmitted and received is the same. This occurs because, with two nodes in the MSC, the information transmitted from one node will be received by the other one. Please note that there is a little redundancy in the LTE channel as well in

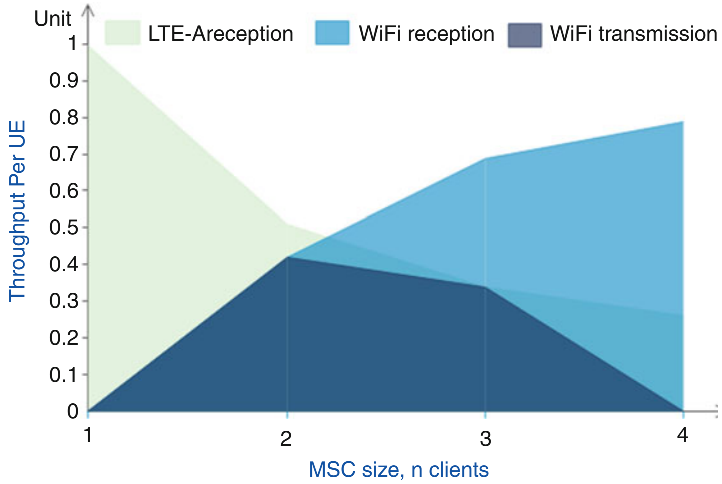


Fig. 13.19 User throughput per UE in MSC [34]

order to provide full resilience against errors in the LTE channel. Hence, the Wi-Fi throughput does not exactly match the LTE-A throughput for n = 2. To summarize, we observe similar results to the analytical ones in [19].

Figure 13.20 shows the average energy consumption per UE. To calculate the energy consumption, the testbed used the model proposed in [9]. The power consumption of LTE and Wi-Fi is tuned as explained in Table 13.4. We observe a huge impact in the average energy consumption, dropping to 38% of its original value when four nodes cooperate in the MSC. We observe a steep reduction at the beginning, followed by a linear and constant reduction as the number of nodes in the MSC increases. The main contributor to energy consumption is the LTE antenna since the reception consumes twice as much power as Wi-Fi. Hence, it is expected that a major drop in LTE reception will promote a major energy consumption reduction. Taking the Eq. (13.1), we observe that the major reduction is between $n = 1$ (100%) and $n = 2$ (50%). From that moment, the LTE contribution reduces by a factor of n (33%; 25%; 20%, etc).

13.7 Conclusion

5G is expected to be a highly mobile, massive, and heterogeneous service environment, and virtualization is key to support these requirements in an efficient manner. Technologies such as SDN and NFV have arisen to provide the virtualization framework to support cloud-based services for UDNs.

In this chapter, we developed the notion of energy-efficient content distribution for UDNs by introducing a new experimental tool for testing cloud-based services

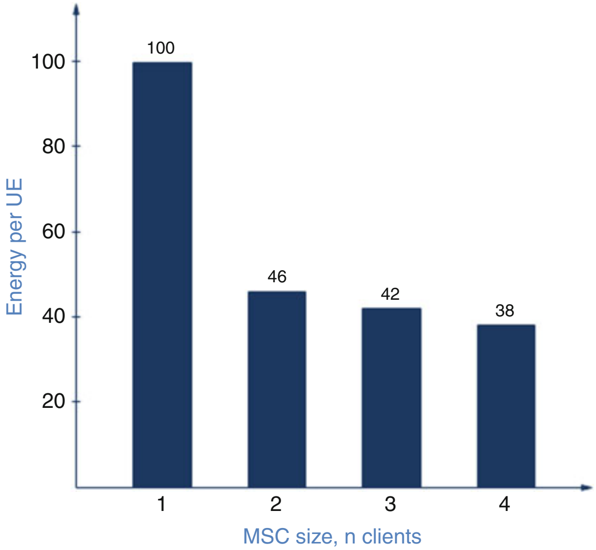


Fig. 13.20 Energy consumption per UE in MSC [34]

within a virtual MSC environment. We introduced a testbed developed in the MSCA H2020 SECRET project that emulates virtual mobile networks and devices. Moreover, we showcase the testbed to demonstrate the eMBB use case and the concept of massive content delivery of streaming services in highly dense virtual mobile networks, i.e., how to offload cell traffic from the LTE macro network to the MSCs. We introduced the protocol, called NCC that enables resilient connectivity, and presented the benefits and drawbacks of the current dissemination technologies. However, there are current bottlenecks that still need to be solved regarding the widespread practical implementation of NCC; these include the following: some smartphones today may not have two interfaces to provide multi-connectivity. Nevertheless, single connectivity can still be used. A second issue is related to the cooperative feature of NCC since the protocol needs the UEs to share resources that require strong user incentive policies.

Moreover, there are two lines of research to emanate as future work; this includes system level studies where the small cell offloading capability/streaming over cooperative NCC takes into account interference between MSCs and, secondly, how NCC can benefit further from the new virtualization frameworks that adds another degree of freedom to their usage, i.e., NCC servers can be strategically placed on the network to attend to different user's requirements (e.g., one MSC streaming the video in HD and other MSC in 4 K); this will enable each NCC server to play a specialized role to deliver a predefined QoS (energy consumption, throughput, latency, etc.). This could even be on the network slice level, which is playing a predominantly role in the B5G paradigm.

Acknowledgments This research work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 722424 (SECRET).

References

1. Cisco. (2017, March 28). *Cisco visual networking index: Forecast and methodology, 2016–2021*.
2. Ho, T., & Lun, D. (2008). *Network coding: An introduction*. Cambridge University Press.
3. Ahlswede, R., Cai, N., Li, S. R., & Yeung, R. W. (July 2000). Network information flow. *IEEE Transactions on Information Theory, 46*(4), 1204–1216.
4. Ho, T., Medard, M., Shi, J., Efiros, M., & Karger, D. R. (2003). On randomized network coding. In *Proceedings of the 41st annual Allerton conference on communication, control and computing* (pp. 11–20).
5. Jones, A. L., Chatzigeorgiou, I., & Tassi, A. (2015). Binary systematic network coding for progressive packet decoding. In *Proceedings of the IEEE international conference on communications (ICC)* (pp. 4499–4504). IEEE.
6. Pandi, S., Gabriel, F., Cabrera, J. A., Wunderlich, S., Reisslein, M., & Fitzek, F. H. P. (2017). PACE: Redundancy engineering in RLNC for low-latency communication. *IEEE Access, 5*, 20477.
7. Torre, R., Pandi, S., & Fitzek, F. H. P. (2018). Network-coded multigeneration protocols in heterogeneous cellular networks. In *SpringerLink Digital Library*, Faro.
8. Fitzek, F. H. P., Katz, M., & Zhang, Q. (2006). Cellular controlled short-range communication for cooperative P2P networking. In *Wireless world research forum (WWRF) 17*. WWRF.
9. Fitzek, F. H. P., & Katz, M. D. (2014). *Mobile clouds. Exploiting distributed resources in wireless, mobile and social networks*. Wiley.
10. Laneman, J. N., Tse, D. N. C., & Wornell, G. W. (2004). Cooperative diversity in wireless networks: Efficient protocols and outage behavior. *IEEE Transactions on Information Theory, 50*, 3062–3080.
11. Ioannidis, S., Chaintreau, A., & Massoulie, L. (2009). Optimal and scalable distribution of content updates over a mobile social network. In *IEEE INFOCOM 2009* (pp. 1422–1430). IEEE.
12. A. Radwan and J. Rodriguez, "Cloud of mobile small-cells for higher data-rates and better energy-efficiency,", 2017. VDE-Verlag.
13. Torre, R., & Fitzek, F. H. P. (2019). A study on data dissemination techniques in heterogeneous cellular networks. In *Broadband communications, networks, and systems*. Springer International Publishing.
14. Viavi Solutions. (2015). LTE multimedia broadcast multicast services (MBMS). *White paper*.
15. Keller, L., et al. (2012). MicroCast: Cooperative video streaming on smartphones. In *MobiSys'12, June 25–29, 2012, Low Wood Bay, Lake District, UK.*
16. Keshtkarjahromi, Y., Seferoglu, H., Ansari, R., & Khokhar, A. (2018). Device-to-device networking meets cellular via network coding. *IEEE Transactions on Networking, 26*, 370.
17. Wang, L., Yang, Z., Xu, L., & Yang, Y. (2016). NCVCS: Network-coding-based video conference system for mobile devices in multicast networks. *Ad Hoc Networks, 45*, 13.
18. Leiva-Mayorga, I., Torre, R., Pandi, S., Nguyen, G. T., Pla, V., Martinez-Bauset, J., & Fitzek, F. H. P. (2018). A network-coded cooperation protocol for efficient massive content distribution. In *IEEE GLOBECOM 2018 Conference Proceedings, Abu Dhabi, United Arab Emirates*. IEEE.
19. Leiva-Mayorga, I., Torre, R., Boscà, V. P., Pandi, S., Nguyen, G. T., Martinez-Bauset, J., & Fitzek, F. H. P. (2020). Network-coded cooperation and multi-connectivity for massive content delivery. *IEEE Access, 1*, 1.

20. Torre, R., Leiva-Mayorga, I., Pandi, S., Nguyen, G. T., & Fitzek, F. H. P. (2020). Implementation of network-coded cooperation for energy efficient content distribution in 5G mobile small cells. *IEEE Access (submitted), 8*, 185964.
21. Pandi, S., Torre, R., Nguyen, G., & Fitzek, F. H. P. (2018). Massive video multicasting in cellular networks using network coded cooperative communication. In *CCNC Las Vegas* (pp. 1–2). IEEE.
22. 3GPP. (2017). *Physical channels and modulation, TS36211.*
23. Ho, T., et al. (Oct 2006). A random linear network coding approach to multicast. *IEEE Transactions on Information Theory, 52*, 4413.
24. Torre, R., Pandi, S., Nguyen, G. T., & Fitzek, F. H. P. (2019). Optimization of a random linear network coding system with newton method for wireless systems. In *2019 IEEE International Conference on Communications (ICC): Communication QoS, Reliability and Modeling Symposium, Beijing* (pp. 1–6). IEEE.
25. Yanikomeroglu, H. (2012). Towards 5G wireless cellular networks: Views on emerging concepts and technologies. In *2012 20th Signal Processing and Communications Applications Conference (SIU), IEEE, Turkey* (pp. 1–2). IEEE.
26. G. I. Tsiropoulos, . A. Yadav, M. Zeng and O. A. Dobre, "Cooperation in 5G HetNets: Advanced spectrum access and D2D assisted communications," IEEE Wireless Communications*,* 24, 5, 110–117, 2017.
27. Ge, C., Wang, N., Foster, G., & Wilson, M. (2017). Toward QoE-assured 4K video-on-demand delivery through mobile edge virtualization with adaptive prefetching. *IEEE Transactions on Multimedia, 19*(10), 2222–2237.
28. Han, T., Han, Y., Ge, X., Li, Q., Zhang, J., Bai, Z., & Wang, L. (2016). Small cell offloading through cooperative communication in software-defined heterogeneous networks. *IEEE Sensors Journal, 16*(20), 7381–7392.
29. A. Papazafeiropoulos, . P. Kourtessis, . M. Di Renzo, J. M. Senior and S. Chatzinotas, "SDN-enabled MIMO heterogeneous cooperative networks with flexible cell association," IEEE Transactions on Wireless Communications*,* 18, 4, 2037–2050, 2019.
30. Bukhari, J., & Yoon, W. (2018). Multicasting in next-generation software-defined heterogeneous wireless networks. *IEEE Transactions on Broadcasting, 64*(4), 915–921.
31. I. SECRET. (2017). *ITN SECRET H2020 – ETN*. [Online]. Available: http://h2020-secret.eu/index.html. Accessed 31 Aug 2020.
32. OpenStack. (2020). *OpenStack*. [Online]. Available: https://www.openstack.org/. Accessed 4 Aug 2020.
33. OpenDayLight. (2020). *OpenDayLight*. [Online]. Available: https://www.opendaylight.org/. Accessed 4 Aug 2020.
34. Rosado, T., & Bernardino, J. (2014). An overview of openstack architecture. In *Proceedings of the 18th International Database Engineering & Applications Symposium* (pp. 366–367). Association for Computing Machinery.
35. Irum, S., Torre, R., Salah, H., Nguyen, G. T., Schulte, G., & Fitzek, F. H. P. (2019). Network-coded cooperative communication in virtualized mobile small cells. In *2019 IEEE 2nd 5G World Forum (5GWF)* (pp. 264–268). IEEE.
36. Lucani, D. E., Medard, M., & Stojanovic, M. (2010). Systematic network coding for time-division duplexing. In *2010 IEEE international symposium on information theory* (pp. 2403–2407). IEEE.

Chapter 14
SDN-Based Resource Management for Optical-Wireless Fronthaul

Michail Dalgitsis, Mohammadreza Mosahebfard, Eftychia Datsika, and John S. Vardakas

Abstract Going beyond the 5G milestone, the next-generation architecture aims to provide an integrated communication platform as a service, in order to handle the different types of devices and varied traffic loads. For this, many operators are moving to software-defined networking (SDN) and network function virtualization technologies (NFV). These technologies help softwarize and virtualize the network architecture and management plane to create enhanced communication capabilities and resource optimization techniques, such as network slicing and functional split. In addition, network softwarization helps to reduce the huge investments implied by 5G, due to high capacity and low latency requirements. In this chapter, we present how an SDN-based architecture represents a feasible solution to effectively address 5G traffic demands over an optical-wireless network, and demonstrate network slicing and functional split use-cases through simulations.

14.1 Introduction

This chapter aims to provide insights on network softwarization in modern mobile networks by exploring the application of SDN technology in converged optical-wireless networks. Network softwarization has revolutionized how networks are designed and operated to deliver network functionality via software running on industry-standard commercial off-the-shelf hardware with greater agility and cost-effectiveness, and SDN is one application [1]. Fundamentals of the SDN paradigm are the separation between the network control logic and the forwarding devices, as well as centralized network intelligence in software components. Due to these key characteristics, SDN is believed to work with network virtualization to radically

M. Dalgitsis (✉)
Centre Tecnològic de Telecomunicacions de Catalunya, Castelldefels, Spain
e-mail: michail.dalgitsis@cttc.es

M. Mosahebfard · E. Datsika · J. S. Vardakas
Iquadrat Informática SL, Barcelona, Spain
e-mail: m.mosahebfard@iquadrat.com; edatsika@iquadrat.com; jvardakas@iquadrat.com

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_14

change the networking landscape toward more flexible, programmable, and highly automated next-generation networks [2].

SDN is shifting the network through the ideas of programmable physical infrastructure and decoupling control and data planes. Network management becomes a simpler task and new services are introduced easily into the network. By leveraging the SDN concept in beyond 5G networks (B5G), advantages like seamless subscriber mobility through a common control plane, and more efficient radio resource allocation through centralization can be foreseen, resulting in a software-defined mobile network (SDMN) [3].

In this chapter, we investigate the SDN paradigm as a 5G enabler for better resource utilization and management in converged optical-wireless networks and consider SDN-based slicing and functional split architectures with multiple services as application examples. The rest of the chapter is structured as follows: Sect. 14.2 contains a more detailed description of the SDN-based converged optical-wireless networks focusing on the mobile optical fronthaul architecture which exploits the analog radio-over-fiber (ARoF) paradigm; Sect. 14.3 demonstrates the provision of network service in users of different requirements and traffic profiles leveraging the capabilities of SDN; Sect. 14.3 also contains descriptions of functional split as a service in a sliced network, monitored by a number of SDN controllers; Sect. 14.4 presents the performance analysis of two use-cases related to network slicing and functional split paradigms, via network simulations, considering a configuration with and without an SDN controller; and finally, Sect. 14.5 summarizes the importance of SDN and its application toward integrated optical-wireless networking infrastructures, and presents future work directions with respect to cloud-based networks.

14.2 SDN Control Plane for Converged Optical-Wireless Networks

5G networks on one hand are expected to provide high data rates (up to 10 Gbps), one millisecond latency, 1000 times bandwidth per unit area compared with long-term evolution (LTE), and 99.999% of network availability [4]. On the other hand, during the last decade, the need for developing novel network abstraction layers in order to perform the network control functions was discerned. This abstraction helps the network management tasks to be more simplified and automated.

To provide the required data rate for 5G networks, the enormous amount of spectrum in the millimeter-wave (mmWave) bands will be employed for enhancing communication capacity [5]. In order to take the advantage of the mmWave, the passive optical networks (PONs), which provides low power consumption and high bandwidth connections, have been employed in both inter-data center connections and access networks [6]. A PON is made of an optical line terminal (OLT), which is in charge of connecting multiple optical network units (ONUs) to metro networks.

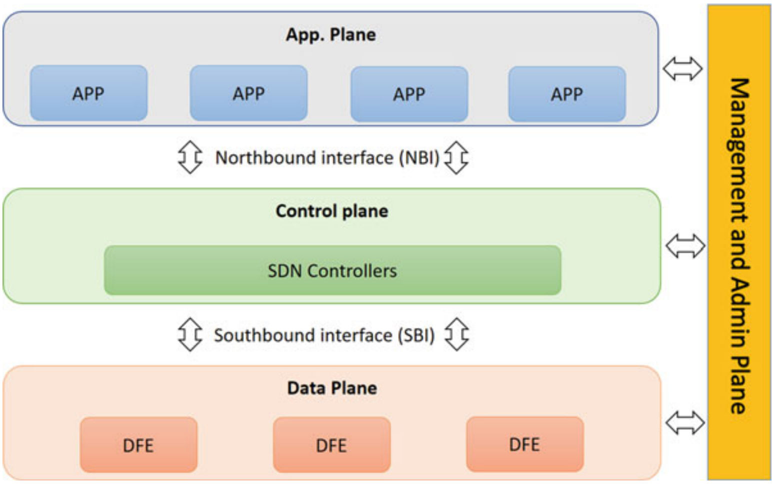

App. Plane
APP
APP
APP
APP
Northbound interface (NBI)
Control plane
SDN Controllers
Southbound interface (SBI)
Data Plane
DFE
DFE
DFE
Management and Admin Plane

Fig. 14.1 A high-level overview of SDN architecture [9]

The OLT broadcasts transmissions in the downlink (DL) direction while, in the uplink (UL), it should manage all the transmitted messages from the physically distributed ONUs for avoiding the possible collisions on the shared fiber link connecting the ONUs to the OLT. To this end, in general, a medium access control (MAC) protocol should be exploited. More specifically, the bandwidth demand of each ONU is reported to the OLT. In the next step, the OLT assigns UL transmission time windows based on a dynamic bandwidth allocation algorithm [7].

Moreover, SDN has been proposed as the promising technology for achieving the desired new abstraction layers. Furthermore, by leveraging SDN, the transformation from the current inflexible, vendor-locked network infrastructures into agile and programmable platforms will be feasible [7, 8]. As is illustrated in Fig. 14.1, there are three different layers in SDN abstraction including the infrastructure layer, the control layer, and the application layer [9]. The northbound interface (NBI) is considered for providing the connection between the application and control layers. On the other hand, the southbound interface (SBI) is considered between the control and the infrastructure layer. As an example of the SBI, we can mention the OpenFlow (OF) protocol [10]. Considering ARoF in the converged optical-wireless fronthaul enables an optimal use of resources. The resource management and allocation within the network are offered by a virtual orchestrator based on SDN’s centralized control plane and NFV [11].

14.2.1 SDN-Based PONs

The management task of PONs is frequently done by non-flexible management systems. In the recent years, as SDN provides more efficient management facilities, the integration of SDN in PONs is becoming more significant. In the current section, we discuss different proposed SDN solutions for PON architectures and scenarios in the literature. The authors in [12] take the advantage of SDN accurate quality of service (QoS) control and propose a management system for gigabit PON (GPON) network services with OpenFlow. It was done by employing an external OpenDayLight (ODL) controller and several OF virtual switches utilized as the OLT and the ONUs. In their architecture, the controller manages the DL and UL traffic channels in order that the services meet the required QoS inside the GPON. Apart from providing the QoS requirements to the users, the proposed testbed brings more flexibility and dynamicity to the control part of the network.

In [13], a software-defined (SD) energy-saving time wavelength division multiplexing PON (TWDM-PON) is proposed, which enhances existing PON by taking advantage of SDN properties. To this end, the authors consider SD PON's elements instead of the legacy ones. By deploying essential applications such as SD-controller and energy-saving, the TWDM-PON network is able to alter the wavelength, link rate, packet-classification, queue threshold, and QoS configurations of the devices. These applications based on the traffic or operator's requirements provide independent, but, at the same time, coordinated energy-saving approaches. More specifically, the proposed SD applications define the energy-saving thresholds, which are being used by the devices to decide when to change their energy-saving modes.

Khalili et al. in [14] propose a new SDN-based architecture for service interoperability for Ethernet PON, which decouples, virtualizes, and simplifies the OLT in terms of management and operation. In addition, their novel architecture, apart from multi-tenancy, provides an optical access network in which various service providers exploit the shared infrastructure. More specifically, some of the control plane-related tasks of the OLT are moved to an OF controller, while the data plane (forwarding tasks) part of the OLT is built around an OF switch, which is more in charge of the Ethernet PON (EPON) service path functions. The OF switch conducts the following tasks: (i) classifying incoming packets based on the matching rules, (ii) adjusting packet header fields if needed, (iii) flow's QoS scheduling, and (iv) packet forwarding tasks. To this end, the OF switch should emulate a number of the OLT's tasks, such as control multiplexer, control parser, and multipoint transmission control.

The authors in [15] have proposed an SDN-based EPON architecture, which enables OF on a group of access technologies including point to multipoint devices by applying small modifications. As is shown in Fig. 14.2, the primary element in their architecture is the hardware abstraction layer (HAL), which is in charge of linking the OLT management port. The main task of HAL is mapping port pairs/VLAN on the physical system. In other words, the HAL behaves as an

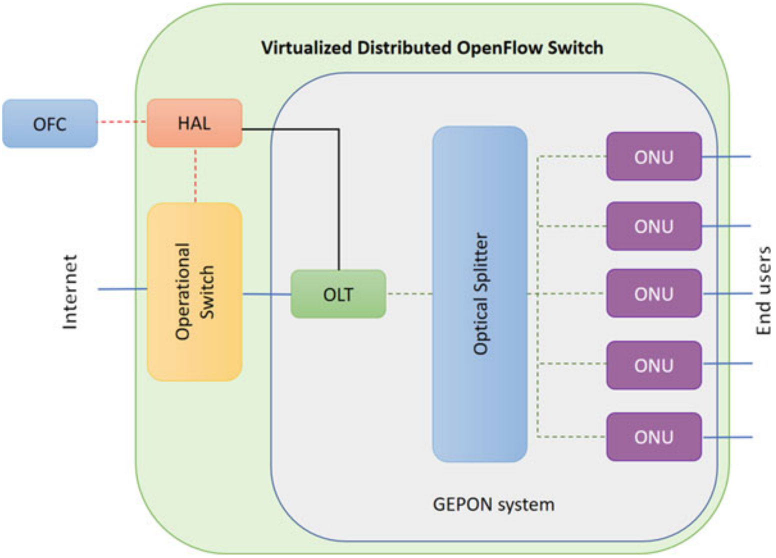


Fig. 14.2 The proposed architecture in [15] for SDN-based GEPONs

intermediate OF controller (OFC). The external OFC, which is directly connected to the HAL, merely sees a single distributed switch with a number of ports. Among these ports, one is for connecting to the OLT and one for each ONU. For transmitting the OF commands to the gigabit EPON (GEPON) switch from the external OFC, the HAL changes the message to determine which port it is related to. For the opposite OF commands transmissions, it is in charge of mapping the message to the corresponding port.

14.2.2 SDN-Based Converged Optical-Wireless Fronthaul

As was mentioned in the previous section, 5G network is expected to provide massive device connectivity and high throughput, and simultaneously should minimize the latency, costs, and the power consumption. To achieve the abovementioned goals, apart from employing the SDN-based PONs, 5G utilizes the mmWave frequencies, which are mainly around 28 GHz and 60 GHz, instead of the sub-6-GHz radio spectrum. As two main advantages of mmWave frequencies, we can mention supporting of unlicensed operations and providing broader user channels, up to 400 MHz (according to the 3rd generation partnership project (3GPP) Rel. 15 standards) [16].

The integration of PONs and mmWave frequencies forms an extremely efficient fronthaul infrastructure for 5G mobile networks and provides a solid connectivity for high number of users and guarantees QoS in terms of data rate and latency [17, 18]. All the advantages of the SDN-based PONs can be also realized in a converged optical-wireless network utilizing mmWave, provided that the problem of being bandwidth consuming of the currently deployed common public radio interface (CPRI) standard is solved. One of the most promising fronthauling schemes to tackle this problem is employing low-layer transport schemes, such as ARoF [16, 19].

ARoF has recently emerged as a more spectrally efficient and cost-effective transport technique capable of fitting enhanced traffic capacities on an intermediate frequency (IF) carrier using low-cost intensity modulation/direct detection schemes. At the same time, it supports efficient frequency aggregation and de-aggregation of multiple radio streams on IF using digital signal processing on fully programmable gate array baseband unit prototypes [16, 20, 21].

5GSTEP FWD project [22], which is funded by the European Commission, aims at proposing a new optical-wireless networking solution to provide high-speed connectivity to end users. The proposed ring architecture in 5GSTEP FWD interconnects a number of ultra-dense WDM-PONs (UDWDM-PONs). In this structure, the OLTs are allowed for interchanging information by using different reconfigurable optical add drop multiplexers [20]. The major idea of the 5GSTEP FWD is integrating the mm-wave RoF links with the UDWDM PONs. More specifically, the combination of the ARoF fronthaul with the flexible carrier aggregation provides an optimal use of network resources either in the optical domain or wireless domain. Two promising tools to perform the resource management and allocation tasks in such networks are SDN and NFV. More specifically, a virtual SDN controller conducts the resource management task in this architecture [11, 22].

14.3 SDN-Based Dynamic Network Service Provisioning for 5G and Beyond

As defined by 3GPP, 5G system architecture is constituted by a core network (CN) and one or more access networks, for instance a radio access network (RAN) or an optical access network. The role of CN is to serve heterogeneous mobile networks and carry the data that is exchanged among different services. It consists of network functions (NFs), NF services, and the interaction between them, while enabling deployments based on SDN concepts [23].

SDMN has been proposed to enhance the performance of core and RAN through advanced joint coordination of resources, spectrum, and mobility, and by boosting the cooperation among converged networks. For the optical access domain though, another SDN-based approach software-defined access (SDA) has been proposed specializing on the convergence between access and transport networks [24]. This integration brings extra benefits to the network such as decreased operational expenditure (OPEX) due to programmable and automated operations, advanced resource utilization with guaranteed QoS, flexibility in fronthaul with flexible functional splitting, scalability, and resiliency with multiple logical isolated network instances (slices).

As a result, SDN controllers – the major element of SDN architecture –have started integrating in various network parts for central control, and network elements have been installed with agents interacting with the controllers, making them feasible for programming. Examples are 5G PPP phase 1 and 2 projects like BLUESPACE [25], which introduces child SDN controllers for the backhaul and fronthaul segments, while one parent SDN controller on top is acting as the transport network controller [26]. To handle the optical resources of the data plane on the fronthaul segment, SDN node agents are running at cell sites and the central office (CO). Another project, working on the converged optical-wireless network solution able to flexibly connect small cells to the core network, is 5G-XHaul [27]. There, by developing a software-defined cognitive control plane, it is able to forecast traffic demand in time and space, and with inherent ability to reconfigure network components to allow the dynamic allocation of network resources to actual and predicted hotspots.

Moreover, users with multiple requirements and diverse traffic profiles are adopted to SDN-based converged optical-wireless networks to optimize the usage of the physical infrastructure through virtualization and resource sharing techniques, while guaranteeing high levels of flexibility in the provisioning of dedicated services with customized QoS. Section 14.3.1 describes flexible functional splits in SDN-based networks and Sect. 14.3.2 shows how SDN can enable slicing in a multi-5G environment.

14.3.1 Flexible Functional Split Selection for Converged Optical-Wireless Networks

Although great advantages can occur by utilizing the traditional split between base band unit (BBU) and radio remote head (RRH) for all the connections, such as pool shared resources and less complex and scalable remote units, at the same time it has very strict latency requirements and high bit rates. Thus, a variety of different functional splits are presented by different names and targeting different flows [28]. The functional split decides how many base station (BS) functions to localize closer to the end-user, promoting benefits such as multiplexing gain on the fronthaul and relaxing bitrate and delay requirements, and how many functions to leave centralized with the multiplexing gain on the central office harnessing greater processing benefits and more shared resources [29].

Distributed RAN and centralized RAN (C-RAN) are the two extremes in a plethora of heterogeneous deployments in 5G networks. 5G diverse service requirements redefined RAN functions and gave birth to a new degree of freedom: functional splitting options. This enables operators to determine the centralization level or functional split of baseband processing functions for each RRH. There are

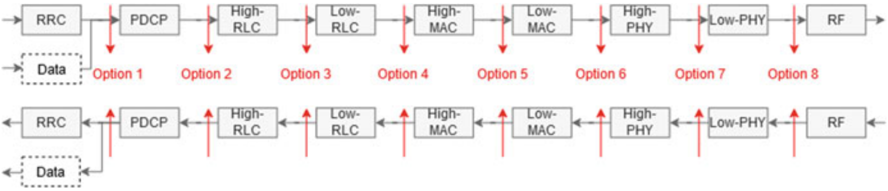


Fig. 14.3 3GPP functional split between central and distributed units

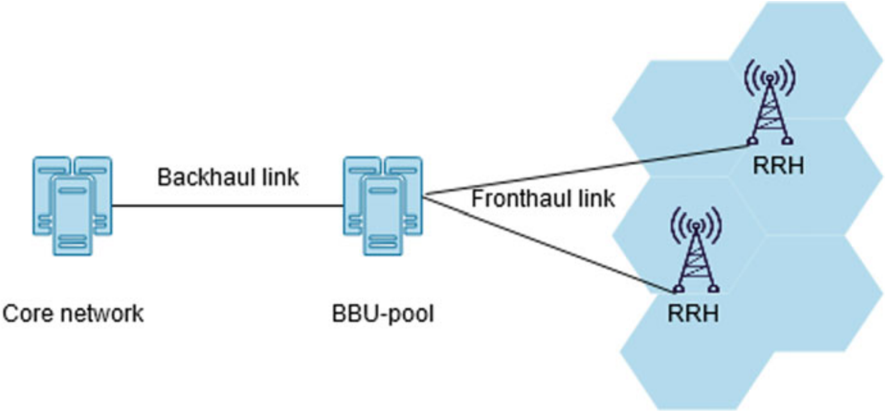


Fig. 14.4 Network overview illustrating backhaul and fronthaul links

several levels of centralizations and major standardization bodies such as 3GPP [30], enhanced common public radio interface (eCPRI) [31], *Small-Cell* forum [32], and the next-generation mobile networks (NGMN) [33] that investigate different options for functional splits. Figure 14.3 illustrates the work and the options of 3GPP.

Moreover, the evolution of RAN part from analog to digital signal processing units and from dedicated hardware components to the general-purpose processor with software-defined radio (SDR) systems [34] is now extended with virtualization through abstraction of the execution environment. Therefore, radio functions turn to a general-purpose application that runs on top of a virtualized environment interacting with physical resources [35]. By employing the technologies of SDN and NFV for RAN [36], controllers manage virtualized RAN functionalities and optimize it on demand. Figure 14.4 shows such a network architecture but with the control and data plane unified. In order to bring higher elasticity to RAN, vertical splitting (decouple control from data plane) and horizontal splitting with the concept of flexible functional splits in different service demands have been considered, since bandwidth and delay constraints vary with each functional split [37].

Flexible architectures or functional splits require appropriate advanced resource allocation mechanisms in both mobile, access and core networks. A possible solution to the physical fronthaul segment is by deploying PONs due to the high availability in connecting small cells and low OPEX because RRH can share the

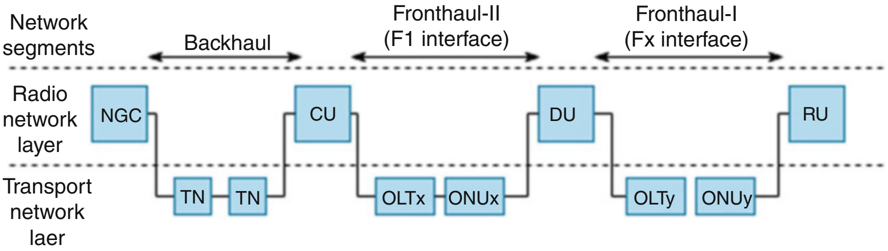


Fig. 14.5 Concept of layered structure showing radio network layer and TNL (OLT, ONU) by NGFI

optical fiber that connects to the BBU [38]. Figure 14.5 presents how a PON can support 5G new radio (NR) fronthaul and how the elements of network radio layer (NRL) can be mapped to the elements of transport network layer (TNL) in PON.

In 5G, a main change is that the original BBU is now split into three parts as defined in TR 38.801:

- Centralized unit (CU);
- Distributed unit (DU);
- Radio unit (RU).

This redesign of the functions of BBU leads to the exploitation functional splits. As a result, this redesign allows also to insert a new network segment between backhaul and fronthaul, the so-called midhaul:

- **Backhaul**: Connection from 5G core to CU or the integrated CU/DU/RU;
- **Midhaul**: Connection between CU and DU;
- **Fronthaul**: Connection from CU or DU or integrated CU/DU to RU.

Similarly, next-generation fronthaul interface (NGFI) divides fronthaul into two fronthaul segments Fronthaul-II between CU and DU, and Fronthaul-I between DU and RU [39]. As a result, standardization bodies are focusing on determining a high layer (localized functions) functional split and one low layer (centralized functions) split between CU and RU, which tends to be more practical to build dedicated PONs specifically to provide wireless services in order to avoid any degradation to fixed user services.

In the current state of the art, most research efforts have been assigned to primarily the lower layer splits, i.e., split 6–8, but recent papers are focusing more on the opportunity of flexible functional splits. In theoretical surveys, flexible functional splitting is also described as RAN-as-a-service (RANaaS) in which RAN functionality is flexibly centralized, based on a cloud infrastructure [40]. In the RANaaS implementation, all RAN functionalities are flexibly centralized. As a result, RANaaS can choose an optimal functional split between the full centralization and local execution on the radio site. In [41] a new RAN concept referred to as flexible RAN (F-RAN) is presented with DWDM fronthaul transport, in which baseband processing functions are strategically distributed within the

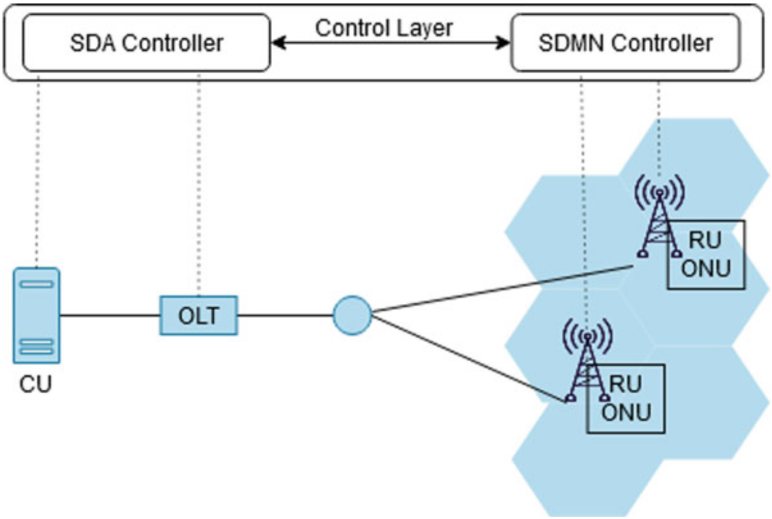


Fig. 14.6 Converged SD-WBA and flexible functional split [43]

RAN in order to optimize the trade-off between radio performance maximization and transport capacity requirement minimization. The performance evaluation of F-RAN concept shows a better utilization of transport resources compared to conventional C-RAN. Another approach in RAN optimization can be found in [42], where the SDN principle of decoupling control and data planes can be implemented in fronthaul networks in cloud RAN and its potential and challenges are discussed.

Moving to simulation-related functional splitting works, in [43] a software-defined converged approach to manage the adaptive flexible functional split in 5G networks is proposed upon an optical access network. Optimal split selection among the options 2, 6, 7, and 8 is achieved with regard to bandwidth management in the optical domain. At the optical domain, as shown in Fig. 14.6, the SDA controller communicates with the SDMN controller informing about the available bandwidth amount in the PON. From the mobile network perspective, the SDMN controller performs a functional split calculation algorithm that determines the optimal functional split based on the currently available bandwidth. A fronthaul command is sent to the SDA controller informing the requirements in terms of bandwidth and maximum latency. The SDA controller also interacts with the agents at the OLT to run an appropriate bandwidth and wavelength allocation scheme for the ONUs serving the RUs.

Conversely, compared to the above-mentioned cell-centric works, [44] demonstrates a user-centric approach that jointly optimizes and orchestrates the heterogeneous radio, link, and computational resources, leveraging the SDN and NFV capabilities. The split decision is occurring on a user basis, optimizing the heterogeneous resource usage outperforming cell-centric methods. Key role for this

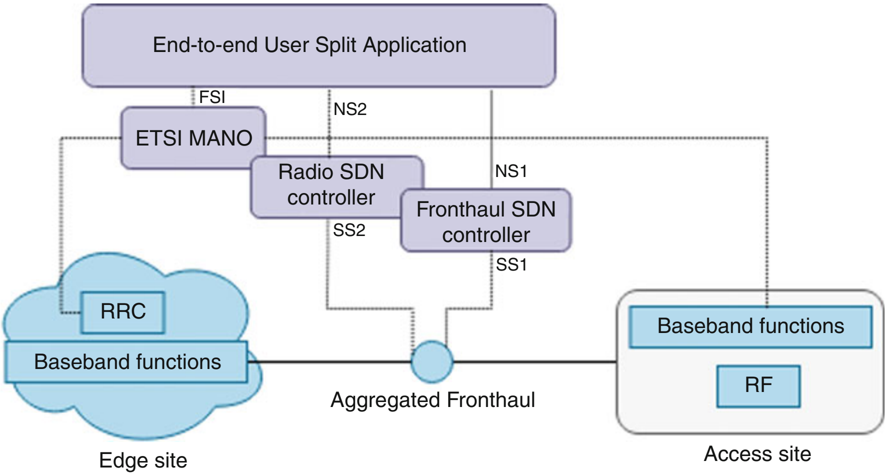


Fig. 14.7 End-to-end user split infrastructure [44]

approach is the integration of SDN controllers handling the radio and fronthaul resources, leveraging its centralized and abstract network view.

As shown in Fig. 14.7, two SDN controllers, a fronthaul and a radio SDN controller, are managing the link and radio resources, respectively.

First, the optimal decision for user functional split is communicated to the orchestrator through the functional split interface. Then, through the NBI1, the fronthaul SDN controller is activated and via the SBI1, it allocates link resources in the aggregated fronthaul network. Once the virtual infrastructure is established, the decision for radio resource allocation is sent to the radio SDN controller through the NBI2 interface. The latter, via SBI2 interface, coordinates the radio resource control (RRC) layer of many cells. Further details about this architecture are provided in [45].

Besides the simulations, practical experiments demonstrate a flexible functional split with OpenAirInterface (OAI) platform [46]. OAI is a flexible experimentation SDR platform by EURECOM which is fully open-source and functionalities of a transceiver like base stations can be implemented with a software radio front end connected to a host computer for processing. The work in [47] focuses on the lower splits, option 7.1 and option 8 for both ARoF and digital radio-over-fiber fronthaul. ARoF integration can reduce the latency by more than 15% to support ultra-reliable and low-latency communication (URLLC) applications. More works illustrating the potential of using OAI software are FlexRAN in [48] and FlexCRAN [49]. The former work demonstrates the capabilities of FlexRAN and acts as an open-source software-defined RAN platform, enabling centralized and distributed schemes of service and baseband functions. FlexRAN decouples the RAN control and data planes via a custom-tailored southbound API. The application of [44] that was mentioned above on simulation-related works relies on such FlexRAN SDN controller for radio resource management providing an API for RAN controlling over multiple RUs. The latter approach, FlexCRAN, acts as a flexible functional split framework over Ethernet fronthaul built based on OAI, evaluating various scenarios by considering fronthaul, RRH/BBU processing, and data plane-related key performance indicator (KPI) measurements.

14.3.2 Network Slicing for Multi-Service Environment

In previous mobile network generations, increased data rate was the vital point. Although the need for increased data rates continues to be of greatest importance moving forward to 5G, the requirements are far more diverse than anything before [50]. 5G is pictured to maintain multiple services such as high-definition video streaming, digital health services and control automation (for instance industrial), and the services are classified mainly into three types by international telecommunication union (ITU); enhanced mobile broadband (eMBB), URLLC, and massive machine-type communication (mMTC) [51]:

- Enhanced Mobile Broadband: Mobile Broadband addresses the human-centric use cases for access to entertainment services, and data such as mobile ultra-high definition video, augmented and virtual reality with peak data rates at home and on the move. The demand for mobile broadband will continue to increase, leading to enhanced Mobile Broadband. Also, eMBB offers extreme capacity, enhanced wide-area coverage, uniformity, and deep network awareness.
- Ultra-reliable and low latency communications: This use case is mainly characterized by ultra-low latency (at millisecond levels) with further strict requirements for throughput and availability. Examples include eHealth, autonomous vehicles, and industry automation. This use case is also known as critical IoT too where monitoring and control occur in real time, and the need for mobile reliability is essential.
- Massive machine type communications: This use case can also be addressed as massive internet of things (IoT) and a key characteristic is the vast number of connected devices typically transmitting a relatively low volume of non-delay-sensitive data. Devices are required to be low cost and have long-lasting battery. Examples include smart cities, smart building, smart agriculture, transport and logistics, smart agriculture, logistics, tracking, and fleet management.

Figure 14.8 depicts the three main requirement dimensions: throughput/capacity, number of devices/low cost, and latency/reliability. Some use cases might require more than one dimension for optimization where others focus only on one KPI.

One of the main challenges for 5G will be to support such diverse use cases in a flexible and reliable way. Table 14.1 contains a high-level overview of the expected traffic characteristics for the various 5G services.

With the myriad of use cases in 5G and RAN evolution, the management plane must adapt accordingly. With high QoS demand, the management of network

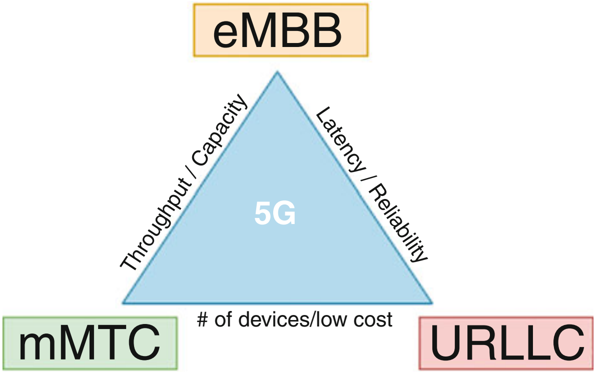


Fig. 14.8 ITU 5G use cases

Table 14.1 High-level overview of expected traffic characteristics for various 5G services from ITU

Radio technology	Peak rate	Average rate	e2e delay (service level)
Enhanced Mobile Broadband (eMBB)	5–10/20 Gb/s (UL/DL)	100 Mb/s per user in urban/suburban areas 1–4 Gb/s (hot spot areas)	10 ms
Ultra-reliable and low latency communications (URLLC)	Much lower than in eMBB: N × Mb/s	Much lower than in eMBB: n × Mb/s	1–2.5 ms
Massive machine type communications (mMTC)	Much lower than in eMBB: N × Mb/s	Much lower than in eMBB: n × kb/s – n × Mb/s	1–50 ms

resources becomes a challenging task. A key driver to induce flexibility on the management domain is network slicing [52], by dividing one physical network into multiple logical networks, referred to as network slices. A slice is a pool of network resources (physical or virtual), selected to satisfy the demands of a service in which the slice delivers. In this way, one slice can have specific capabilities related to one specific service or 5G use case, as Fig. 14.9 illustrates.

However, 5G envisions high data rates and lower communication delays, leading to various services, co-existing all together. To simultaneously support eMBB, URLLC, and mMTC service, even more flexible deployment solutions need to be implemented. In addition to flexible software-defined and virtualized fronthaul, network slicing paradigm offers such extra and additional flexibilities and SDN is one of the key enablers to achieve the realization of network slices. Indeed, many 5G research and demonstration projects (such as 5GSTEPFWD, 5GSTEPFWD [53], 5GNORMA [54], 5GTRANFORMER [55]) are addressing the realization of 5G slicing through the principles of SDN.

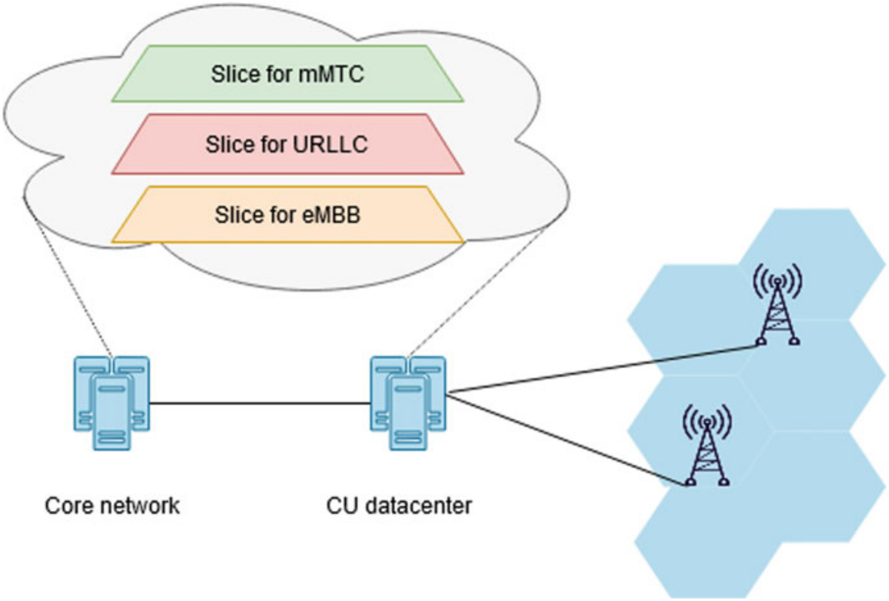


Fig. 14.9 5G network slicing architecture [52]

According to open network foundation (ONF) [56], the SDN architecture is an appropriate tool for supporting the key principles of slicing. This is because 5G network needs to satisfy a wide range of service requests in an agile and cost-effective manner. By separating the control from the data plane, the resources from the underlying forwarding plane can be dynamically configured and managed to deliver tailored services to clients located in the application plane. Therefore, SDN controllers, which are the basic components, act as mediators between resources and clients, enabling a server-client relationship [57], as depicted in Fig. 14.10. This server-client relationship consists of two conceptual components, the server and client contexts. At the server context, an SDN controller interacts with all the underlying resources (infrastructure resources and network functions) and assembles them into a Resource Group, through one of its southbound interfaces. When an end-user (client) requests a service, another SDN controller in the client context abstracts and customizes the resources in order from the Resource Group to deliver the service through one of its northbound interfaces.

The SDN architecture also includes an administrator to configure the controllers, create the contexts, and install their associated policies. As a result, the ONF SDN network architecture enables slicing in terms of client context, as this provides all sets of resources (as Resource Group) in an abstract way.

To sum up, network slicing in 5G systems can support the multifaceted cases and SDN provides the tool to assist slicing. Nevertheless, network slicing can also arise issues such as security issues, since open interfaces that support the programmability

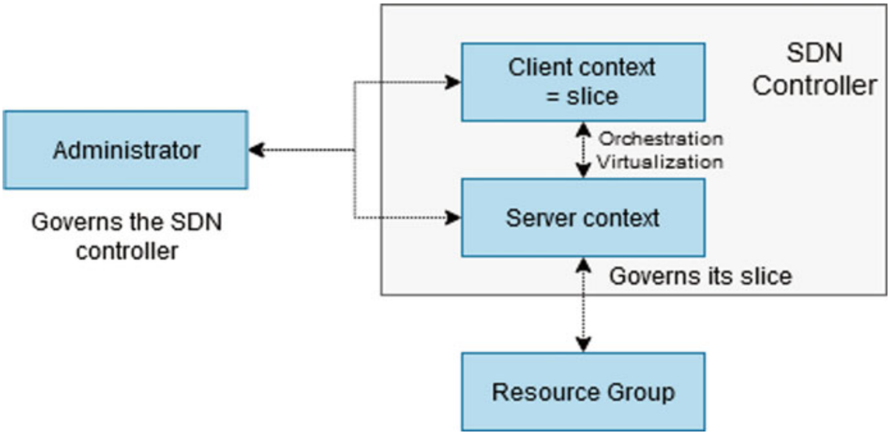


Fig. 14.10 ONF SDN Network Slicing architecture [58]

of the network bring new potential attacks to softwarized networks as well as management issues given by the dynamism and scalability that slicing brings. Furthermore, since the slices are deployed over a common and shared infrastructure, functional split must dynamically adjust service demands.

From a network slicing point of view, it is preferred to use split options that the fronthaul load is user-dependent and runs over a packet-switched network. Taking into consideration the physical and functional architecture of 5G and the drivers on softwarization and virtualization such as SDN and NFV respectively, different use cases will benefit from different functional splits [59]. Hence eMBB with high transmission rates will benefit from a high degree of centralization on the fronthaul to implement the efficient allocation of resources and coordinated multipoint. Moreover, in the URLLC case due to strict delay requirements, a network with good traffic management is required and will benefit from a centralized MAC scheduler to determine faster optimal routes. In contrast, mMTC will benefit from variable bitrate on the fronthaul as it has relaxed bitrate requirements, but because of the large numbers of devices at a single CU, centralizing some functionalities could achieve efficient utilization of computing resources.

The work in [60], implements an architecture, in which slices are constructed for the three main service types, and chooses an appropriate functional split option, considering bandwidth and delay requirements. This approach achieves high flexibility and large scalability by utilizing an open network operating system SDN controller [61], which manages the network configuration among network components. In conclusion, the one network needs to be compatible with all the requirements in all slices, in such a way that different functional splits will be used for different slices, in order to enhance system's performance and resource utilization.

14.4 Service Provisioning Use-Cases and Design Guidelines

Moving from theory to practice, the last part of this chapter contains two use-cases addressing the above-mentioned issues related to network slicing and functional splitting. First, we attempt to optimally determine the network slices so that the specific delay and bandwidth requirements of multiple services are met, by considering both the optical and wireless networks resources over an SDN-based resource management scheme.

The second use-case intends to highlight the flexibility in an SDN-based converged optical-wireless fronthaul in terms of splitting the baseband and radio processing functionalities between a central office and distributed entities, and the impact of this flexibility on delivery of functional split as a service (FSaaS) by meeting user's requirements. Two 5G services have been modeled (eMBB and mMTC) over an optical-wireless converged network, and the impact of each of these services on different functional split options is examined. In addition to slicing the network, an appropriate functional split option is chosen, based on an SDN split decision controller.

14.4.1 Network Slicing Use-Case

In this section, we discuss our work [62] in which an SDN-based PON configuration has been taken into account. In the mentioned work, the goal is to jointly manage the resources in the optical and wireless resources of the converged network in an efficient way. Firstly, we divide the network into a number of isolated network slices each with different QoS requirements. Afterward, the communicational resources assignment task to each slice is done by the virtual SDN controller, which is installed in a server at a data-center. The task of the virtual SDN controller is satisfying the QoS requirements of the user equipment (UEs) at each network slice.

14.4.1.1 System Model

A converged optical-wireless network fronthaul is considered. There are M ONUs, which are present at the same places as M RRHs. As Fig. 14.11 illustrates, the ONUs are connected to a virtual OLT (vOLT) through fiber links. A WDM-PON approach is considered, where the vOLT assigns different wavelengths to every single ONU/RRH by using a wavelength router. From the SDN point of view, the vOLT installed as a software in one of the data-centers is the control-plane part of legacy OLT, which manages the wavelength assignment task of the router, while the router is assumed as the data-plane entity of the legacy OLT. Furthermore, the available bandwidth of each RRH is indicated as W. Each RRH serves either the connected UEs or small cells supporting a number of UEs. We consider an ARoF

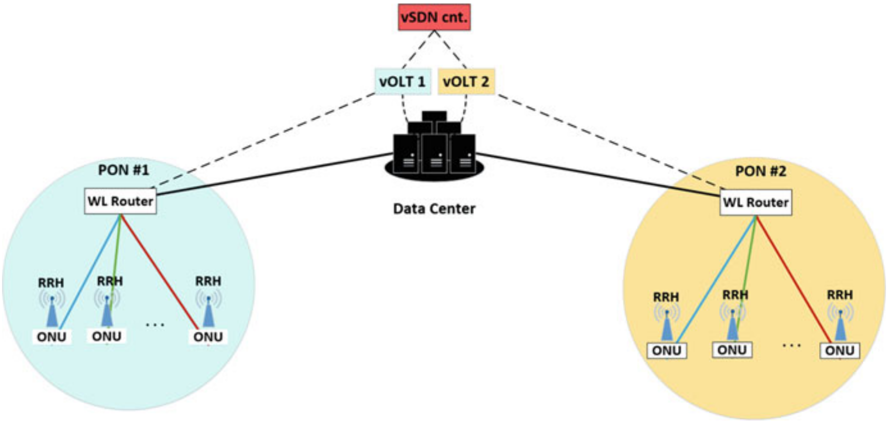


Fig. 14.11 System architecture

approach for transmitting the data via fiber links; hence, there is exactly the same available bandwidth in the optical domain. Thus, we can consider *C* channels with equal bandwidth of *B* (*W*=*C***B*) in both optical and wireless domain in the UL direction. Additionally, we divide the network into *K* different E2E slices.

The UEs belonging to one slice, which are served by a specific RRH, generate UL traffic related to the slice characteristics in terms of throughput and latency. It is assumed that the generated traffic by each UE follows a Poisson distribution. Consequently, the total UL traffic of each slice follows a Poisson distribution with the arrival rate of $\lambda_{m,k}$, where $m = 1,\ldots,M$. Regarding the role of NFV in the considered system model, all the physical resources including CPU, memory, and storage form the NFV infrastructure (NFVI). Apart from this, both the ubiquitous network functions such as routers and switches and the common network applications including firewall and load balancer can be virtualized on the top of the NFVI as softwares. In order to create a service function chaining (SFC), the input traffic should be steered through a set of ordered virtual network functions (VNFs), which are interconnected through virtual links. For a specific SFC, the VNFs' type and their order are completely predefined.

On the other hand, concerning the SDN part, the virtual SDN controller (vSDN controller) is in charge of managing the corresponding vOLT and the virtual BBU for every single RRH in the network, and also interconnects different VNFs of the SFC. Furthermore, the forwarding tables of the networking devices, which support the OF protocol, are installed by the vSDN controller. By perceiving specific QoS thresholds for the users of each slice, we attempt to satisfy the QoS requirements of different services.

To this end, the different minimum required data rates and maximum acceptable E2E latency are set for the services of each slice. It is assumed that the vSDN controller receives the real-time statistics of all the VNFs and slices, thus it can check the defined thresholds in real time. To be more specific, provided that the threshold of a specific slice is exceeded, the vSDN controller applies the required changes to the resource allocation policies in order that the QoS requirements of all slices are met. In the case that the statistics from the slices are not sent to the controller for any reason, the bandwidth channels will be equally assigned to all the slices until the controller receives the required data again for making more efficient decisions.

14.4.1.2 Simulation Results

In this section, the simulation parameters and results of the considered system model of a converged optical-wireless configuration are presented. Furthermore, we assess the performance in terms of the experienced average throughput and E2E latency experienced by the user. We compare our results with a non-SDN based network, in which similar amounts of communications resources are assigned to different slices with various QoS requirements.

The total length of the fiber link starting from the ONU to the vOLT inside the DC is considered to be $d = 40$ km and the speed of light in the fiber is defined as $\mathrm{C_f} = C/n$, where C is the speed of light in the vacuum and $n = 1.4475$ is the refractive index of the fiber. We employ the NR operating band n258 with 24.25 GHz as lower band and 27.5 GHz as the higher band. Thus, the total available bandwidth at each RRH will be $W = 3.25$ GHz and by deploying suitable modulation techniques, we would have total data rate of $R = 10$ Gbps. By considering channel bandwidth $B = 100$ MHz, a total of $C = 32$ channels are available for the UL direction.

A converged network with $M = 1$ RRH is deployed, which supports a maximum of 50 UEs or small cells connected to it. Moreover, three slices exist in the network ($K = 3$). First slice ($k = 1$) provides service to the users with data rate of 350 Mbps and maximum acceptable delay of 10 ms, while slice $k = 2$ offers medium amounts of throughput (100 Mbps) and 10 ms delay. Finally, users in the last slice ($k = 3$) have strict latency requirements (less than 2 ms) and minimum required data rate of 50 Mbps. Moreover, the first slice supports maximum 20 UE, while slice $k = 2$ has 25 active groups of users connected to the RRH via different small cells/lampposts, which results in a massive number of connected devices. The number of UEs in last slice ($k = 3$) is considered to be up to 5 UEs. Each user belongs to merely one slice. We vary the generated UL traffic by each UE/small cell from 5 Mbps to 350 Mbps. At the beginning, the number of channels assigned to each slice is proportional to the corresponding number of UEs of the slice. In all simulation results, dashed lines present non-SDN approach while the solid lines refer to the SDN-based approach.

Figure 14.12 shows the comparison between the SDN-based and non-SDN-based resource management methods, and their impact on data rate and latency of UE in a particular slice.

Figure 14.12g and h show the non-SDN-based-non-sliced case, in which all the resources are equally allocated between all the UEs and consequently they experience the same amount of data rate and latency. In Fig. 14.12a and b, by

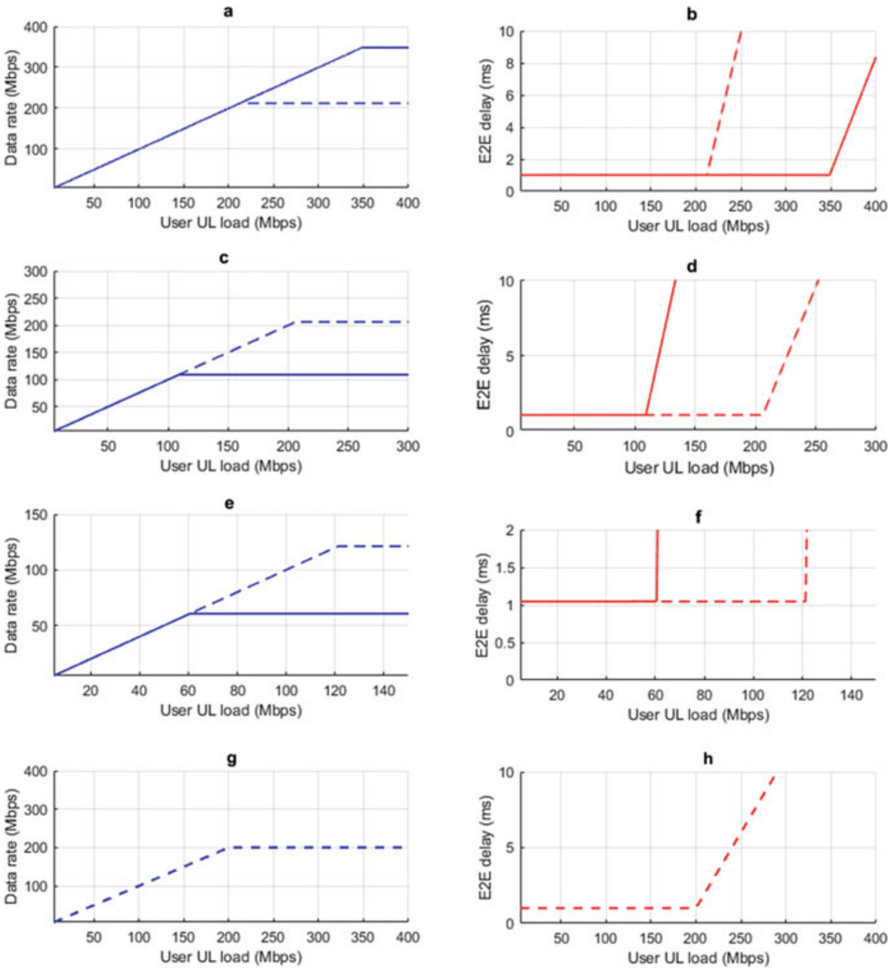


Fig. 14.12 SDN-based vs. non-SDN-based resources management in a sliced optical-wireless network

applying the network slicing concept to the network (dashed plots), UEs in slice $k = 1$ have the maximum data rate of around 212 Mbps with guaranteed latency below 10 ms. Although, by dividing the network into different slices the data rate improves, it is still less than the slice requirements. In contrast, for the second and third slices (Fig. 14.12c and d for slice $k = 2$, and Fig. 14.12e and f for the third slice), the experienced data rates for the UEs are much higher than the slice QoS requirements.

To this end, the controller steps in and checks the predefined thresholds for each slice, while the UL traffic load is increasing. It changes the number of allocated channels to the network slices. As is illustrated in Fig. 14.12, applying

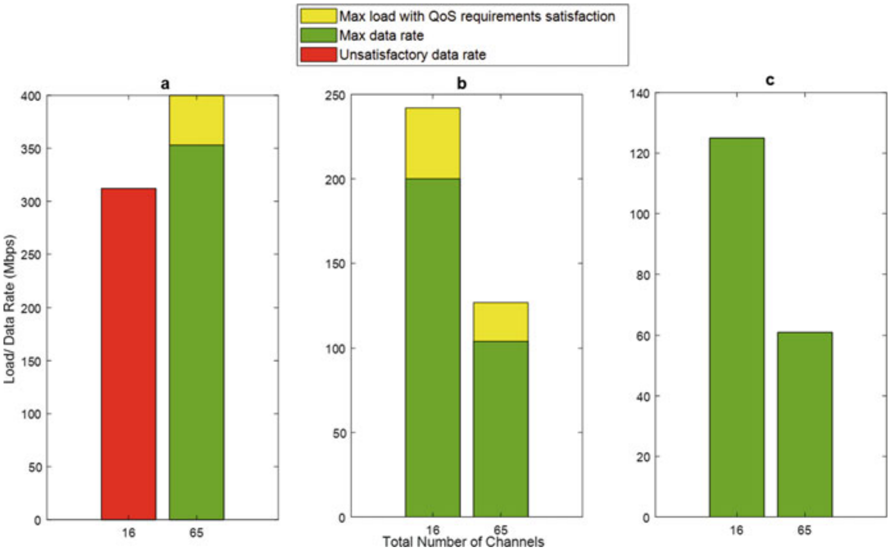


Fig. 14.13 Impact of the total number of channels (C) on the resource management in slice (**a**) $k = 1$, (**b**) $k = 2$, (**c**) $k = 3$

SDN to the sliced network results in significantly performance improvement. To be more specific, for slice $k = 1$ the data rate will saturate at 350 Mbps, while the latency remains below 10 ms up to UL load of 400 Mbps. Nevertheless, the UEs belonging to slice $k = 2$ lose around 100 Mbps of data rates, but still meet the QoS requirements of the slice compared with the non-SDN-based case.

Moreover, we assess the impact of the number of channels (C) on the performance of the slices. To do this, we perceive two values for the total number channels in each wavelength as $C = 16{,}65$ resulting in $B = 200{,}50$ MHz of channel bandwidths, respectively. In Fig. 14.13, all the values for the UE's load and the maximum UE's data rate are gained for the case that the UE's E2E latency is below 10 ms for slices $k = 1{,}2$ (Fig. 14.13a and b, respectively) and for the slice $k = 3$ less than 2 ms (Fig. 14.13c). As is depicted in the Fig. 14.13a, for $C = 16$ the user's data rate requirement serviced by the first slice is not satisfied (the red bar). On the other hand, for this amount of C, both data rates and maximum supported load per UE for the two other slices are even over-satisfied. By considering narrower the channel bandwidth (or equivalently by increasing the total number of channels to $C = 64$), the data rate and maximum supported load for the UE corresponding to the slice $k = 1$ increases and is fully satisfied. In contrast, the reverse trend for the other two slices is observed, while the QoS requirements of all services offered by all the slices are met. Therefore, it can be concluded that by increasing the number of channels in wavelength, our SDN-based resource allocation algorithm will be more efficient.

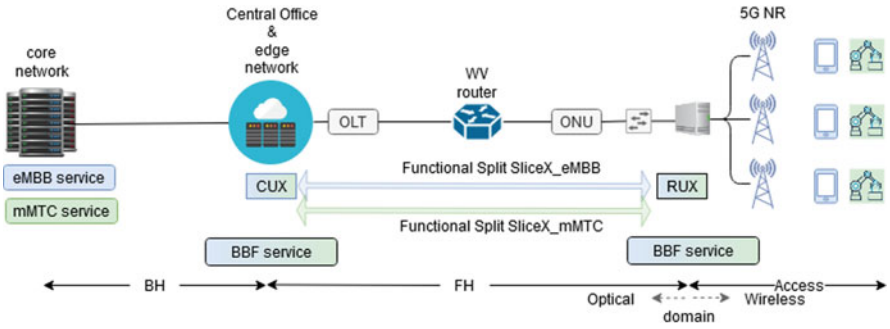


Fig. 14.14 Converged optical-wireless 5G architecture with functional split per slice

14.4.2 Flexible Functional Split Decision Use-Case

Flexible network architecture is envisioned as one of the features of the next-generation mobile networks able to support diverse 5G services, such as URLLC, mMTC, and eMBB traffic profiles. To address the requirements of a multi-service environment, two enablers flexible functional split and network slicing techniques hold a key role. This case study intends to study the flexibility in a SDN-based converged optical-wireless configuration in terms of splitting the baseband and radio processing functionalities between a central office and distributed entities, and the impact of this flexibility on the delivery of 5G services.

As Fig. 14.14 illustrates, users with different traffic profiles and requirements are divided into different slices (service-based slice). Users request a service, which can be located either in the core or edge network. The service to be delivered has to meet the user's requirements in terms of bandwidth and depending on the fronthaul load, an SDN controller who manages both radio and optical resources decides an optimal split of the baseband functionalities between the CO and RU per each slice.

From the above-mentioned network example, the wireless and optical domains can be clearly identified. The former is the last leg between 5G NR and the end users, in which the admission control is responsible to accept or reject the user, and the latter contributes to an optical mobile fronthaul, connecting all the 5G type base stations to a central office through fiber links. Base stations on next-generation mobile communications have a smaller footprint and run a set of functionalities enabling different functional split options on the fronthaul. Thus, after a user is accepted based on the available radio resources in the access part, the fronthaul generates various bitrate loads in the DL depending on which split option is enabled. An SDN controller is necessary to determine which split option should be activated to meet user's requirements. The use-case is divided into three phases for simplification:

Table 14.2 Simulation configuration

Variables	Values	
Service-based slice	eMBB	mMTC
Arrival rate	6 users/min	3 users/min
Service time	3 min	1 min
Bandwidth requirement	100 Mbps	10 Mbps
Simulation time	20 min	

- Phase 1 – users' arrival, and admission control on the wireless domain;
- Phase 2 – calculation of DL fronthaul load for functional split options 6, 7, and 8 on the optical domain;
- Phase 3 – functional split decision by an SDN controller, with respect to users and PON bandwidth requirements.

In phase 1, the arrival of users follows the Poisson distribution and the interarrival time is modeled by an Exponential distribution; as in the modeling of teletraffic systems, there are many instances where the assumption is made that the offered traffic is a Poisson process [63]. Table 14.2 contains the values of the arrival rates and service times of eMBB and mMTC users, respectively.

Also, each service differentiates from the other on the downlink data rate requirement. Users that request eMBB service require much high data rates compared to IoT and sensors which need only a few Mbps. eMBB service time is higher as well, as a user approximately needs 3 min in the case of video streaming. Figure 14.15a shows the times the users of eMBB and mMTC service arrive to the system. If users are accepted (meaning there are enough radio resources), the service is provided and when completed, the users depart from the system (Fig. 14.15b).

A user to be accepted or rejected depends on the available resources on the access part. The bandwidth of the radio channel in the RU determines the maximum number of served users. In the interest of this case study, two scenarios have been investigated, the first with one radio channel of 40 MHz per RU and shared resources to both services, and the second scenario assigns two radio channels of 40 MHz in each RU. However, in the latter scenario where network slicing is enabled, only one radio channel is available for each service type. Rather than channel bandwidth, there are more RAN elements, which affect the maximum data rate on the downlink. Especially in 5G NR, with higher modulation formats and multiple-input multiple-output (MIMO) schemes, the maximum throughput is increased.

In 5G NR, the approximate data rate for one carrier in a band is computed as in (14.1) based on the 3GPP TS 38.306 standard [64], and Table 14.3 summarizes the configuration of the given use-case:

$$data\ rate\ \ (in\ Mbps) = 10^{-6} \bullet v_{Layers} \bullet Q_m \bullet R_{max} \bullet f \bullet \frac{N_{PRB}^{BW,\mu} \bullet 12}{T_s^{\mu}} \bullet (1 - OH) \tag{14.1}$$

wherein:

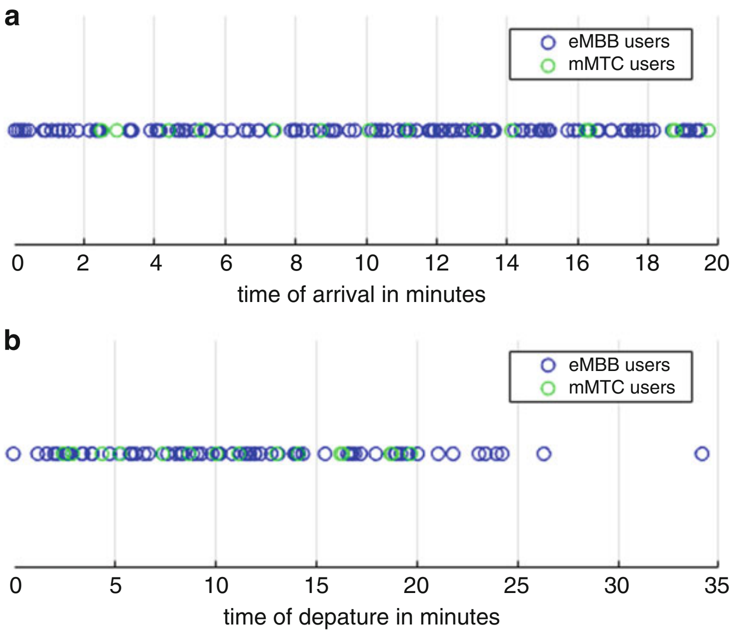


Fig. 14.15 (**a**) User's arrival time, (**b**) User's departure time

Table 14.3 RAN configuration

Variable	Value
Number of radio channels per RU	2
Radio channel bandwidth	40 MHz
Modulation format	256QAM
MIMO scheme	2×2
5G NR numerology	0
Frequency range in DL	FR1

v_{Layers} is the maximum number of supported MIMO layers.

Q_m is the maximum supported modulation order, quadrature amplitude modulation (QAM) format is used.

$R_{max} = 948/1024$.

f is the scaling factor and can take the values 1, 0.8, 0.75, and 0.4.

μ is the numerology as defined in TS 38.211 [65].

$N_{PRB}^{BW,\mu}$ is the maximum resource block allocation in bandwidth BW with numerology μ, as defined in TS 38.101 [66], where BW is the UE-supported maximum bandwidth in the given band.

T_s^{μ} is the average orthogonal frequency division multiplexing symbol duration in a subframe for numerology, i.e., $T_s^{\mu} = \frac{10^{-3}}{14\bullet 2^{\mu}}$,

note that normal cyclic prefix is assumed

Table 14.4 Fronthaul configuration

Split options	Baseband functionalities		Comments/observations/description
	Central office	Radio unit	
Option 8	RRC RLC MAC PHY	RF	User independent – Constant bit rate
Option 7	RRC RLC MAC PHY	PHY RF	User independent – Constant bit rate
Option 6	RRC RLC MAC	PHY RF	User dependent – Variable bit rate

Table 14.5 PON configuration

Entities	Quantity
OLT	1
ONU	1
RU per ONU	3
Downlink wavelength (λ_{wv})	1 ($BW_{\lambda\_wv} = 10$ Gbps)

OH is the overhead and takes the following values

0.14, for frequency range FR1 for DL
0.18, for frequency range FR2 for DL
0.08, for frequency range FR1 for UL
0.10, for frequency range FR2 for UL

Moving forward to phase 2, on the optical domain, according to the functional split option, various loads are generated on the fronthaul network. Split options 8 and 7 are user independent and have a constant bit rate because the resource element mapping function is executed in the CO and this function is necessary to detect unused subcarriers and thereby achieve variable bitrate. These options are cell-based, and again the RAN configuration imposes very high fronthaul load requirements. On the other hand, split option 6 is user dependent with variable bit rate. All the physical layer (PHY) now is in the RU closer to the user. The functionalities of each option can be shown in Tables 14.4 and 14.5 that contain information on the current use-case PON topology. In phase 2, the fronthaul bitrate is calculated by looking at what type of data is being transmitted on the fronthaul link, which is different depending on the functional split option.

The traditional split between the RF and physical layer (option 8) demands extremely high capacity on the fronthaul network and this leaves room for improvement. To state an example of the huge dynamic range in bitrate among the different functional splits, the fronthaul bitrates for the split options 8, 7, and 6 are illustrated in Fig. 14.16 and calculated based on (14.2, 14.3 and 14.4) from [67]:

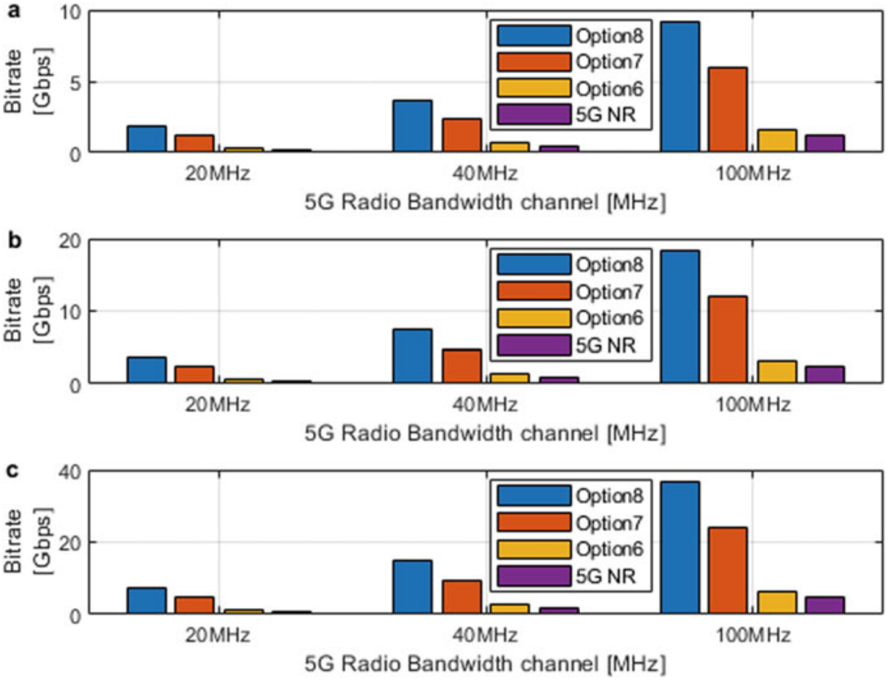


Fig. 14.16 Fronthaul bandwidth requirements versus 5G radio configuration with (**a**) 2×2 MIMO, (**b**) 4×4 MIMO, (**c**) 8×8 MIMO

$$FH\ load\ option\ 8\ \ (in\ Mbps) = f_{sampling} \bullet 2 \bullet bit_{width} \bullet N_{ports} \bullet \frac{BW_{ch}^{5G}}{BW_{ch}^{ref}} \tag{14.2}$$

$$FH\ load\ option\ 7\ \ (in\ Mbps) = bitrate_{option7}^{noOH} + OH_{option7}^{\%} \bullet bitrate_{option7}^{noOH} \tag{14.3}$$

$$FH\ load\ option\ 6\ \ (in\ Mbps) = bitrate_{option6}^{noOH} + OH_{option6}^{\%} \bullet bitrate_{option6}^{noOH} \tag{14.4}$$

wherein.

$f_{sampling}$ is the sampling rate 30.72 MS/s

bit_{width} is the radio over fiber bit-width of an I/Q symbol (15 bits in both downlink and uplink)

N_{ports} is the number of antenna ports (two for the considered scenario)

BW_{ch}^{5G} is the bandwidth in MHz of the 5G radio channel

BW_{ch}^{ref} is the reference LTE 20 MHz bandwidth

$bitrate^{noOH}_{option7} = \frac{10^{-6} \bullet 12 \bullet 2 \bullet bit_{width}}{T^{\mu}_{s}}$ is the bitrate of option 7 without the overhead information

$OH^{\%}_{option7}$ is overhead percentage for option 7 due to extra overhead from synchronization and the Ethernet frame and is approximately 8% DL overhead according to [67]

$bitrate^{noOH}_{option6} = bitrate^{ref}_{option6} \bullet \frac{BW^{5G}_{ch}}{BW^{ref}_{ch}} \bullet \frac{Q^{5G}_{m}}{Q^{ref}_{m}} \bullet \frac{v^{5G}_{Layers}}{v^{ref}_{Layers}}$ is the bitrate of option 6 without the overhead information

$bitrate^{ref}_{option6}$ is the bitrate of the reference LTE option 6 signal (211 Mbps)

$OH^{\%}_{option6}$ is the overhead percentage for option 6 due to extra overhead from scheduling control, synchronization, and the Ethernet frame and is approximately 14.1% DL overhead according to [68]

Q^{5G}_{m} is the modulation order of the 5G signal (256QAM for the considered scenario)

Q^{ref}_{m} is the modulation order of the reference LTE signal (64QAM)

v^{5G}_{Layers} is the number of MIMO layers (two for the considered scenario)

v^{ref}_{Layers} is the number of MIMO layers of the reference LTE signal (two)

It clearly demonstrates how the 5G NR further challenges the fronthaul bandwidth requirements, since a 20 MHz bandwidth channel with 2×2 MIMO technology and 256 QAM requires around 2.5 Gbps fronthaul load, in contrast to a 100 MHz radio channel with 8×8 MIMO scheme and 256 QAM, which can cater for almost 40 Gbps.

In the first simulation scenario where both services share the same resources, Fig. 14.17 shows the generated fronthaul load results for all the three split options. Although the number of RUs can vary in this case study, the ONU relates to three RUs.

It can be observed how a fully centralized architecture (option 8), with all the functionalities on the CO, requires high data rates on the fronthaul. However, wavelength restrictions of the PON do not allow the implementation of option 8 for almost all the simulation time. In the scenario, there is one wavelength from OLT to ONU with 10 Gbps bandwidth and when all three antennas are active (serving users), option 8 exceeds the PON bandwidth threshold of 10 Gbps. Option 7 in contrast, which is also a user-independent option with pool gains, can be implemented throughout the simulation period. Compared to flat bitrates (option 7 and 8), option 6 which is user-dependent produces the lowest fronthaul load and is analogous to the user's requirements in addition to the extra overhead of the MAC layer. In this context, it is worth mentioning that the calculation for the fronthaul load in split option 6 is based on (14.5), which is the sum of the user's request in terms of bandwidth plus the additional overhead for this option.

$$FH\ load\ option\ 6\ (in\ Mbps) = \sum BW^{users}_{req} + OH^{\%}_{option6} \bullet \sum BW^{users}_{req} \tag{14.5}$$

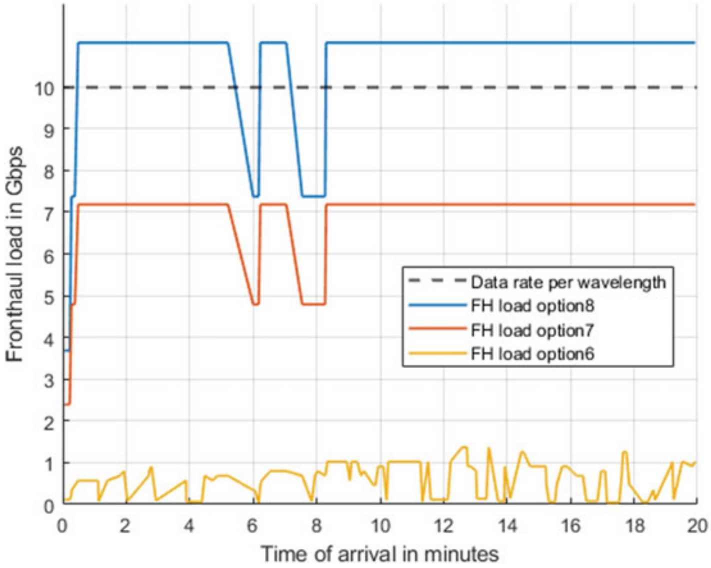


Fig. 14.17 Fronthaul load calculation for split options 6, 7, and 8

In the second scenario, a network slicing mechanism is enabled, and the resources now are divided logically for each service. Furthermore, the system has to implement now two functional splitting mechanisms, one for each service-based slice. Each user traffic type performs the admission control with its own radio channel on the access network. Figure 14.18 illustrates such a scenario calculating the fronthaul load of each user type separately.

The peaks and valleys for the mMTC users appear because the arrival rate is lower than for eMBB (1 user/min for mMTC compared to 6 users/min for eMBB) and the RUs are not active all the time, as well as option 6 fronthaul load for mMTC is insignificant:10 Mbps versus 100 Mbps bandwidth requirement for eMBB. Simulation results show that none of the mMTC users have been rejected in comparison to the eMBB users. To avoid such a situation, especially if the number of served users increase, dynamic real-time network slicing in the fronthaul can be applied as in [69], to satisfy service level agreements.

By slicing the network for each service, and slicing the resources on the radio part, a similar mechanism needs to be done on the fronthaul link as well. The one functional split among the CO and the RUs must be sliced into the number of slices (or services) and a functional split per slice needs to be selected. This approach increases the computational resources but entails multiplexing gains on the fronthaul. Figure 14.19 presents a high-level RAN architecture with the controller layer providing functional split programmability as a service.

The RAN architecture enhances the baseline architecture by functional models emerging from the 5GPPP innovations as outlined in [70]. One such extension is the

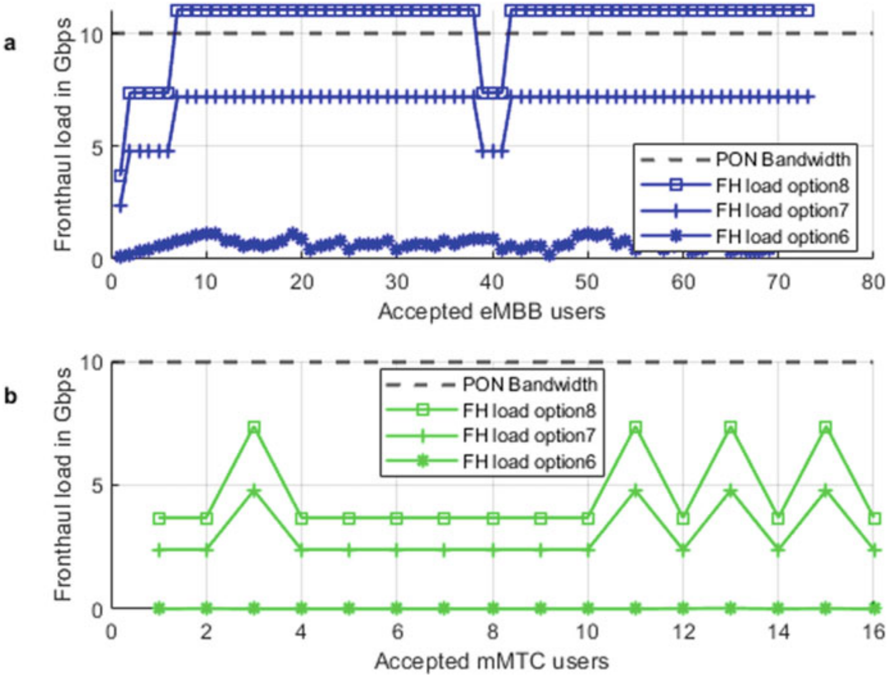


Fig. 14.18 (**a**) Fronthaul load for eMBB users, (**b**) Fronthaul load for mMTC users

controller layer, which enables fronthaul programmability, in terms of functional split control functions, as specific application (FSaaS App) implementations. The application can run on the NBI over the SDN controller, and the communication with the RAN can be maintained over the SBI. It can be envisioned that such a service provides slow-scale control functionality and can support the RAN control functions, such as the radio resource management.

Phase 3 is the last phase of the simulation, in which the SDN controller must decide which split option is optimal for each slice. Since the eMBB slice requires more resources and because of the mobility of the users, they undoubtedly will benefit from a more centralized option where management of resources is performing on the CO. Therefore, the controller gives priority to the eMBB slice, and according to (14.6) is trying to find the most centralized split option with respect to the total PON bandwidth.

$$N_{act\,RU}^{eMBB} \bullet BW_{FSi} < BW_{\lambda wv} \bullet N_{ONU} \tag{14.6}$$

wherein

$N_{act\,RU}^{eMBB}$ is the number of active RUs in the eMBB slice
BW_{FSi} is the bandwidth requirement of a functional split i in Mbps

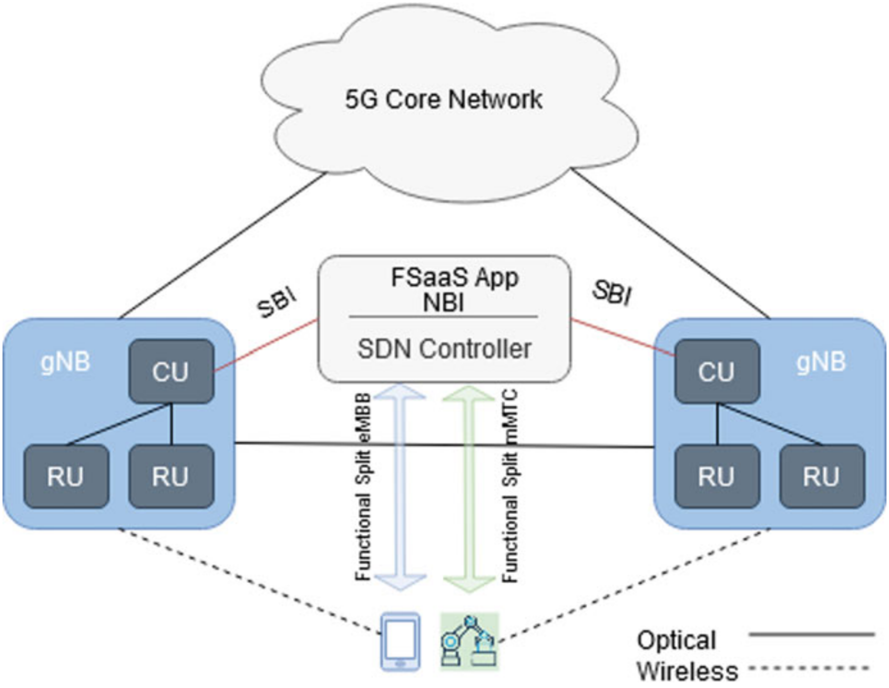


Fig. 14.19 High-level RAN architecture with the controller layer providing functional split programmability

BW_{λ_wv} is the bandwidth of the wavelength in DL in Mbps
N_{ONU} is the number of ONUs in the system

After the controller determines the split option for eMBB users, the available bandwidth of eMBB slice is calculated based on the optimal split selection (14.7)

$$BW_{av}^{eMBB} = N_{act\,RU}^{eMBB} \bullet BW_{FS}^{chosen} \tag{14.7}$$

wherein

BW_{av}^{eMBB} is the total available bandwidth for eMBB slice after the optimal functional split decision in Mbps
BW_{FS}^{chosen} is the bandwidth requirement for the chosen optimal functional split in Mbps

The rest of the bandwidth according to (14.8) remains for the mMTC slice and following (14.9), the controller determines again the optimal split for the mMTC slice, respectively.

$$BW_{rem}^{mMTC} = BW_{\lambda wv} \bullet N_{ONU} - BW_{av}^{eMBB} \tag{14.8}$$

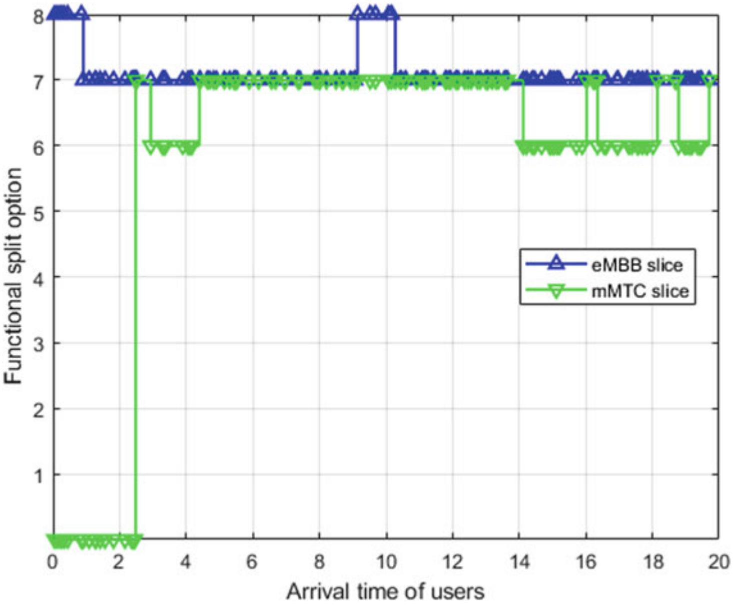


Fig. 14.20 Functional split options for eMBB and mMTC slices

$$N_{act\,RU}^{mMTC} \bullet BW_{FSi} < BW_{rem}^{mMTC} \tag{14.9}$$

wherein

BW_{rem}^{mMTC} is the available bandwidth for the mMTC slice in Mbps
$N_{act\,RU}^{mMTC}$ is the number of active RUs in mMTC slice

Thus, during the simulation time, a functional split option is selected and given as a service to each slice with respect to bandwidth requirements based on the available options 8, 7, and 6. Figure 14.20 demonstrates such an example with the eMBB slice to obtain the option 8 at the beginning where less RUs are active, and later throughout the simulation period option 7 is activated. At the beginning, the slice for the mMTC service does not implement any functional split as the RU does not serve any user. When mMTC users arrive, the controller based on the remaining bandwidth selects the option 7, and subsequently option 6, when all the RUs are active. When some RUs do not serve users, the mMTC slice can obtain a higher split value, like option 7, for instance.

14.5 Conclusion

This chapter discussed the introduction of new technologies in mobile communication networks. Network softwarization enables SDN-based networks resulting in a novel way of managing the communication and IT resources by separating the control logic from hardware and its centralization in software-based controllers. SDN controllers are the new pieces in the puzzle of B5G, complementing the resource utilization and optimization with more efficient techniques on the control plane and orchestrating both the back and front part of the network. Currently, both backhaul and fronthaul networks utilize PONs to solve the densification problem and to provide an ultra-high-speed gateway to the core network. However, the evolution of RAN bridges the optical infrastructure with the wireless access domain of the end users, providing 5G services like eMBB, mMTC, and URLLC through small cell technology and enabling centralized management of resources promoting higher capacity networks. This chapter investigated applications of softwarization based on fronthaul functional split service through programmable networks, and network slicing either as a service (network slicing use-case) or as a technique (flexible functional split decision use-case). By employing these technologies in synergy through the SDN paradigm, significant gains could be achieved in terms of efficient QoS provisioning. In conclusion, on one hand, by merging SDN and network slicing, network slices can be fully distinguished and the SDN controller assigns the efficient amount of resources to each slice while the UL traffic loads of each slice changes. Moreover, the impact of the channel bandwidth on the resource allocation procedure was investigated and it was shown that the resources allocation will be more efficient and dynamic as the bandwidth unit gets smaller inside both the optical and wireless domains. On the other hand, the functional split paradigm with the SDN integration showed the benefits and feasibility of applying dynamically various functional splits per service-based slices (such as mobile broadband and IoT). Thus, it is clear that network provisioning needs to be compatible with all the service requirements in all slices, in such a way that different functional splits will be used for different slices in order to enhance system's performance and resource utilization.

Acknowledgments This research work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 722429 (5GSTEPFWD).

References

1. Shah, S. A. R., Kim, J., Bae, S., et al. (2016). Network softwarization: A study of SDN and NFV integration. *International Conference on Convergence Technology, 6*(1), 834–835.
2. Bannour, F., Souihi, S., & Mellouk, A. (2018). Distributed SDN control: Survey, taxonomy, and challenges. *IEEE Communications Surveys & Tutorials, 20*(1), 333–354.

3. Li, L., Mao, Z., & Rexford, J. (2012). Toward Software-defined Cellular networks. In *Software Defined Networking (EWSDN), 2012 European Workshop on* (pp. 7–12).
4. Thales Group, "Introducing 5G Technology and Networks (Definition, Use Cases and Roll-out)". https://www.thalesgroup.com/en/markets/digital-identity-and-security/mobile/inspired/5G. Accessed 25 July 2020.
5. Al-samman, A. M., Azmi, M. H., & Rahman, T. A. (2018). A Survey of Millimeter Wave (mm-Wave) Communications for 5G: Channel Measurement Below and Above 6 GHz. In *Proceedings of international conference of reliable information and communication technology* (pp. 451–463).
6. Lara, A., Kolasani, A., & Ramamurthy, B. (2014). Network innovation using OpenFlow: A survey. *IEEE Communication Surveys & Tutorials, 16*(1), 493–512.
7. Thyagaturu, A. S., Mercian, A., McGarry, M. P., et al. (2016). Software defined optical networks (SDONs): A comprehensive survey. *IEEE Communication Surveys & Tutorials, 18*(4), 2738–2786.
8. Kosmetos, E., Matrakidis, C., Stevdas, A., et al. (2018). An SDN architecture for PON networks enabling unified management using abstractions. In *Proceedings of ECOC* (pp. 1–3).
9. Nguyen, V.-G., Brunstrom, A., Grinnemo, K.-J., et al. (2017). SDN/NFV-based Mobile packet core network architectures: A survey. *IEEE Communication Surveys & Tutorials, 19*(3), 1567–1602.
10. Alqahtani, A. M., Mohamed, S. H., El-Gorashi, T. E. H., et al. (2020, May 2). PON-based connectivity for fog computing, arXiv:2005.00839.
11. Konstantinou, D., Bressner, T. A. H., Rommel, S., et al. (2020). 5G RAN architecture based on analog radio-over-fiber fronthaul over UDWDM-PON and phased Array fed reflector antennas. *Optics Communications, 454*, 124464.
12. Azofra, J., Merayo, N., Aguado, J. C., et al. (2018). Implementation of a testbed to analysis a SDN based GPON. In *Proceedings of European Conference on Optical Communication (ECOC)* (pp. 1–3).
13. Pakpahan, A. F., Hwang, I. S., & Liem, A. T. (2019). Enabling agile software-defined and NFV based energy-efficient operations in TWDM-PON. In *7th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1–7).
14. Khalili, H., Rincon, D., Sallen, S., et al. (2018). An integrated SDN-based architecture for passive optical networks. In A. Haidine & A. Aqqal (Eds.), *Broadband communications networks – Recent advances and lessons from practice*. Intech.
15. Clegg, R. G., Spencer, J., Landa, R., et al. (2014). Pushing software defined networking to the access. In *Third European workshop on software defined networks* (pp. 31–36).
16. Kanta, K., Pagano, A., Ruggeri, E., et al. (2020). Analog Fiber-wireless downlink transmission of IFoF/mmWave over in-field deployed legacy PON infrastructure for 5G Fronthauling. *IEEE/OSA Journal of Optical Communications and Networking, 12*, D57–D65.
17. Kalfas, G., Vagionas, C., Antonopoulos, A., et al. (2019). Next generation fiber-wireless fronthaul for 5G Mmwave networks. *IEEE Communications Magazine, 57*(3), 138–144.
18. Datsika, E., Kartsakli, E., Vardakas, J. S., et al. (2018). QoS-aware resource management for converged fiber wireless 5G fronthaul networks. In *2018 IEEE Global Communications Conference (GLOBECOM)* (pp. 1–5).
19. Raddo, T. R., Rommel, S., Cimoli, B., et al. (2019). The optical fiber and mmWave wireless convergence for 5G fronthaul networks. In *2019 IEEE 2nd 5G World Forum (5GWF)* (pp. 607–612).
20. Habib, U., Aighobahi, A. E., Quinlan, T., et al. (2017). Demonstration of radio-over-fiber-supported 60 Ghz MIMO using separate antenna-pair processing. In *2017 International Topical Meeting on Microwave Photonics (MWP)* (pp. 1–4).
21. Kalfas, G., Agus, M., Pagano, A. et al. (2019). Converged analog fiber-wireless point-to-multipoint architecture for eCPRI 5G fronthaul networks. In *IEEE Global Communications Conference (GLOBECOM)* (pp. 1–6).

22. Vardakas, J. S., Monroy, I. T., Wosinska, L. et al. (2017). Towards high capacity and low latency backhauling in 5G: The 5G STEP-FWD vision. In *2017 19th International Conference on Transparent Optical Networks (ICTON)* (pp. 1–4).
23. 3GPP TS 38.108: “NR; Base Station (BS) Radio Transmission and Reception”.
24. Elbers, J. P., Grobe, K., Magee, A. (2014). Software-defined access networks. In *The 2014 European Conference on Optical Communication (ECOC)* (pp. 1–3).
25. BLUESPACE 5G PPP Phase 2 project. https://bluespace-5gppp.squarespace.com. Accessed 25 July 2020.
26. Muñoz, R., Vilalta, R., Fàbrega, J. M., et al. (2018, June 18). BlueSPACE’s SDN/NFV architecture for 5G SDM/WDM-enabled fronthaul with edge computing. In *2018 European Conference on Networks and Communications (EuCNC)* (pp. 403–9).
27. 5G-XHAUL 5G PPP Phase 1 project. https://www.5g-xhaul-project.eu. Accessed 25 July 2020.
28. Larsen, L. M., Checko, A., & Christiansen, H. L. (2018). A survey of the functional splits proposed for 5G mobile crosshaul networks. *IEEE Communications Surveys & Tutorials, 21*, 146–172.
29. Maeder, A., Lalam, M., De Domenico, A., et al. (2014). Towards a flexible functional split for Cloud-RAN networks. In *2014 European Conference on Networks and Communications (EuCNC)* (pp.1–5).
30. 3GPP TR 38.801 V14.0.0: “Study on New Radio Access Technology: Radio Access Architecture and Interfaces”, 2017.
31. Ericsson, A. B. (2015). Common Public Radio Interface (CPRI); Interface Specification v7.0. *Huawei Technologies Co. Ltd, NEC Corporation, Alcatel Lucent, and Nokia Networks*.
32. Small Cell Forum. (2015). *Solving the HetNet Pzzle small cell virtualization functional splits and use cases*.
33. NGMN Alliance. (2015). *Further study on critical C-RAN technologies*.
34. Mitola, J. (1995). The software radio architecture. *IEEE Communication Magazine, 33*(5), 26–38.
35. Nikaein, N. (2015). Processing radio access network functions in the cloud: Critical issues and modeling. In *Proceedings of the 6th international workshop on mobile cloud computing and services* (pp. 36–43).
36. Nguye, V. G., Brunstrom, A., Grinnemo, K. J., et al. (2017). SDN/NFV-based Mobile packet core network architectures: A survey. *IEEE Communications Surveys & Tutorials, 19*(3), 1567–1602.
37. Harutyunyan, D., & Riggio, R. (2018). Flex5G: Flexible functional Split in 5G networks. *IEEE Transactions on Network and Service Management, 15*(3), 961–975.
38. Redana, S., Bulakci, Ö., Zafeiropoulos, A., et al. (2019). *5GPPP architecture working group: View on 5G architecture*. European Commission.
39. Chih-Lin, I., Li, H., Korhonen, J., et al. (2017). RAN revolution with NGFI (xHaul) for 5G. *Journal of Lightwave Technology, 36*(2), 541–550.
40. Alimi, I. A., Teixeira, A. L., & Monteiro, P. P. (2017). Toward an efficient C-RAN optical fronthaul for the future networks: A tutorial on technologies, requirements, challenges, and solutions. *IEEE Communications Surveys & Tutorials, 20*(1), 708–769.
41. Monti, P., Li, Y., Mårtensson, J., et al. (2017). A flexible 5G RAN architecture with dynamic baseband split distribution and configurable optical transport. In *2017 19th International Conference on Transparent Optical Networks (ICTON)* (pp. 1–1).
42. Arslan, M. Y., Sundaresan, K., & Rangarajan, S. (2015). Software-defined networking in cellular radio access networks: Potential and challenges. *IEEE Communications Magazine, 53*(1), 150–156.
43. Marotta, A., Cassioli, D., Kondepu, K., et al. (2019). Exploiting flexible functional split in converged software defined access networks. *Journal of Optical Communications and Networking, 11*(11), 536–546.
44. Matoussi, S., Fajjari, I., Aitsaadi, N., et al. (2019). Joint functional split and resource allocation in 5G Cloud-RAN. In *2019 IEEE International Conference on Communications (ICC)* (pp. 1–7).

45. SDR LAB, "Functional Split Platform". https://sdr-lab.u-pem.fr/splitting-C-RAN.html. Accessed 25 July 2020.
46. OpenAirInterface. http://www.openairinterface.org. Accessed 25 July 2020.
47. Alfadhli, Y., Xu, M., Liu, S., et al. (2018). Real-time demonstration of adaptive functional Split in 5G flexible mobile Fronthaul networks. In *Optical fiber communication conference* (pp. 1–3).
48. Foukas, X., Nikaein, N., Kassem, M. M., et al. (2017). FlexRAN: A software-defined RAN platform. In *Proceedings of the 23rd annual international conference on mobile computing and networking* (pp. 465–467).
49. Chang, C. Y., Nikaein, N., Knopp, R., et al. (2017). FlexCRAN: A flexible functional split framework over ethernet fronthaul in Cloud-RAN. In *2017 IEEE International Conference on Communications (ICC)* (pp. 1–7).
50. Lashgari, M., Natalino, C., Contreras, L. M., et al. (2019). Cost benefits of centralizing service processing in 5G network infrastructures. In *Asia Communications and Photonics Conference (ACP)* (pp. 1–3).
51. Series, M. (2015). IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond, *Tech. Rep. Recommendation ITU-R M.2083-0*.
52. Guan,W.,Wen, X.,Wang, L., et al. (2018). A service-oriented deployment policy of end-to-end network slicing based on complex network theory. *IEEE Access*, 6, 19691–19701.
53. 5GSTEPFWD ITN project http://www.5gstepfwd.eu. Accessed 8 Jun 2021.
54. 5GNORMA 5GPPP Phase 1 project. http://www.it.uc3m.es/wnl/5gnorma. Accessed 25 Jul 2020.
55. 5G-TRANSFORMER 5GPPP Phase 2 project http://5g-transformer.eu. Accessed 25 Jul 2020.
56. ONF TR-521. (2016). SDN architecture, Tech. Rep.
57. Ordonez-Lucena, J., Ameigeiras, P., Lopez, D., et al. (2017). Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges. *IEEE Communications Magazine, 55*(5), 80–87.
58. ONF TR-526. (2016). Applying SDN architecture to 5G slicing, Tech. Rep.
59. Osseiran, A., Monserrat, J. F., & Marsch, P. (2016). *5G Mobile and wireless communications technology*. Cambridge University Press.
60. Tsukamoto, Y., Saha, R. K., Nanba, S., et al. (2019). Experimental evaluation of RAN slicing architecture with flexibly located functional components of Base Station according to diverse 5G services. *IEEE Access, 7*, 76470–76479.
61. ONOS. https://onosproject.org. Accessed 25 Jul 2020.
62. Mosahebfard, M., Vardakas, J., Ramantas, K., et al. (2019). SDN/NFV-based Network Resource Management for Converged Optical-wireless Network Architectures, In *21st International Conference on Transparent Optical Networks (ICTON)*, pp. 1–4.
63. Brown, T. C., & Pollett, P. K. (1991). Poisson approximations for telecommunications networks. *The ANZIAM Journal, 32*(3), 348–364.
64. 3GPP TS 38.306 Version 15.3.0, 5G; NR; User Equipment (UE) Radio Access Capabilities, 2019.
65. 3GPP. NR; Physical channels and modulation. TS 38.211 Version 15.2.0 Release 15.
66. 3GPP TS 38.101-1, NR; User Equipment (UE) Radio Transmission and Reception, 2017.
67. 3GPP TSG RAN WG3. Transport Requirement for CU and DU Functional Splits Options, 2016.
68. NGMN, Further Study on Critical C-RAN Technologies, *White Pap. Version 1.0*, 2015.
69. Maule, M., Mekikis, P., Ramantas, K., et al. (2019). Real-time dynamic network slicing for the 5G radio access network, In *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6.
70. 5GPPP AWG, View on 5G Architecture, *Version 3.0.* 2019.

Chapter 15
Cloud-Based Content Management for B5G

Santiago Sánchez, Tadege Mihretu Ayenew, and Mahshid Mehrabi

Abstract In this chapter, we present cloud-based content management for beyond 5G (B5G) cellular network, where we consider the general scope as how multimedia contents and related computational tasks can be managed effectively over the emerging mobile network that is predominantly becoming a virtual machine. In this context, we explore content management approaches in terms of how content can be offloaded and cached and computational offloading. Our aim lies towards reducing the service latency and backhaul congestion, which directly enhances the network performance, and takes a step towards reducing energy consumption in the network system and devices. We harness the cloud-based infrastructure to push further the boundaries of current solutions to enable more dynamic approaches according to real-time scenarios and requirements using the inherent virtual architecture that 5G and beyond offers.

15.1 Introduction

For the past decade, multimedia applications have been extremely advancing, causing an explosive increase in broadband content traffic. This advance has placed enormous demand on bandwidth and computationally limited backhaul networks. In addition, these contents demand for preprocessing before they are passed through the Internet. Hence, the ability to precache this content offers multiple benefits such as network offloading, service latency and cost reduction that

S. Sánchez
Universitat Oberta of Catalonia (UOC), Barcelona, Spain
e-mail: ssanchezcor@uoc.edu

T. M. Ayenew (✉)
National and Kapodistrian University of Athens (UOA), Athens, Greece
e-mail: tadegem@di.uoa.gr

M. Mehrabi
Technische Universität Dresden (TUD), Dresden, Germany
e-mail: mahshid.mehrabi@tu-dresden.de

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3_15

result in enhanced cellular network performance. However, it is also clear that to find the optimal content caching strategy is a challenge that is often modelled as an optimization problem to increase the QoE-related parameters (such as service latency, throughput, cache hit probability, energy, etc.) under resource (such as cache size, computation, bandwidth, etc.)-limited networks.

Nowadays, 5G and beyond architecture is provisioning intelligent vertical solutions that enhance content caching based on the estimation of content attributes such as popularity, location, mobility and ephemerality by resorting to artificial intelligence techniques such as deep learning (DL). Specifically, the DL is a novel vertical service helper that manages over-complicated decision processes, enhancing wireless network management and resource optimizations. This service is applied to the network edges, providing low-cost policy control based on intelligent network status predictions. To understand the optimal management of DL techniques as a service, there is a need to understand the learning time consumption, available resources in the specific period and the transmission cost. In this context, in Sect. 15.2, we explored learning-based content management and proposed a content caching technique using DL. As a case study, we investigated the impact of mobility and popularity estimations on performance and to obtain the optimal content update technique, thus taking a step towards alleviating the traffic load on the backhaul network. Furthermore, in Sect. 15.3, we revisit the traffic offloading scenario where we model the problem as generalized knapsack problem, which is simplified using relaxation technique and solved using dynamic programming (DP). This scheme increases the availability of content at the helper node (computing node) in a balanced fashion. Such straightforward and very efficient optimization methods boost the performance of both the content caching process and the enabling technologies.

To deploy the enabling technologies for content and traffic management, there is a need for disruptive and efficient network architectures that provide service platform on which novel services can be orchestrated. Among many enhancing technologies, multi-access edge computing (MEC) is proposed to improve the performance of content management in cellular networks. This new technology helps to manage newly emerging application requirements by offloading the computational capability to devices that are now also considered as part of the edge network. The task offloading, harnessing on MEC capability, aims to reduce power consumption by eventually enabling truly virtualized handsets as well as to enhance network resource utilization. Although, this topic has been well investigated, there are still some open issues to be foreseen for practical uses and within mobile small cell context that is investigated in Sect. 15.4, to what we refer to as "Device-Assisted Computation Offloading for B5G Network".

15.2 AI-Caching for Traffic Offloading

15.2.1 Introduction

Nowadays, the connectivity of intelligent devices in cellular network is massively increasing, which is causing a huge impact on the transport network (backhaul/fronthaul). These mobile devices produce a huge amount of traffic over the transport network, which means a huge load for mobile network operators (MNO). According to Cisco, by the end of 2022, the average number of connected devices will be 3.6 per capita [1]. Additionally, by the end of 2022, the traffic will grow to nearly 396 exabytes per month. This extensive increase in traffic is mainly attributed to the wide usage of virtual reality and video streaming platforms, which are expensive to MNOs. Concretely, to deal with this traffic, MNOs are adapting their structures with computing and storage resources to interact with content providers (CP) to satisfy these aspects.

However, the fifth-generation wireless communication (5G) technology promises to answer ambitious requirements of end users, which caused high complexity to alleviate this traffic burden. To facilitate the management of this traffic load, new practical technologies have been deployed on the network edges, over which intelligent entities optimize the network in a decentralized manner to intelligent nodes, called mobile helpers (MH) [2]. Concretely, at the edges manage storage and computing resources available to mobile users through wireless connections. These nodes enable several mobile services, such as machine learning services, augmented reality and cognitive assistance [3]. Furthermore, these computational techniques allow MNOs to provide specific services for traffic offloading in different scenarios.

Among other complementary services, content caching aims at compromising the traffic demand over the network edges. It assists the video platforms and social media streaming to control heavy sized files that gain or lose popularity fast among users [4, 5]. The caching process is usually based on recovering the old IP caching concept and migrating this to the network edges, storing content intelligently to reduce backhaul costs. This can be done through storage units of the MHs, located over the base stations (BS), improving the content download. Basically, it satisfies the high traffic demand but in a limited cache size of the MHs, called storage unit (SU). However, maximizing the storage utilization is a challenge to address [6]. The optimal utilization of the caching storage to minimize the backhaul traffic involves the analysis of the traffic dynamics to optimally update contents at the storage. Specifically, two of the main factors that influence traffic offloading are user mobility and user preferences.

The impact of both factors can be addressed using machine learning (ML) and DL techniques to estimate the user mobility trajectory [7, 8] and content preferences [9, 10]. However, DL and ML have several challenges to address on these estimations to reach optimal performance on the content updates. The main challenge of these estimations is the periodicity from which the method predicts the user position and the popularity of the content, and then the periodicity from which the optimization method will apply the optimal content update (OCU). Essentially, it is necessary to know how much information is being periodically captured and how much is being lost, due to remaining errors through predictions [11–13]. This existing error tends to increase when the data is non-well sampled in both dependencies (mobility and content preferences) [14–16]. Then, this error will be reflected in the traffic offloading due to inaccurate content updates on the SU. To analyse this, it is important to learn how to model the network and which techniques lead to optimal predictions for caching. However, that depends on the dynamics of the user mobility and the content preferences, and the prediction will not avoid errors, but this error must be quantified on specific periodicity to reach a maximum accuracy on traffic offloading. Usually, this error varies depending on the time horizon of the network to apply the content update.

All the arguments mentioned so far synthesize the existing situation on traffic offloading, which are the challenges on backhaul. We present how to model caching and several learning techniques applied to content caching in Sect. 15.2.2. The content placement and the network analysis are presented in Sect.15.2.3. How to reach optimal traffic offloading and perform caching analysis are presented in Sects. 15.2.4, 15.2.5 of this chapter. In Sect. 15.2.6, we conclude by giving insights of DL-based caching to minimize the backhaul traffic.

15.2.2 Traffic Offloading Techniques

In practice, there exist two key aspects to obtain an optimal content update. The first consists in distinguishing between local and global content popularities. In particular, the mobile users aim user preferences, and these preferences impact all the content popularity of network locally and globally. Measuring the local popularity at each BS, the MNO can observe particular file's popularity on the BS [17]. However, globally measuring the overall BS's popularity, the MNO will obtain inaccurate popularity of all users [18]. These measurements are crucial to obtain an accurate content update and thus to reach optimal traffic offloading.

The second key aspect lies in the optimal traffic analysis, which means how fast the traffic changes and how fast the learning method estimates an OCU [19]. This analysis requires adequate sampling techniques to estimate user mobility and user request over time. This issue is of major importance in crowded (hotspot) scenarios, where the resources are limited, and the decisions must be taken on time. Usually, traffic rapidly varies in crowded places (metro stations, stadiums, main streets) where mobility and request are highly dynamic depending on specific day times. In literature, several works have sought to address these ideas. Specifically, in [20] the authors use the AdaBoost model to track the users' positions along time by defining a Markov decision process (MDP). By this approach, they perform accurate traffic predictions per BS in long horizons without including short horizon dynamics.

The reasons mentioned so far particularly emphasize on long term horizons where the network popularity does not frequently vary. It is extremely important to predict the content popularity in the short term where the traffic is dynamic and crucial for MNOs. To mitigate its effects, it is necessary to estimate the probability of having an accurate short horizon prediction. MNOs must interpret and introduce intelligent AI entities at the BS to step up content updates, to analyse the costs of learning and the benefits of the update with this estimated information.

15.2.3 Network Analysis for Content Placement

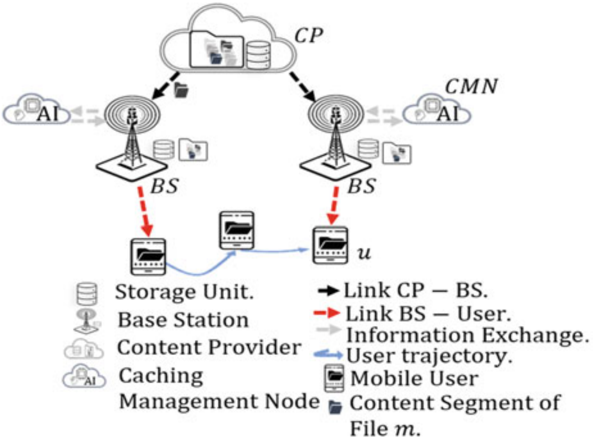
Based on the abovementioned estimation insights to analyse an OCU, we divide the content placement analysis in two key factors: the network modelling and the traffic analysis. The former is to understand the limitations that content update is facing on real network scenarios, and the latter one is to show the impact of the traffic analysis to perform content updates.

15.2.3.1 Network Modelling

In the field of wireless communications, network analysis plays a decisive role to support different traffic scenarios with specific conditions. As we mentioned earlier, the traffic offloading relies on the mobility and user preference estimations. To make an OCU over the network SUs, given the network conditions, we define a network composed of BSs equipped with SUs and supported by a caching management node (CMN). Concretely, the CMN will be operating on the BSs optimizing the content update on the SU, to offload the network traffic. Furthermore, these SU have a storage capacity denoted by c_b. Consequently, the accuracy of the content update will depend on the user mobility and content popularity estimations that are operating over the CMNs. Apart from that, the BSs are connected to the core network through a backhaul link, with a limited capacity ψ_b, in (bps). The library of contents, namely, M, is stored in the CP and downloaded to any BSs upon a request. Thanks to the storage capacity of BSs, contents can be partially or completely cached into the SU of one or several BSs. This alleviates the backhaul link requirements. Furthermore, each content has a specific size s_m/Δ segments. The size of the segments in which contents are divided, i.e. Δ, is the same for all contents, and it is the minimum content size that can be stored in a SU. The set of users served by the described 5G network is denoted by U. Any given user $u \in U$ is characterized by the position at time t, namely, p_u^t, the content request rate λ_u,in requests/s, and by the vector of content preferences $\theta_u = (\theta_{u,1}, \theta_{u,2}, \ldots, \theta_{u,|M|})$, where $\theta_{u,m}$ is the probability of user u requesting content m when there is a request. Hereafter, θ_u and λ_u are assumed to remain constant over time.

Focusing on the request and delivery process, when a user u requests for a content to a serving BS, it is directly served by the BS if it is already stored in its SU. Conversely, if the content is not stored in the SU, the content is downloaded from the CP through the backhaul link and finally delivered to the user by the BS. Likewise,

Fig. 15.1 System model



the content can be partially stored in the SU. That is, some segments with size Δ are stored in the SU and some others are not. In that case, only missing segments are downloaded from the CP, while segments stored in the SU of the BS are directly delivered to the user. Regarding the user mobility, the content can be delivered by segments when the segments are available over several BSs over time.

Besides the abovementioned advantages of ML, we illustrate our system model in Fig. 15.1, including several ideas to overcome limitations on the content update. To further clarify, Fig. 15.1 illustrates our 5G network scenario providing the content placement from the CP to the SU of the BS. Specifically, the content is successfully delivered from two BSs: the first segment is delivered making usage of the backhaul link due to the non-updated segment over the SU, and the second which delivers the segment available on the SU during the user trajectory.

15.2.3.2 Backhaul Traffic Estimation

The estimation of backhaul traffic depends on the requests of users, as well as on the contents stored in the BSs' SUs. That is, when the requested content is stored in the SU, then the content need not be downloaded through the backhaul link. Let us define the request rate at BS b at time t, denoted by λ_b^tas the aggregate of the request rate of all users served by b:

$$\widetilde{\lambda_b^t} = \sum_{u=1}^{U} \lambda_u \int_{A_b} f_P^u(p,t)\,dp, \tag{15.1}$$

where A_b is the coverage area of BS b and $\int_{A_b} f_P^u(p,t)$ is the surface integral of the user location probability distribution function (*pdf*) over A_b. It is worth noting that the location of users is characterized by the *pdf* of their locations. Hence, a user can impact on the estimate of the request rate of several BSs. Therefore, $\int_{A_b} f_P^u\,(p,t)\,dp \le 1$ and $\sum_{b=1}^{B} \int_{A_b} f_P^u\,(p,t) = 1$. In the same manner, the estimate of the popularity of contents at BS b is $\overset{\sim}{\Theta^t_b}$ $=(\overset{\sim}{\Theta^t_{b,1}}, \ldots, \overset{\sim}{\Theta^t_{b,|M|}})$, where:

$$\widetilde{\theta^t_{b,m}} = \frac{\sum_{u=1}^{u} \lambda_u \theta_{u,m} \int_{A_b} f_P^u\,(p,t)}{\widetilde{\lambda^t_b}}. \tag{15.2}$$

As we discussed before, these two estimations are crucial to define an optimal backhaul traffic minimization. To obtain the OCU, the CMN must process all the estimations submitted to generate an optimal segment allocation over the BS. Based on this, we define a content allocation vector of BS b at time t, denoted by $\overset{\sim}{v_b^{t^*}} = (v_{b,1}^{t^*}, \ldots, v_{b,|M|}^{t^*})$, where $\overset{\sim}{v_b^{t^*}}$ is the number of segments of content m stored in BS b from the estimations of $\widetilde{\lambda^t_b}$and $\overset{\sim}{\Theta^t_b}$. The values of this vector are $\overset{\sim}{v_b^{t^*}} = 0, \ldots, s_m/\Delta$. Furthermore, note that the amount of content is limited by the storage capacity of the SU's $\sum_{m \in M} \overset{\sim}{v_b^{t^*}} \le c_b/\Delta$. Which means that for a given $\int_{A_b} f_P^u\,(p,t)$, it is possible to calculate the backhaul traffic minimization as:

$$\min_{\overset{\sim}{v_b^{t^*}}} \overset{\sim}{\zeta^t_b} = \widetilde{\lambda^t_b}\, \overset{\sim}{\Theta^t_b} \left(s - \Delta \overset{\sim}{v_b^{t^*}} \right)^T \tag{15.3a}$$

$$s.t. \qquad \left\| \overset{\sim}{v_b^{t^*}} \right| |_1 \le \frac{c_b}{\Delta}, \tag{15.3b}$$

$$\overset{\sim}{v_b^{t^*}} = 0, 1, \ldots, \frac{s_m}{\Delta}, \qquad \forall m \in M \tag{15.3c}$$

where $s = (s_1, s_2, \ldots, s_{|M|})$ is the vector of contents' sizes and $(\cdot)^T$ is the vector-transpose operator. The $\| \cdot ||_1$ is the 1-norm operator. Note that $\overset{\sim}{v_b^{t^*}}$is not the optimal content allocation vector for the backhaul traffic, however, is the optimal content allocation vector for the estimate of the backhaul traffic. The minimization of the estimated backhaul traffic is constrained by the capacity of the SU, as shown by (15.3b). As defined above, contents are divided into equal Δ size segments, which can be stored in the SU or not. In any case, the total number of segments into which a content m is divided $s_m/\mathit{\Delta}$ as stated by (15.3c). The explanation of how we reach the $\int_{A_b} f_P^u\,(p,t)$ definition is addressed in the subsequent section.

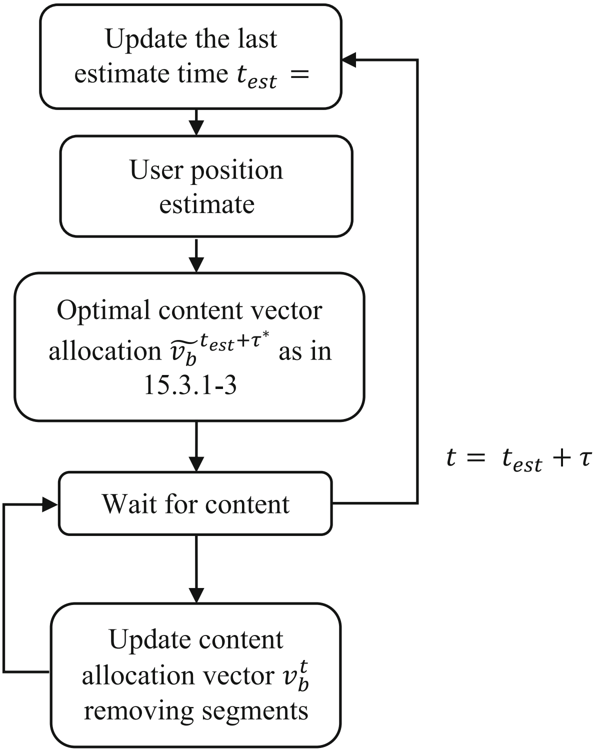

Update the last estimate time t_{est} =
User position estimate
Optimal content vector allocation $\widetilde{v_b}^{t_{est}+\tau^*}$ as in 15.3.1-3
Wait for content
Update content allocation vector v_b^t removing segments
$t = t_{est} + \tau$

Fig. 15.2 Flowchart of the proposed content update

15.2.4 *Optimal Traffic Offloading: A Learning Approach*

15.2.4.1 Content Placement Management

Based on the abovementioned section, the content estimation must be correlated with the *pdf* estimation of the position of the users at time $t_{est} + \tau$, namely, $\int_{A_b} f_P^u (p, t_{est} + \tau)\, \forall u \in U$. It indicates that the optimal content allocation vector at $t_{est} + \tau$, $\overset{\sim}{v_b}^{t_{est}+\tau^*}$ can be calculated using (15.3a). The aim of the caching update is to turn the current content allocation vector $\overset{\sim}{v_b}^{t_{est}}$ into the estimated content allocation vector $\overset{\sim}{v_b}^{t_{est}+\tau^*}$. This update must be carried out within the period $[t_{est}, t_{est} + \tau)$, and it is detailed in the flowchart given by Fig. 15.2.

In Fig. 15.2, it is possible to observe the flowchart with several processing steps: initially, the user position estimate and the optimal content vector allocation. Then, the system waits for the m content request, and two decision steps are imposed to analyse the estimation. The former observes the time elapsed since the last prediction $t_{est} + \tau$ and the latter to measure the segment difference among the optimal content vector allocation and the existing allocated vector. Then, a check decision C_b storage unit step is evaluated, to process the content segment removal and then pass to *wait* mode for a new content m request.

15.2.4.2 Deep Learning Analysis Estimation

To optimally learn from the network, the information that influences the OCU is required to understand how fast the network traffic varies. An appropriate technique to estimate the user mobility is the recurrent neural networks, which aims to obtain sequential predictions of the user trajectory. In particular, the long short-term memory (LSTM) networks aim to reach optimal time series predictions. We are interested to assist the LSTM with a mixture density network (MDN) to rehearse the learning phase and obtain a user position estimation, as a *pdf* to support the traffic estimations. To support the performance of estimations as a *pdf*, we use the formulation proposed in [9]:

$$\begin{aligned} f^t &= \varphi\left(W_f \cdot \left[h^{t-1}, x^t\right] + b_f\right) \\ o^t &= \varphi\left(W_{out} \cdot \left[h^{t-1}, x^t\right] + b_{out}\right) \\ i^t &= \varphi\left(W_{in} \cdot \left[h^{t-1}, x^t\right] + b_{in}\right) \\ \tilde{C}^t &= \tanh\left(W_c \cdot \left[h^{t-1}, x^t\right] + b_c\right) \\ C^t &= f^t * \left(C^{t-1}\right) + i^t * \tilde{C}^t \\ h^t &= o^t * \tanh\left(C^t\right) \end{aligned} \tag{15.4}$$

where f^t, o^t, i^t, C^t and h^t are the forget gate, output gate, input gate, memory cell and the hidden representation, respectively. Where $f^t \in \{0, 1\}$, denoting 1 as stored information and 0 as rejected. The learning parameters that need to be estimated during the training phase are W_z and b_z, where $z \in \{f, i, o, c\}$ are the weight matrices and the bias vector and φ is the sigmoid function. Then, x^t is the input P_u^t, where x and y are the u-th user positions over t. Then, the MDN takes the target vector h^t of the LSTM network and completes with a Gaussian mixture model which has the flexibility to model probability distribution functions. The MDN function can be described as follows:

$$y^t = \left\{\left(\overset{\sim}{\mu_c^t}, \overset{\sim}{\alpha_c^t}, \overset{\sim}{\sigma_c^t}, \overset{\sim}{\varrho_c^t}\right)\right\}_{c=1}^{C} = W_y * h^t + b_y, \tag{15.5}$$

where the $\mu_c^t = \overset{\sim}{\mu_c^t}$are the mean values, $\alpha_c^t = \frac{\exp\left(\overset{\sim}{\alpha_c^t}\right)}{\sum_{k=1}^{c} \exp\left(\overset{\sim}{\alpha_c^t}\right)}$ the weights of the probability distribution function (pdf), $\sigma_c^t = \exp\left(\overset{\sim}{\sigma_c^t}\right)$ the variance of the output and $\varrho_c^t = \tanh\left(\overset{\sim}{\varrho_c^t}\right)$ depict the correlation of the c^th Gaussian component in the respective time horizon t. Here, the SoftMax function generates the outputs corresponding to the α_c^t parameters while σ_c^t represents the scale parameters that avoid variances going to zero. Finally, y^t is the Gaussian mixture distribution generated by the mentioned parameters, and the predicted value of the P_u^t at each time horizon is obtained.

The MDN is a full description of the probability distribution; due to this, we use the loss function of the distribution with the training data. In our problem we are interested in using several mixtures to satisfy $\sum_{k=1}^{k} \alpha_k (x; w) = 1$, to estimate the learning adaptability in different mobility scenarios. The validation of this technique is estimated using the negative logarithm of the likelihood, defined as:

$$L_{MDN}(p_u, y_u) = \sum_{t=0}^{T} -\log \left[\sum\nolimits_{c=1}^{C} \alpha_t^c \phi \left[y_u^t; \mu_c^t, \sigma_c^t, \varrho_c^t \right] \right]. \tag{15.6}$$

Finally using this formulation, we reach the *pdf* of the user location as $y_u^t = f_P^u(p, t)\, \forall u \in U$, which is used to estimate the traffic over time; and it is calculated with the usage of the μ_u^t and the σ_u^t of the predictions. Furthermore, this error metric allows us to calculate the impact of the learning accuracy on the content update, depending on the size of the variance obtained from the prediction.

15.2.5 Caching Performance Evaluation on Traffic Offloading

To assess the network performance, we define an experiment, where we tune the learning method to the periodicity of the user mobility. We estimate the optimal learning sequence and update time to observe how significant are the predictions on the content update. We set a total of 100 users located over the area following a spatial uniform distribution. Besides, we set 16 BSs, which are uniformly distributed over the area. Ten m contents are considered and their content size s_m decreasing from [1000,100] Mb, respectively. Three user profiles are defined to enhance the diversity over the network. Two user profiles use decreasing and increasing popularity profiles, respectively, with regard to the size, and the other is a mixture of those profiles – we define 50 users that follow the decreasing profile, 20 that are increasing and 30 that adopt a mixed profile. The network traffic is generated using a request rate $\lambda_u = 0.01$ request/s. This means that each user generates a request every 100 s on average. The distribution of the time between requests follows an exponential distribution. The region of study is a 400 × 400 meter square-shaped area. The base stations are equipped with SU, and all of them have the same size. The SU will increase in value with respect to the backhaul offloading and cache hit ratio (CHR) for the simulation range, until the traffic load is satisfied.

In mobility, we use two scenarios: one with high aggregation of users emulating a hotspot (HS)-crowded area and another with low aggregation (Lagg) to show disperse distribution users over the area. Furthermore, the user's velocity is uniformly defined *as* $r_u \sim U(0, 10)$m/s. User's movement is initialized selecting a uniform angle in $[0, 2\pi]$. Additionally, 4 hotspots are equally distributed for the total users; it means each hotspot will hold 25 users in 2 mobility cases.

We adopt the next learning parameter configuration as set learning rate to 1e-4, the stepwise decay method with a decay rate of 0.95, and the decay period is 1000 iterations of the learning system. We use adaptive moment estimation (Adam) as our batch gradient descendent method. The batch size is 40. The three main layers of the LSTM side are composed of 64 units, and its input is a vector feed with $\xi_{p_u^t}$. We apply a sampling technique that supports the optimal sequence estimation from the sampling depending on the user's velocity and content request. The MDN Gaussian distributions are composed by eight parameters, which are the mean μ_u^t and variance σ_u^t, of the 2-D user positions. Additionally, the Gaussian distribution possesses two other correlation coefficients, consisting in a total of 16 MDN values. The calculation of the optimal error is composed of several nodes per layer that are 16, 8 and 2, respectively.

The prediction horizons will be defined in term of seconds, due to high impact on the learning method and the backhaul offloading. We use a control environment for managing the learning time and segment update using short-term horizons of [50, 100, 150, 200, 250, 300] seconds, to find the optimal learning accuracy.

In Fig. 15.3, it is possible to address the results of the sampling technique for each Gaussian mixture, providing optimal sampling values for a defined velocity. Moreover, we notice an increase in sampling seconds per mixture that provides less computational cost at each increment. Additionally, we demonstrate the accuracy of the learning diagnosis using the sampling technique for the two mobility scenarios. Measuring the accuracy over 300 epochs (cycles through the full training dataset), we obtain higher accuracy values for the Mix_4 in both scenarios. However, the higher accuracy values belong to the HS scenario, where the network can learn fast optimal user positions due to highly correlated user behaviour.

Furthermore, we measure the efficiency of our OCU compared with the content classifier technique proposed by [18], where they propose a content update technique estimating the optimal content popularity using a Content Classifier Policy (CCP). Firstly, we measure the efficiency of both algorithms using the popularity estimation of both techniques. Furthermore, we use the setup for the two mobility scenarios (HS and Lagg), using content segmentation with a s_m/Δ size of 12.5 Mb. Additionally, we set the simulation to the shortest time horizon which is 50 s to predict an optimal content update for optimal backhaul offloading.

In Fig. 15.4a, we notice that our technique outperforms the CCP during the offloading due to the optimal traffic estimation. Moreover, the traffic offloading experiences high variability due to the mobility dynamics. It is also possible to observe coherent variations when the content update based on this phenomenon. In this context, both techniques consistently reduce the overall network load; however, our OCU structure provides optimal results in both mobility scenarios. In Fig. 15.4b, the CHR consistently increases along with the optimal content update policy. The motives are that both techniques take cache decisions according to the content popularity. However, our solution is more accurate with traffic estimation along the

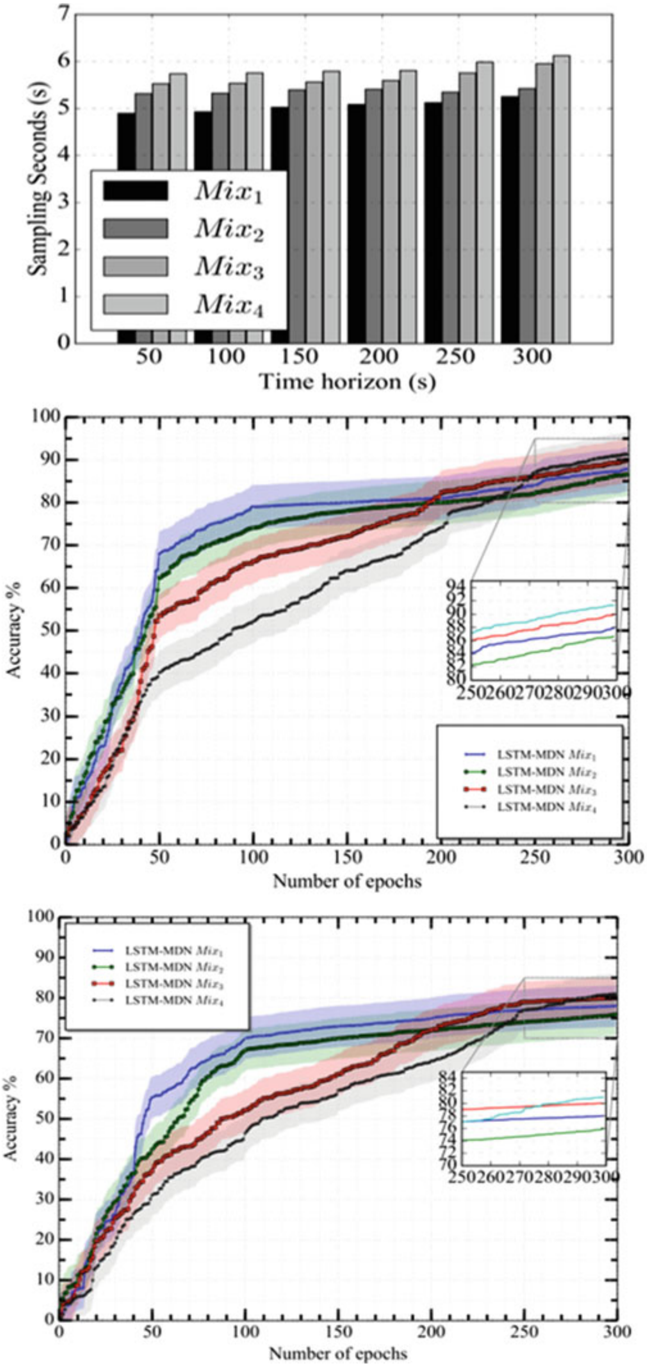


Fig. 15.3 (**a**) Sampling mobility, (**b**) accuracy over HS aggregation mobility scenario and (**c**) accuracy over Low aggregation mobility scenario

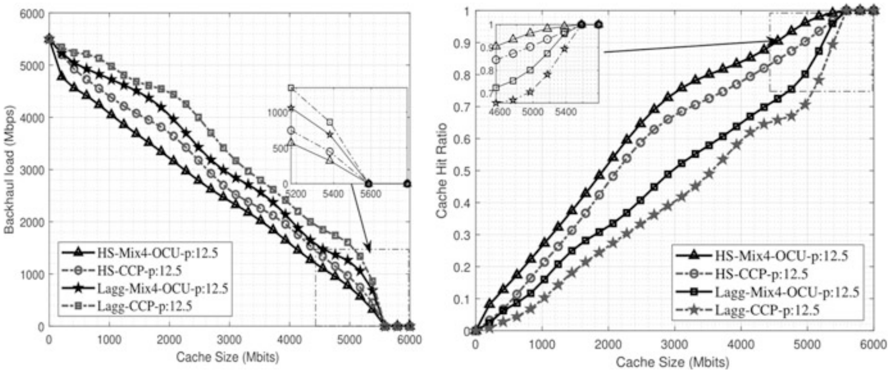


Fig. 15.4 (**a**) Backhaul offloading (left) and (**b**) cache hit ratio (right) for two learning approaches

shortest horizon 50s. Our OCU provides high accuracy in the shortest time horizon, where the other methods eventually fail. In fact, it can be explained by considering the high mobility scenario, which causes dynamic traffic variations and increases the learning difficulties in short time periods.

15.2.6 Conclusion

In this section, we have discussed an accurate proposal to obtain optimal content updates based on estimating the traffic dynamics and its dependencies over time. As a case study, we investigate the impact of mobility and popularity estimations on performance and to obtain the optimal content update technique. The performance assessment of the OCU has notable learning trade-offs between the hotspot and low aggregation mobility scenarios. Our results clearly show the importance of considering realistic mobility models for content placement analysis that allows the MNO's to analyse the accuracy of the learning methods in highly dynamic scenarios. Finally, these baselines motivate the MNO's to improve learning traffic approaches for content caching based on their effectiveness over several time periods.

15.3 Mobile Edge Content Placement Strategy for B5G Networks

15.3.1 Introduction

In current cellular network, the monthly content traffic volume is increasing very fast. In the cellular network, there are three key challenges that network operators should mitigate to meet the required quality of experience (QoE). The first challenge is that the network backhaul has limited and inflexible capacity; therefore the increase in traffic volume causes network congestion. The second challenge is, that network bandwidth is mainly used by most frequently requested, known as *popular*, contents. These contents are frequently transmitted from the backhaul server to the end-user equipment (UE), which incurs higher transmission costs [21]. The third bottleneck is that the 5G and beyond networks have a milestone of extremely low service latency. However, contents served from the backhaul network must travel a long distance and pass through multiple routers before they are delivered to users, which forces the network to exhibit a high response delay. To address these challenges, *content caching* is being extensively used. It is a disruptive way of prefetching contents to the radio access network for further use. The main goal of edge caching is to reduce congestion and service delay so that the QoE is improved by trading off communication costs for storage resource [22].

Several edge devices such as macro base station (MBS), micro(sub) base stations (SBS), femto base stations (FBS) and smart UEs are deployed in the current cellular network. We call these devices as cache-enabled cellular network edges or mobile helpers (MH). Now, they are having high computational capacity to process the required information in the radio access network (RAN). In addition, they have sufficient storage capacity to cache contents [23]. We can further enhance the performance of the caching using new technologies such as multi-access edge computing (MEC) and the device-to-device technologies. Hence, these MHs can run applications and information within RAN without need of central server. These technologies help to solve challenges of ultra-low latency requirement, real-time processing, location ware and personalized high QoE to a large crowd of users [24]. It also helps to apply a dynamic data rendering and preprocessing tasks, which eases content delivery, mainly for mission-critical services.

15.3.2 Enhancing Techniques for Content Caching

As stated in subsection 15.3.1, there are several technologies proposed to enable the content caching process and enhance the QoE. However, optimization of content management, whether during consumption by the network or by the enabling technologies in the cellular networks, should be addressed to further boost the network performance; thus, network resource optimization will ultimately bring a double benefit. Moreover, it is worthy to note that in content management, both *content caching* and *content delivery* are interdependent and one step may affect the other. The placement can be uncoded (where the entire or part of a file is placed to a caching edge, but they are not transcoded [25]) or coded (where the content is transcoded by some information coding techniques such as fountain codes [26, 27]).

Content placement is a resource-constrained process, which can be modelled by the knapsack problems like in [28, 29]. To solve it optimally, different algorithms are proposed. In [21], the authors used integer programming, with an end-to-end approach, which is impractical to solve optimally. Similarly, [30] uses dynamic programming (DP) to maximize network performance gain in terms of customer satisfaction. Recent work in [31] has used DP to solve content placement by modelling the placement as knapsack problems for a given but different popularity and size of contents. The survey work in [27] has explored the most widely used content caching problems in terms of modelling and solving algorithms them. For each content placement model, many algorithms are used to solve them. However, most of them follow greedy and heuristic solutions [22]. Although simpler, they are not optimal solutions, and they cannot deal with some constraints such as the content size heterogeneity, link capacity, caching capacity and heterogeneity of caching edges.

In this section, we focus on a content caching strategy by modelling it as a multiple knapsack problem and solving using the DP. Multiple MHs are clustered around a central head, called the MBS, where the objective function is to maximize the cache hit probability (CHP) value of the content within the cluster.

15.3.3 Content Placement at Caching Edges

We consider a downlink direction of a heterogeneous cellular network, logically represented in Fig. 15.5. The first tier contains MBS which feeds contents from the backhaul network to several SBSs. Both the MBS and SBS have different caching capabilities. All types of SBS serve as MH and have different and finite cache sizes, in bits. The MHs are found at the second tier and deliver contents to associated UEs, found at the third tier. The MBS has list of video contents, defined as $\mathcal{M} = \{(f_i, l_i) : 1 \leq i \leq |\mathcal{M}|\}$, where f_i is the unique identifier of a content in the library and ℓ_i is its size, in bits. The MHs are assumed to be centrally controlled by the MBS while placing contents in their cache, where the placed content resides in its entirety, or as fragmented subcontent. These MH edges also cooperate while delivering the contents to the UEs that are found in their coverage area.

In each computation period, each content is assumed to have a known popularity vector, $\mathcal{P}$=$\{\rho_1,\ \rho_2,\cdots,\rho_{|\mathcal{M}|}\}$ where $\sum_{i=1}^{|\mathcal{M}|}\rho_1 = 1$, as seen by the MBS. The popularity is defined as the probability that a content to be requested by UEs, associated to an MH. This $\mathcal{P}$ can be derived through localized popularity estimation techniques, for example, the DL-based estimation proposed in Sect. 15.2, applied at the MBS. Mainly, the popularity is calculated based on the frequency of requests that the contents received from users in the speculated period. The MBS collects the content request profile from each MH and make necessary preprocessing on the information.

We focus on the content placement process that takes place at the MHs, which are monitored by the MBS. We assume that the content selection process from library $\mathcal{M}$ is done by the central MBS in such a way that neither a subcontent nor a

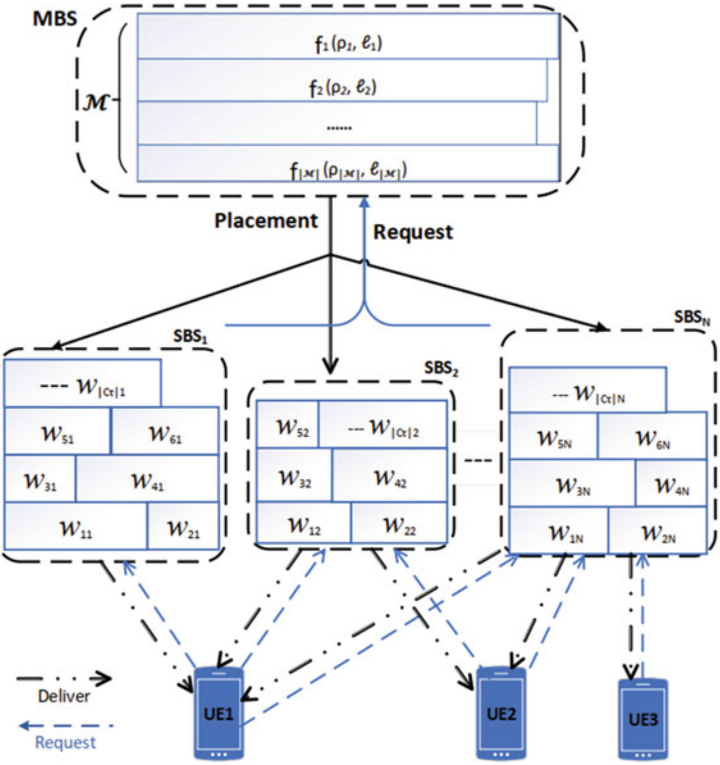


Fig. 15.5 The MBS places non-overlapping subcontents to cooperating multiple MHs

complete content is placed in more than one MH. During the placement step to the MHs, contents can be coded into smaller parts (*subcontents, w*) based on the ratio of requests they receive from each MH. Once the (sub)contents are cached at the MHs, and when a new request comes from a UE, the MHs communicate to each other and serve the respective (sub)content to the requesting device, through the appropriate path that the MHs decide. If no trace of the requested content is found in any of the MHs, then it is directly served from the MBS. We denote the set of contents that are eventually placed at j^{th} tagged MH by $C_j, j=1, \ldots, N$.

The aim is to select non-overlapping subsets C_j of popular contents for each MH that maximizes the objective function. While doing maximizing the CHP, the sum of sizes of cached contents in each MH should not exceed the cache size of respective MH. Here, our objective function is called the *cache hit probability* (CHP). The CHP is the sum of probabilities that a file requested by a UE is found cached in the cluster and can be retrieved from any of the MHs. Given C as a set of all cached sub)contents in all the MHs and vector ρ_{ij} for popularity values when subcontent i cached at j, the CHP value is:

$$\Psi_{\mathcal{C}} = \sum_{j=1}^{N} \sum\nolimits_{i=1}^{|\mathcal{C}_j|} \rho_{ij}, \quad f_i \in \mathcal{C}_j \tag{15.7}$$

To maximize the objective function ($\varPsi_C$) of the entire set of MHs, we optimally select contents whose subcontents fit to every MH, at a minimized time complexity. Hence, the objective function is defined on $\mathcal{M}$ as:

$$\Psi_{\mathcal{M}} = \max\left(\Psi_{C}\right), \quad j = 1, 2, \ldots N \tag{15.8}$$

subjected to:

$$\sum\nolimits_{i \in \mathcal{C}_j} \ell_i \le L_j, \quad j = 1, 2, \ldots N$$

where the $2^{\mathcal{M}}$denotes all the possible subset of contents taken from $\mathcal{M}$. In this content placement formulation (15.8), each MH wants to maximize its objective function without having a duplicated content. This makes the problem quite complex to solve so that there is no optimal solution yet. Instead, by taking the assumption that contents are portioned to subcontents based on the interest of MHs, we devise a simple technique that provides a suboptimal result, in two steps, as shown in the following sections.

Step 1. Content Selection by Relaxation

The straightforward task to maximize problem (15.8) is to select most popular contents from the initial library $\mathcal{M}$, assuming that we are going to place them in a single cache. Hence, we apply a combinatorial selection process at the MBS by redefining problem (15.8) as:

$$\Psi_{\mathcal{M}} = \max \sum_{j=1}^{N} \sum\nolimits_{i=1}^{|\mathcal{M}|} \rho_i \; x_{ij} \tag{15.9}$$

subjected to:

$$\sum_{i=1}^{|\mathcal{M}|} \ell_i x_{ij} \le L_j, \quad j = 1, 2, \ldots N$$

$$\sum_{j=1}^{N} x_{ij} \le 1, x_{ij} \in \{0, 1\}, i = 1, 2, \ldots |\mathcal{M}|$$

where x_{ij} is a decision parameter whether f_i is to be cached at j^{th} MH. To solve (15.9), we can apply simple relaxation to the first constraint and get the upper bound.

Surrogate relaxation is very practical because the contents can be easily partitioned, and we have many items to be placed at a few MHs. Let (λ_1,...,λ_j) be a vector of positive multipliers which satisfies: $\sum_{j=1}^{N}\lambda_j \sum_{i=1}^{|\mathcal{M}|}\ell_i x_{ij} \leq \sum_{j=1}^{N}\lambda_j L_j$. As proved in [32], the optimal vector of multipliers that give us tightest upper bound is found when λ_1=λ_j =κ, $\kappa > 0$. Hence, problem (15.9) is reduced to:

$$\Psi_{\mathcal{M}} \approx \max \sum_{i=1}^{|\mathcal{M}|} \rho_i \; y_i \tag{15.10}$$

subjected to:

$$\sum_{i=1}^{|\mathcal{M}|} \ell_i \; y_i \leq L, \quad y_i \in \{0, 1\}$$

where $y_i = \sum_{j=1}^{N} x_{ij}$ is aggregated decision parameter and $L = \sum_{j=1}^{N} L_j$ is the cluster cache size. Placement problem in (15.10) is simple 0/1 knapsack problem that can be optimally solved using DP. Since we partition these selected contents into fitting subcontents to each MH, except the last which might not fit, this value is almost surely equal to the optimal value of (15.9).

Step 2. Balanced Content Assignment to MHs

Assume that MHs can cooperatively deliver subcontents to users and each content has unique global popularity ρ_i at the MBS. Further we assume that partitioning a content in terms of its size has the same partitioning effect on its popularity. Hence, once we optimally select the contents to be cached from the MBS, we have the freedom to partition them using any coding scheme and assign subcontents to the MHs without transcoding. They are filled to MH caches in random size order.

1. *Minimized content partitioning:* We iteratively choose contents to place into all caches, in any order of cache size. After fitting as many contents as possible to cache L_j, the overhead portion of the first unfitting content and other free contents are placed to L_{j+1}. Therefore, we have maximum of $N - 1$ files to be partitioned, while this number can be further reduced by reselecting subset whose sum of content sizes exactly fits to the cache size of j^{th} MH, by using the *subset sum problem (SSP)*. In this case, if all contents are assigned, no content is partitioned, and this gives the optimal solution of the entire problem. However, this suppresses the interests of MHs towards each content, and traffic load will be unbalanced because the SSP deals only with size fitting.
2. *Full content partitioning:* We partition almost all contents to N subcontents (w_{ij}), as shown in Fig. 15.5. For that, we use the ratio of number of requests in the system to every content (where assumed that there is no redundant request to a single content from same user through different MHs). Let r_{ij} be the number of requests to content f_i ($f_i \in \mathcal{M}$) through j $^{\text{th}}$ MH, in specific period. Then, the total request to the content, at the MBS, is $r_i = \sum_{j=1}^{N} r_{ij}$. We can use

this request history to determine subcontent parameters: size ($\hat{\ell}_{ij}$), popularity ($\hat{\rho}_{ij}$) and decision parameters ($\hat{y}_{ij}$), for each subcontent using a partitioning rate $R_i = \frac{r_{ij}}{r_i}$, where $\sum_{j=1}^{N} R_i = 1$. This directly infers that the largest portion of the content is served to a UE from the MH where the content is the most popular, while no portion might be cached at the MH where no request is made through. A proportional popularity vector $\mathcal{P}_j = \{\hat{\rho}_{1j}, \hat{\rho}_{2j}, \ldots . \hat{\rho}_{ij}\}$ is created for each j MH where w_{ij} is assigned. *Hence*, the CHP value for each MH is:

$$\Psi_{\mathcal{C}_j} = \sum_{i=1}^{|\mathcal{C}_j|} \hat{\rho}_{ij}, \ j = 1, 2, \ldots \mathrm{N} \tag{15.11}$$

Subject to:

$$\sum_{i=1}^{|\mathcal{C}_j|} \hat{\ell}_{ij} \le L_j,$$

$$\sum\nolimits_{j=1}^{N} \hat{\rho}_{ij} = \rho_i; \sum\nolimits_{j=1}^{\mathrm{N}} \hat{\ell}_{ij} = \ell_i$$

In whatever assignment we make, the cluster CHP is not changed, i.e., $\sum_{j=1}^{N} \Psi_{\mathcal{C}_j} = \Psi_{\mathcal{M}}$; which means, the partitioning preserves the availability of contents in the set of MHs. This content assignment approach creates a fair availability of contents of interest to each MH. Upon a request, the associated MHs cooperatively serve subcontents to UEs.

Service Reward The content placement scheme has direct impact on the instantaneous service reward and on the amount of data downloaded while serving a request. The service reward measures the amount of data capacity required while serving UEs. Since the popularity distribution is known and contents have no redundancy in the cluster, we can adapt the average service reward formulation (Φ) of all cached contents, from [33], as follows.

$$\Phi = \sum\nolimits_{j=1}^{N} \sum\nolimits_{i=1}^{|\mathcal{C}_j|} \hat{\rho}_{ij} \hat{\ell}_{ij} \tag{15.12}$$

Compared to assignment without partitioning scheme, the proposed strategy rewards by decreasing the downloading capacity demand by nearly a factor of N. This is particularly useful when the content caching happens on the loaded backhaul network.

15.3.4 Numerical Results

In this section, we evaluate the performance of proposed content selection and subcontent assignment strategy by comparing with iterative strategies, which fill the caches of the MHs one by one. They use the *greedy*, DP and the *random* selection methods. In all strategies, the MHs are filled in increasing order of their cache size, to avoid biasing.

In the analysis, the number of contents is fixed to $|\mathcal{M}| = 100$, and their popularity is assumed to follow a *Zipf* distribution [5], while size of contents follows an exponential distribution. We consider a mean value (μ) of the content size distribution to be 1.23 Gb, which corresponds to a video duration of 4 min with 720p resolution at rate of 5 Mbps, according to [6]. We use three MHs where total available cache size (L) increases from 10 Gb to 100 Gb, wherein the three cache sizes share is $L_1 = 0.2{*}L$, $L_2 = 0.35{*}L$ and $L_3 = 0.45{*}L$.

Figure 15.6a, evaluates the performance of three algorithms used to solve the relaxed problem (15.10). Since we can partition the contents and the entire cache size of the cluster (L) is used, this gives a nearly optimal solution when using the DP. The result shows that the DP algorithm outperforms other widely used baseline strategies: greedy and random. Given this result, Fig. 15.6b depicts that the proposed content selection strategy (by relaxation and solved using DP) outperforms other three iterative strategies. The first iterative baseline (*iterative-DP*) strategy uses the DP itself to fill the MH caches one by one, which gives better performance than other two iterative strategies, which use the greedy and random methods. But when the cache sizes are relatively larger, also the iterative strategy using random (*iterative-*

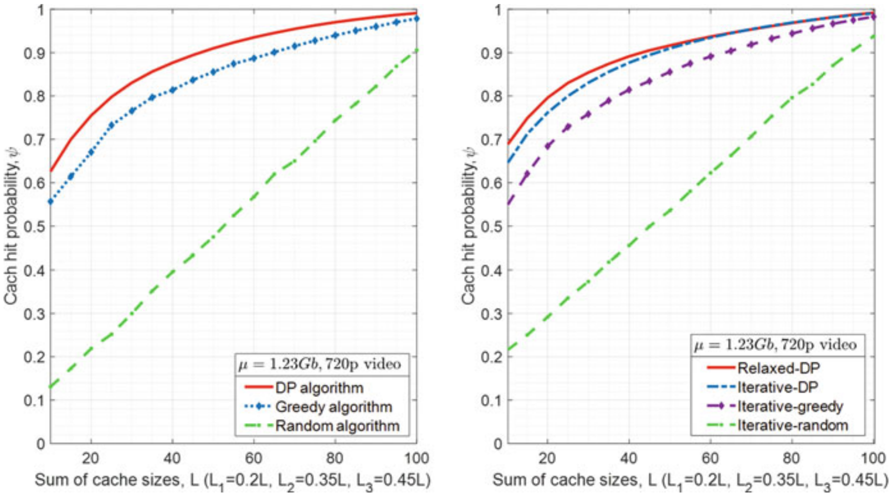


Fig. 15.6 Content selection results: (**a**) the surrogate relaxation solved using different algorithms, (**b**) proposed content selection strategy (the relaxed one and using the DP) compared to iterative baseline strategies (right)

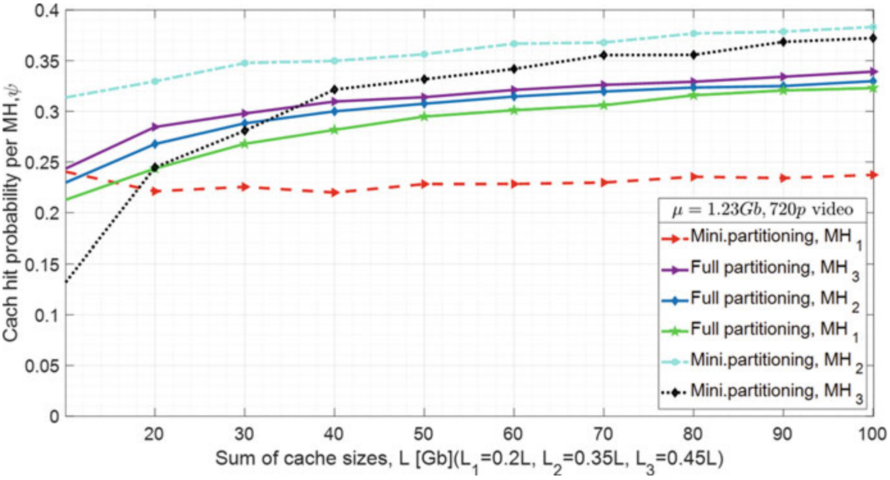


Fig. 15.7 Cache hit probability of individual MHs with different subcontent assignment strategies

random) performs well, while others are nearly equally performing. This is because there is enough space to cache contents. This implies that the role of proposed strategy is not significant for clusters with higher caching capacity, compared to the sum of popular contents' size in the library.

The result in Fig. 15.7 compares the impact of partitioning scheme to assign subcontents to the MHs. It shows that the CHP values for each MHs is close to each other when we use the proposed strategy (*full partitioning*), unlike for the case of *minimized partitioning* technique. This shows that the proposed content selection and assignment strategy places contents in a fair fashion, in their subcontents format that are synthesized based on the interest of the MHs. In contrast, the *minimized partitioning* technique, which places almost all contents without partitioning, incurs a very wide performance gap between the MHs. This shows that not only the contents are biasedly assigned to MHs, but also the performance depends on the MHs cache size and the order of filling them. More interestingly, the performance gaps between MHs for the proposed assignment method is consistent with increasing the cluster cache sizes because the parameters of the assigned subcontents are delimited by the independent request profile from each MH.

Figure 15.8 depicts that the proposed full partitioning strategy highly lowers the link capacity budget to serve the contents through the links to each MH. In this case, the total service rewarding gain is very high (means the downloading demand is reduced) due to extensive partitioning to contents. In addition, the proposed strategy fairly balances the demanded capacity load in contrast to the minimized partitioning assignment scheme, between all the MH cache sizes. But in the minimized partitioning case, the two MHs are overloaded compared to the first cache. The load is steadily increasing at the two prominent cache sizes, regardless of the request interests. Harnessing the proposed straightforward strategy reduces backhaul load of cellular networks.

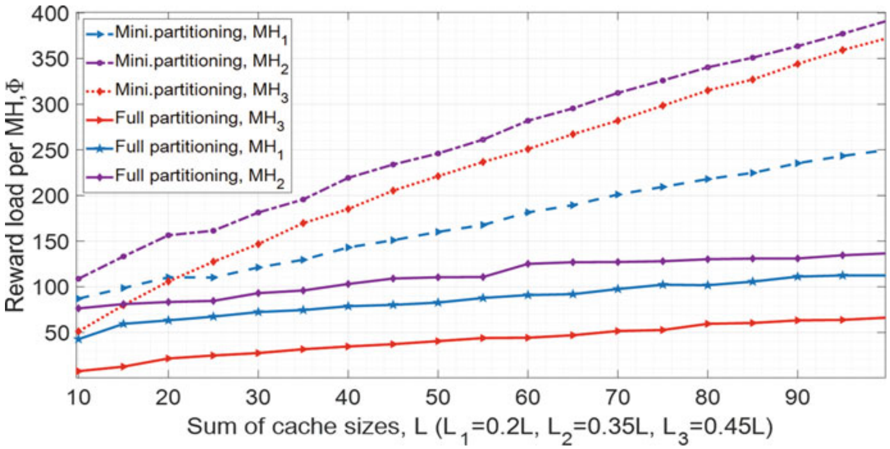


Fig. 15.8 The instantaneous serving reward for subcontents assignment strategies per each MH

15.3.5 Conclusion

In this section, we have studied content caching process where multiple network edges serve as mobile helpers to cache popular contents, of specific period, and offload macrocell base stations. The content caching is modelled by special generalized knapsack problems, where it is reduced to single knapsack problem using simple relaxation. The last form is solved by the DP while the candidate contents are decoded to subcontents and assigned to MHs based on the request rates. This scheme increases the availability of subcontents, or the entire content, at the MHs in a balanced fashion of the service load. With the computational capacity of caching edges, the placement process can be updated when necessary. Such straightforward and very efficient optimization methods boost the performance of both the content caching process and the enabling technologies. The proposed placement strategy can be further analysed in terms of reducing the latency and transmission costs.

15.4 Device-Assisted Computation Offloading for B5G Networks

15.4.1 Introduction

In this section, we introduce the core terms, technologies, and concepts needed for multi-access edge computing (MEC) as the key enabler of 5G and beyond networks, and computation offloading as one of its main use cases. We also introduce recent trends in using device-to-device (D2D) communication, as well as their motivation and use cases, and limitations and future research challenges towards D2D-assisted MEC.

When cloud computing became popular, it disrupted the traditional in-house data- centres in companies worldwide. The easy scalability, operation, and management allowed us to concentrate on software and services. When it was originally deployed, services that reached users' mobile end devices were not widespread, due to the lack of user equipment (UEs). With the advent of smartphones and other smart devices, the mobile cloud computing (MCC) paradigm emerged as an enabling technology to provide computing and communication services on demand. With growing hardware capabilities in processing and storage, innovative use cases evolved and became widespread, that included stringent requirements based on latency and reliability, that includes the AR/VR revolution, the Tactile Internet, as well as further real-time applications.

The solution to harness MCC is called MEC and refers to the concept of moving the data centre closer to the user, cutting away the transfer delay to a minimum. The edge cloud (EC) is instantiated in the base stations (BS) themselves and offers access to computational and storage services. In contrast to the centralized MCC, the MEC paradigm reduces the backhaul data traffic (compared to sending all UE service requests to the core network) [34, 35], extends user's battery life by offloading compute-intensive tasks to edge servers [36], and provides real-time information of UE locations and behaviours, which are helpful for enabling context-aware services [37, 38]. MEC also includes small cell scenarios, where one macro-BS supports multiple cells at different locations. However, with increasing demands for computing and data transmission due to emerging services, ultimately being utilized by large numbers of wireless end devices, additional MEC computing and storage infrastructure are required. However, the high cost for installation and maintenance for the telecommunication industry may limit the deployment of higher capacity MEC systems [39, 40], unless we are able to identify and exploit freely available computing resources in the networking vicinity.

The D2D communication recently emerged as enabling technology that aims to take a step towards solving the network capacity bottlenecks of the classical architecture of mobile communication. D2D communication as a radio technology addresses this shortcoming by enabling the direct data exchanges between two adjacent UEs, without any involvement of central control units or core networks [41, 42]. By exploiting the benefits of devices in a proximity to each other, this concept allows several improvements in spectral efficiency, data rates between devices, power consumption, and end-to-end delay. Notable use cases include public safety in emergency and disaster scenarios [43], energy- and data-efficient storing and sharing of video files and images [44], vehicular communication (V2X, cellular-V2X) and public safety [45], and collision avoidance [46], but also optimization of charging of electric vehicles [47]. The research question arises whether D2D can be a complement legacy to MEC approaches towards enhancing computational and storage capability.

15.4.2 Device-Assisted MEC Application for Task Offloading: Use-Case Scenario

15.4.2.1 Enhancing MEC Task Offloading with End-Devices

Fulfilling the various low latency demands while still saving the battery life of mobile devices, an attractive strategy, is to offload computationally intensive tasks to more powerful MEC servers or even to adjacent users using D2D communication links. This is especially true for idle nearby devices. The procedure also reduces the load on the cellular network infrastructure, and usage of cellular bandwidth. The concept of device-assisted MEC computation offloading via D2D communication is shown by Fig. 15.9, including so-called partial offloading of subdividable tasks from UE4 to both the MEC server and UE5, and binary offloading from UE1 via relay UE2 to helper UE3, where UE2 acts as a mobile small cell.

In the real world, the type of application plays an important role in offloading decisions. Typically, there is a distinction between partitionable and non-partitionable computations, each depending on the application scenario. Partitionable tasks consist of subtasks that may be distributed to different computational entities and are independent. Candidate (sub-)tasks for offloading are sent to a MEC server or adjacent UEs, either directly via D2D communication or via relays, or they are executed locally, depending on the computational and latency requirements of the task, among other factors. However, in typical IoT scenarios, data dependency between the different IoT sensors' tasks is non-trivial. That is, some information from other tasks is needed to execute the subsequent next tasks. A simple scenario of a device-enhanced offloading framework with a general dependency graph is introduced in the following section.

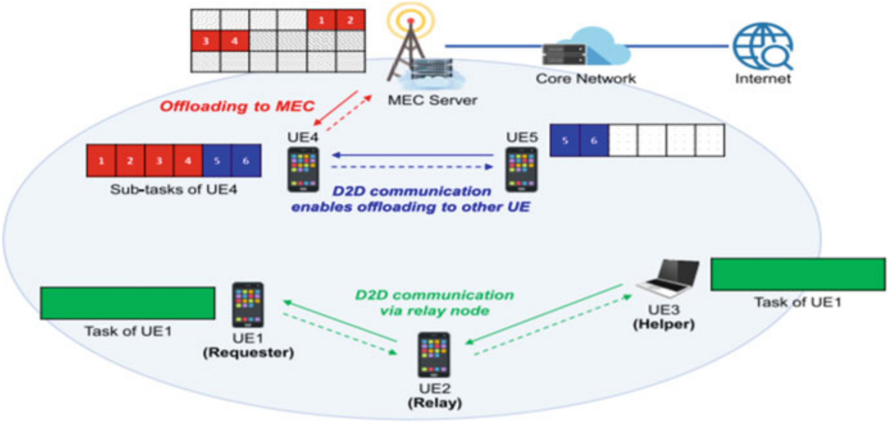


Fig. 15.9 *Device*-assisted computation offloading in MEC networks [48]

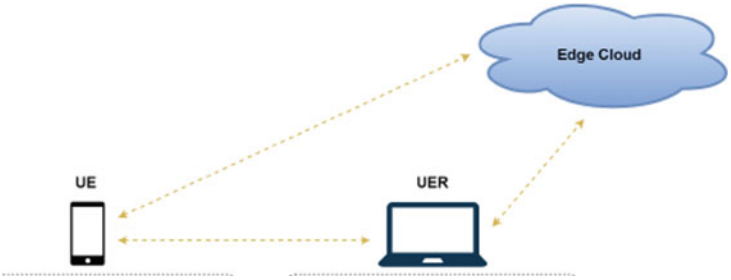


Fig. 15.10 System model [49]

15.4.2.2 Model Description

Consider a two-tier network, featuring two user equipment's (UE and UER, the latter acting either as a helper to execute (sub-)tasks directly, or relay, to forward them to the edge server), which can directly communicate with each other via the D2D network, and a MEC server node attached to a BS as depicted in Fig. 15.10. Then, tasks can be executed cooperatively using D2D communication, where two performance metrics are minimizing the total energy consumption of the device and fulfilling the time deadline constraints of the tasks. In the second layer, the edge server is assumed to have more powerful resources such as CPU frequency and the number of processing units than either of the two mobile devices. Furthermore, edge servers are typically powered by the electrical grid, so the energy consumption of edge sever is not considered at this stage. Both UEs can communicate directly with the edge server via cellular networks. In most cases, the edge server would be connected to cloud servers via wired networks (such as optical fibre network), but only the first two layers are considered, and no cloud server is present in this model.

For simplicity, the wireless channel is assumed to be quasi-static during the execution time, and the channel state information and computation-related parameters are available for devices. The computation results are assumed to be much smaller than the input data, so the time transmitting the result is negligible, and only the transmission time of the task input and the execution time are taken into consideration. There is a time deadline for completing the UE1's tasks, and it is assumed that the data size and computation resource requirements for each task are known in advance.

15.4.2.3 Proposed Offloading Decision Algorithm (OOA)

In the following, we elaborate on the proposed system model for designing our offloading approach that includes:

Task Model
In this scenario, the user's tasks are sequentially dependent meaning that each task requires the result of its predecessor tasks. The task finish time, $FT_{k,}$ is defined as task k (T_k) execution time and starting time [50]:

$$FT_k = RT_k + T_k$$

where the start time is the time that the predecessor of task k execution process is finished and defined as follows:

$$RT_k = \begin{cases} 0 & P(k) = \varnothing, \forall k \in \mathcal{K} \\ \max\limits_{j}\left(FT_j\right) & \forall j \in P(k), P(k) \neq \varnothing, \ \ \forall k \in \mathcal{K} \end{cases} \tag{15.13}$$

where $P(k)$ is a set containing task k-predecessors.

Communication Model
Since in our scenario, the size of the result of the executed tasks is small compared to the former data, the downlink data rate is neglected and we just focus on the up link. Therefore, the achievable up-link data rate for task k transmission is calculated based on the Shannon theorem

$$R_k = B\log_2\left(1 + \frac{P_k H_k}{\Gamma\sigma^2}\right) \tag{15.14}$$

where B is the bandwidth, P_k is the transmission power of task k, H_k denotes the channel power gain from sender to receiver, σ^2 is the noise variance and Γ is the coding gap as a function of bit error rate (BER), which is determined based on the coding schemes and medium access protocol. To simplify the model, we assume $\Gamma = 1$.

Computation Model
Depending on which device or server is executing the task, the time and energy needed for execution of task k are calculated based on the computation capability of the host which can be calculated based on the CPU cycles needed for execution of each bit of task, which is known in advance in our scenario. The time needed for execution of task k can be written as:

$$T_k = \frac{d_k w_k}{f_k} \tag{15.15}$$

where d_k and w_k are data size and required CPU cycles per bit of task k, respectively. Here, f_k denotes the CPU frequency of chip for computing task k, and the CPU frequency is assumed to remain the same during processing task k. The energy consumption for execution of task k can be then expressed as follows:

$$E_{pk} = e(f_k)^2 w_k d_k \tag{15.16}$$

The energy consumption required to offload the task can be formulated as:

$$E_{tk} = P_k \frac{d_k}{R_k} \tag{15.17}$$

where *e(f$_k$)* is the effective switched capacitance of the chip.

There are four possible execution scenarios, namely, local execution on the device, execution on the helper, execution on the server using direct cellular communications, and finally execution on the server through transmission of the task via a relay node.

Scenario 1 – Local Execution
Based on the described models in (15.15), the computation execution time in this mode can be written as:

$$T_k^{loc} = \frac{d_k w_k}{f_k^l} \tag{15.18}$$

And the energy consumption is:

$$E_k^{loc} = e_l \left(f_k^l\right)^2 d_k w_k \tag{15.19}$$

Scenario 2 – Assisted Execution
The task is offloaded from UE to UER, and the UER plays the role of helper for UE to execute its computation task. The data rate from UE to UER is:

$$R_k^{lh} = B \log_2 \left(1 + \frac{P_k H_k^{lh}}{\sigma^2}\right) \tag{15.20}$$

The computation execution time consists of two parts, the communication and computation time:

$$T_k^h = \frac{d_k w_k}{f_k^h} + \frac{d_k}{R_k^{lh}} \tag{15.21}$$

The energy consumption also consists of two parts, the communication and computation cost. The formula can be expressed as:

$$E_k^h = e_h \left(f_k^l\right)^2 d_k w_k + P_k \frac{d_k}{R_k^{lh}} \tag{15.22}$$

Scenario 3 – Remote Execution

Under this offloading strategy, the task is offloaded to the edge server which has more powerful computation ability and energy supply. The UE transmits the subtask to the edge server directly. After remote execution, the result is sent back to the UE.

The data rate from UE to edge server is:

$$R_k^{ls} = B \log_2 \left(1 + \frac{P_k H_k^{ls}}{\sigma^2} \right) \tag{15.23}$$

The time to complete remote execution is:

$$T_k^{ls} = \frac{d_k w_k}{f_k^s} + \frac{d_k}{R_k^{ls}} \tag{15.24}$$

Because the edge server is powered by the grid, here we only consider the energy consumption for communication:

$$E_k^{ls} = P_k \frac{d_k}{R_k^{ls}} \tag{15.25}$$

Scenario 4 – Assisted Remote Execution

The task is still offloaded to the edge server like in remote execution, but the difference is that UE1 does not send the data to the server directly, but with help from UER. This time, the UER plays the role of a relay, receives data from UE, and then transmits the data to the edge server. This can be a good strategy when the wireless channel state between the UE and edge server is far worse than the channel state between the UER and the edge server.

The data rate from UER to edge server is:

$$R_k^{hs} = B \log_2 \left(1 + \frac{P_k H_k^{hs}}{\sigma^2} \right) \tag{15.26}$$

For this offloading strategy, the subtask completion time is:

$$T_k^{lhs} = \frac{d_k w_k}{f_k^s} + \frac{d_k}{R_k^{lh}} + \frac{d_k}{R_k^{hs}} \tag{15.27}$$

And the energy consumption is:

$$E_k^{lhs} = P_k \frac{d_k}{R_k^{lh}} + P_k \frac{d_k}{R_k^{hs}} \tag{15.28}$$

The four different strategies can be expressed as x_k, y_k, z_k, v_k ∈{0, 1}, ∀ $k \in K$, and all possible offloading strategies are included in set S = {x_k, y_k, z_k, v_k}, k = {1,2, . . . , K}. A task can only be executed on one device; therefore, only one of the four variables for subtask k can be 1. The offloading decision should satisfy the following constraint:

$$x_k + y_k + z_k + v_k = 1, \forall k \in \text{Ƙ} \tag{15.29}$$

By considering the offloading strategies and the aforementioned Eqs. (15.15, 15.16, 15.17, 15.18, 15.19, 15.20, 15.21, 15.22, 15.23, 15.24, 15.25, 15.26, 15.27, 15.28 and 15.29), the completion time and energy consumption of subtask k can be expressed as:

$$T_k = x_k T_k^{loc} + y_k T_k^{h} + z_k T_k^{ls} + v_k T_k^{lhs} \tag{15.30}$$

$$E_k = x_k E_k^{loc} + y_k E_k^{h} + z_k E_k^{ls} + v_k E_k^{lhs} \tag{15.31}$$

Now, the optimization problem can be formulated aiming to minimize the total energy consumption of the system. The optimized values are offloading strategies and the start time of each task. The formulation is as follows:

$$\begin{aligned} \text{P1:} \quad & \min_{S} \textstyle\sum_{k=1}^{K} E_k \\ s.t. \quad & x_k,\ y_k,\ z_k,\ v_k \in \{0,1\}, \forall k \in \text{Ƙ} \\ & x_k + y_k + z_k + v_k = 1, \forall k \in \text{Ƙ} \\ & FT_K \leq T_D \\ & RT_k = \begin{cases} 0 & P(k) = \emptyset, \forall\, k \in \mathcal{K} \\ \max\limits_{j}(FT_j) & \forall\, j \in P(k), P(k) \neq \emptyset,\ \forall\, k \in \mathcal{K} \end{cases} \end{aligned} \tag{15.32}$$

In problem P1, the first and second constraints are limitations of the offloading strategies, the third constraint gives a time deadline in which to finish the whole task, the fourth constraint means the start time of task k should be the finish time of its predecessor and the last constraint shows that there is only one start task. Due to the existence of a non-linear constraint, the problem is a non-linear programming (NLP) model, and since the first constraint is integer, the problem then becomes mixed integer. Thus, this optimization problem can be classified as a mixed-integer non-linear programming (MINLP), which is NP-hard problem. To solve such a problem, we first transform it into a homogeneous quadratically constrained quadratic programming (QCQP), and then, using semidefinite relaxation (SDR), we obtain an approximation of the original problem. Finally, a randomization method [50] is applied to obtain an optimum offloading strategy.

Table 15.1 Simulation parameters

Parameters	Value
Number of (sub)-tasks	25
Bandwidth	5 MHz
Noise covariance	10×10^{-9} W
Channel gain from UE1 to UE2	10×10^{-3}
Channel gain from UE1 to server	10×10^{-7}
Channel gain from UE2 to server	10×10^{-5}
Transmitting power	150 mW
CPU frequency of UE1	0.4×10^{9} cycles/s
CPU frequency of UE2	0.5×10^{9} cycles/s
CPU frequency of server	2×10^{9} cycles/s
Effective switched capacitance of UE1	1.5×10^{-27}
Effective switched capacitance of UE2	1×10^{-27}
Time deadline	4 s
Sets of strategies	200

15.4.2.4 Performing D2D Task Offloading

In this scenario, the user's tasks are generally dependent, meaning that each task requires the result of its parents in the graph that forms the whole tasks. The key parameters are shown in the following Table 15.1:

The performance of our optimized offloading decision algorithm OOA is compared according to the following strategies:

- OLA: Only local algorithm, all tasks are executed at UE1.
- OSA: Only server algorithm, all tasks are executed at the edge server.
- EOA: Egoistic offloading algorithm – UE1 uses an egoistic strategy where all tasks are offloaded to UE2 or the edge server.
- SOA: Self-offloading algorithm – UE2 does not help UE1 anymore; UE1 should either execute the task or offload it to the edge server.

One key optimization parameter is the energy consumption for different w (cycles/bit), a measure for the task complexity.

As it can be seen in Fig. 15.11, by increasing the computation resource requirements, the energy consumption of all methods increased, as expected. The energy consumption of OSA is constant because there is only transmission cost (we assume the edge server to be powered by the grid, so we only consider the energy consumption for communication). Our proposed offloading algorithm (OOA) can exploit the border region and shift the offloading strategy, with the least amount of power consumption. The other algorithms slow down increasingly in speed as w grows, and tasks are then offloaded to the edge server. For example, in OOA, all tasks are executed at UE1 while $w < 20$, while EOA and SOA require more energy, which demonstrates how cooperation is saving energy from a holistic perspective.

We analyse the energy consumption with varying data sizes reflecting the size of the computation (task). As can be seen by Fig. 15.12, by increasing the size of

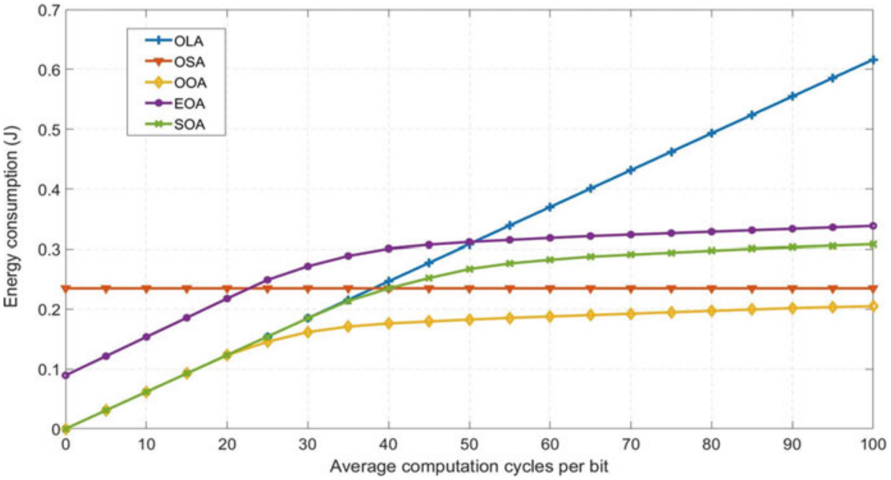


Fig. 15.11 The energy consumption versus average computation cycles per bit (w)

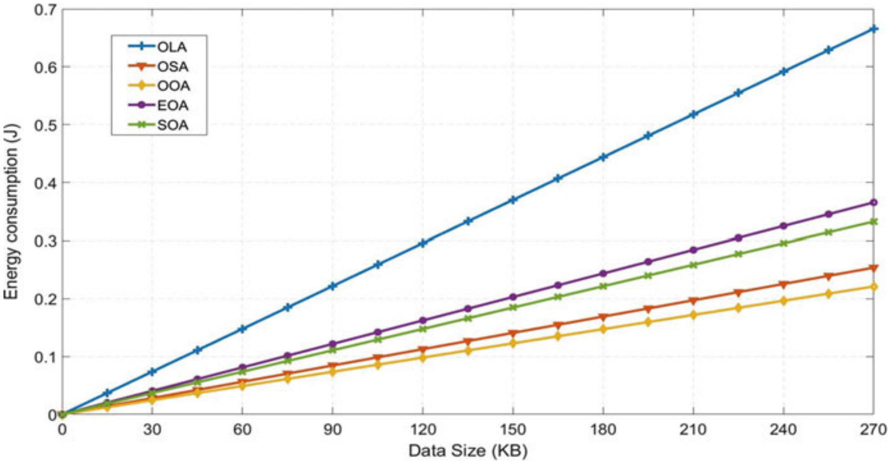


Fig. 15.12 The energy consumption versus data size (KB)

tasks, the energy consumption of all methods increases, since more bits per task are needed to be executed. However, the energy consumption of the new holistic offloading algorithm is still the lowest. The energy consumption of OLA is greater than the other algorithms as data size grows, indicating that on-device computation is infeasible, while OSA is the second-best algorithm, and EOA and SOA cost more energy than OOA and OSA. This plot demonstrates the clear superiority of our optimization scheme that opportunistically makes use of the available resources. In addition, with growing task complexity, which is a realistic assumption for future developments, the benefit of such algorithm will be even more apparent.

15.4.3 Challenges and Future Research Directions

In the proposed offloading scenario, the scenario was considered ideal. However, in practice the following challenges could arise that will undoubtedly require advanced solutions to mitigate their impact:

- *Channel interference*: if multiple UEs offload their tasks to MEC servers, or adjacent end devices that use the same communication resources (time slots, frequency channels, etc.), then interference among multiple ongoing D2D communication links, between D2D communications, and the macro cellular network arises. This interference increases with the growing numbers of UEs within a given cell coverage area. Dedicating exclusive resources to D2D communications can solve the interference problem but at the expense of reduced reuse efficiency. Multiple interference management techniques, such as power control, mode selection, and radio resource allocation are generally used jointly to improve the network capacity as well as spectrum reuse efficiency [48, 49, 50].
- *User mobility*: UE movements, including movements of either requester, relay, or helper UEs, can break the D2D links and the assumptions of quasi-static channels. Should the relay move during the transmission of tasks, for example, any link deterioration between the requester and helper will likely impact in terms of increased latencies and waste battery power. Mobility can also influence social graph information by changing social ties between UEs. Updating and predicting the availability and reliability of computation resources are a key prerequisite for enabling satisfactory user experiences and energy savings. Future research should develop and validate effective and efficient network management methods for assessing and predicting the availability of computational as well as communication resources [51].
- *Security challenges*: Security poses a considerable problem in the context of offloading computation tasks to adjacent devices. Side channel attacks [52] could allow the exploitation of UEs' personal information, and such data security breaches would likely deter users from adopting any task offloading schemes. Moreover, such security breaches could counteract eventual positive effects of offloading incentive mechanisms, which would make users lose interest in participating in the cooperation. User mobility is another issue and requires adaptive security mechanisms that allow and account for varying user locations.

  Future research should comprehensively address the security and privacy aspects of device-enhanced MEC. One path could use the social user communities as a foundation, with UEs divided into different groups based on their social relationships, interests, and locations. Security levels within a UE group could allow users to make an informed decision as to participate in task offloading or resource sharing. Any such grouping of UEs limits the potential of collaboration opportunities, and hesitation to engage in collaborations with a nearby stranger could become a new norm. Throughout, these security considerations generate overhead for the network communication, and, thus, this needs to be carefully traded off against their benefits [51, 53, 54].

15.4.4 Conclusion

In this section, using the device-enhanced MEC concept, we investigated how we can save battery lifetime in user handsets as well considering service latency requirements for task offloading applications. We first described the concept of device-enhanced MEC systems as a promising solution to the ever-increasing Internet traffic. Then, considering the dependency relationship of IoT tasks, we proposed a new offloading method for a delay-sensitive application in an environment including one MEC server and two users. The simulation analysis results were compared with baseline approaches that considered local and server execution policies where all tasks are either implemented in the server or on the cloud. The plot demonstrated the clear superiority of the proposed OOA scheme that makes intelligent use of the available resources to reduce overall power consumption, while taking time dependency between tasks and completion time into account.

15.5 Conclusion

In this chapter, we have presented cloud-based content management for B5G, where we consider how multimedia contents and tasks can be managed effectively over the mobile network towards meeting QoE-related parameters such as network offloading, reducing service latency, enhancing the network efficiency and reducing energy consumption in computing devices.

The 5G and beyond networks are becoming virtual machines, thanks to the paradigms such as MEC and cloud computing, among others, that will offer multiple flexibility in terms of content and task managements. Looking ahead to the future, network management will predominantly rely on machine learning processes and automations to provide a flexible cloud-based network infrastructure that is geared towards enhancing network performance for the operators and vertical business models. In this chapter, we investigated this intelligent networking infrastructure to open new opportunities in terms of content and resource management, pushing further network efficiency gains into the way we deliver rich content services and increasing the QoE to users.

In this context, Sect. 15.2 discussed the requirements and benefit of employing machine learning towards acquiring knowledge on the network attributes such as content request profile, user mobility and user density, in real traffic scenarios. Moreover, we investigated the impact of applying DP-based learning to capture the shortest zone traffic dynamics to achieve an optimal content update. As a contribution, we develop a probabilistic learning approach for dynamic network traffic rates which assists the BSs to offload the backhaul traffic periodically. In particular, the backhaul offloading efficiency is assessed for different mobility scenarios by considering content popularity heterogeneity. It was demonstrated that when edge devices employ collaborative caching schemes based on DL functionalities, it can substantially maximize the SU utilization. Moreover, we analyse the complexity of probabilistic content partition updates, where we assume that each SU updates the content partitions in a decentralized manner depending on the observed predictions. This SU MIP maximization, in turn, minimizes the backhaul traffic, hence, alleviating the short-term transport load in overcrowded mobility cases. In addition, this framework enables MNOs to potentially avoid higher learning expenses in dynamic traffic scenarios, to manage the improper use of backhaul.

In Sect. 15.3, we investigated practical and cost-efficient popular content placement policy to intelligently exploit MHs that are very close to end users. Once the content popularity vector is defined for a specific epoch, using DL-based estimation as elaborated in Sect. 15.2, combinatorial optimization was applied to select the subset of contents that can maximize the cache hit probability in the radio network. We employed surrogate relaxation to reduce the constrained content placement problem, that is initially modelled as special Generalized Assignment Problem, which is an NP-hard, to simple 0/1 knapsack problem and optimally solved it using the dynamic programming (DP). In this optimization policy, we reduced the complex content placement problem, involving cooperative multiple MHs into a unified storage unit format, by taking advantage of content partitioning. The proposed approach is compared to the baseline algorithms, which all iteratively place content without duplicating content copies to each MH. These algorithms include the DP, greedy (that selects in decreasing order of their popularity) and the random (that selects in random order of their popularity) methods. We have shown that the proposed strategy outperforms the baseline approaches in increasing the cache hit probability by effectively exploiting better the storage unit in the system. In addition, the proposed strategy highly reduces the backhaul load by reducing the required content serving capacity from the backhaul network. A study on the utility functions of content delivery is left for future work.

In the last section of the chapter, Sect. 15.4, we have explored the capability of MEC for task offloading in a bid to reduce the energy consumption in handset devices, while reducing service latency. Legacy or baseline approaches adopt task offloading onto the cloud or edge network, the latter in a bid to further reduce service latency and energy consumption by bringing the “processing task” near the end user. In this approach, we revisit the edge networking paradigm by pushing further the boundaries by including actual end-user devices in the near vicinity as part of the edge network. Thanks to D2D connectivity, user devices can form part of the cooperative cluster and share the computing load, a paradigm that is gaining importance in 6G research, that is referred to as dew computing. In this context, we propose a novel offloading decision algorithm entitled “OOA” that takes a significant step towards reducing the energy consumption in cellular networks compared to baseline approaches. The proposed scheme makes intelligent use of the available MEC resources in local nodes to reduce the overall power consumption, while taking time dependency between tasks and completion time into account. Challenges and open issues are also recommended for future research works such as handling the channel interference, user security and privacy, and mobility of users and computing edges.

Acknowledgements This research work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreements No. 722788 (SPOTLIGHT) and No. 722424 (SECRET).

References

1. Barnett, T. J., Jain, S., Andra, U., & Khurana, T. (2018). Cisco visual networking index (VNI) complete forecast update. In *s.l. Cisco* (pp. 1–78).
2. Wang, X., Han, Y., Leung, V. C. M., Niyato, D., Yan, X., & Chen, X. (2020). Convergence of edge computing and deep learning: A comprehensive survey. *IEEE Communication Surveys and Tutorials, 22*(2), 869–904.
3. Wang, F., Zhang, M., Wang, X., Ma, X., & Liu, J. (2020). Deep learning for edge computing applications: A state-of-the-art survey. *IEEE Access, 8*, 58322–58336.
4. Paschos, G. S., Iosifidis, G., Tao, M., Towsley, D., & Caire, G. (2018). *The role of caching in future communication systems and networks*. https://arxiv.org/pdf/1805.11721.pdf
5. Ketabi, R., Al Qathrady, M., Alipour, B., & Helmy, A. (2019). Vehicular traffic density forecasting through the eyes of traffic cameras; a spatio-temporal machine learning study. In *Proceedings of the 9th ACM symposium on design and analysis of intelligence vehicular networks and applications* (pp. 81–88). Association for Computing Machinery (ACM).
6. Salehi, M., Abad, H., Ozfatura, E., Ercetin, O., & Deniz, G. (2015). Dynamic content updates in heterogeneous wireless networks. *IEEE Access*. Vol. 15, pp.107–110
7. Ge, X., Ye, J., Yang, Y. c., & Li, Q. (2016). User mobility evaluation for 5G small cell networks based on individual mobility model. *IEEE Journal on Selected Areas in Communications, 34*(3), 528–541.
8. Hsu, W., Spyropoulos, T., Psounis, K., & Helmy, A. (2008). Modeling spatial and temporal dependencies of user mobility in wireless mobile networks. *IEEE Transactions, 17*, 1–14.
9. Danqing, K., Yisheng, L., & Yuan-yuan, C. (2017). Short-term traffic flow prediction with LSTM recurrent neural network. In *IEEE 20th international conference on intelligent transportation systems* (pp. 1–6). IEEE.
10. Kong, W., Dong, Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2017). Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Transactions on Smart Grid, 10*, 841.
11. Riihijarvi, J., & Mahonen, P. (February 2018). Machine learning for performance prediction in mobile cellular networks. *IEEE Computational Intelligence Magazine, 13*(1), 51–60.
12. Chen, M., Mozaffari, M., Saad, W., Yin, C., Debbah, M., & Hong, C. S. (2017). Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience. *IEEE Journal on Selected Areas in Communications, 35*(5), 1046–1061.
13. Chen, M., Qian, Y., Hao, Y., Li, Y., & Song, J. (2018). Data-driven computing and caching in 5G networks: Architecture and delay analysis. *Wireless Big Data: Technologies and Applications, 25*, 70.
14. Gal, Y. (2017). Uncertainty in deep learning. *IEEE/ACM Transactions on Audio Speech and Language Processing, 1*(1), 1–11.
15. Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports, 7*(1), 1–14.
16. Yao, S., et al. (2018). ApDeepSense: Deep learning uncertainty estimation without the pain for IoT applications. In *Proceedings of the international conference on distributed computing systems* (Vol. 2018-July, pp. 334–343). IEEE.

17. Zhang, J., & Yao, Y. (2019). Interplay of cache sizes and file popularity in coded caching. In *IEEE global communications conference (GLOBECOM)* (pp. 1–6). IEEE.
18. Feng, H., Jiang, Y., Niyato, D., Zheng, F. C., & You, X. (2019). Content popularity prediction via deep learning in cache-enabled fog radio access networks. In *IEEE global communications conference GLOBECOM proceedings*. IEEE.
19. Yuan, D., & Ahani, G. (2020). Accounting for information freshness in scheduling of content caching. In *IEEE-ICC2020*. IEEE.
20. Yu, C. (2015). Modelling user activity patterns for next-place prediction. *IEEE Systems Journal*, Vol. 11, Issue 2, pp. 1060–1071.
21. Jiang, W., Feng, G., & Qin, S. (2017). Optimal cooperative content caching and delivery policy for heterogeneous cellular networks. *IEEE Transactions on Mobile Computing, 16*, 1382–1393.
22. Zhang, S., He, P., Suto, K., Yang, P., Zhao, L., & Shen, X. (2018). Cooperative edge caching in user-centric clustered mobile networks. *IEEE Transactions on Mobile Computing, 17*(8), 1791–1805.
23. Safavat, S., Sapavath Naveen, N. N., & Rawat, D. B. (2019). Recent advances in mobile edge computing and content caching. *Digital Communications and Networks., 6*, 189.
24. Yu, Y. (2016). Mobile edge computing towards 5G, vision, recent progress, and open challenges. *China Communications, 13*, 89–99.
25. Li, X., Wang, X., & Leung, V. C. M. (2016). Weighted network traffic offloading in cache-enabled heterogeneous networks. In *IEEE international conference on communications, ICC* (pp. 1–6). IEEE.
26. Maddah-ali, M. A., & Niesen, U. (2014). Fundamental limits of caching. *IEEE Transactions on Information Theory, 60*, 2856–2867.
27. Li, L., Zhao, G., & Blum, R. S. (2018). A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies. *IEEE Communication Surveys and Tutorials, 20*(3), 1710–1732.
28. Wen, J., Huang, K., Yang, S., & Li, V. O. K. (2017). Cache-enabled heterogeneous cellular networks: Optimal tier-level content placement. *IEEE Transactions on Wireless Communications, 16*(9), 5939–5952.
29. Poularakis, K., Iosifidis, G., Argyriou, A., Koutsopoulos, I., & Tassiulas, L. (2016). Caching and operator cooperation policies for layered video content delivery. In *Proceedings – IEEE INFOCOM* (pp. 874–882). IEEE.
30. Khreishah, A., & Chakareski, J. (2015). Collaborative caching for multicell-coordinated systems. In *Proceedings – IEEE INFOCOM* (pp. 257–262). IEEE.
31. Ayenew, T. M., Xenakis, D., Passas, N., & Merakos, L. (2018). Dynamic programming based content placement strategy for 5G and beyond cellular networks. In *IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)* (pp. 1–6). IEEE.
32. Martello, S., & Toth, P. (1990). *Knapsack problems: Algorithms and computer implementations*. Wiley.
33. Gupta, A., Amuru, S., Tandom, R., Buehrer, R. M., & Clancy, T. C. (2014). Learning distributed caching strategies in small cell cell networks. In *11th international symposium on wireless communications systems (ISWCS)* (pp. 917–921). IEEE.
34. Rehmani, M. H., & Ahmed, E. (2017). Mobile edge computing: Opportunities, solutions, and challenges. *Future Generation Computer Systems, 70*, 59–63.
35. Morris, I. *ETSI drops mobile from MEC. Light reading*. September de 2017. https://www.lightreading.com/mobile/mec-(mobile-edge-computing)/etsi-drops-mobile-frommec/d/d-id/726273
36. Chen, L., Zhou, S., & Xu, J. (2018). Computation peer offloading for energy-constrained mobile edge computing in small-cell networks. *IEEE/ACM Transactions on Networking, 26*(4), 1619–1632.
37. Shi, B., Yang, J., Huang, Z., & Hui, P. (2015). Offloading guidelines for augmented reality applications on wearable devices. In *Proceedings of the ACM international conference on multimedia, Brisbane, Australia* (pp. 1271–1274). Association for Computing Machinery(ACM).

38. Nunna, S., Kousaridas, A., Ibrahim, M., Dillinger, M., Thuemmler, C., Feussner, H., & Schneider, A. (2015). Enabling real-time context-aware collaboration through 5G and mobile edge computing. In *Proceedings of the international conference on information technology-new generations, Las Vegas, NV, USA* (pp. 601–605). IEEE.
39. Al-Turjman, F. (2019). 5G-enabled devices and smart-spaces in social-IoT: An overview. *Future Generation Computer Systems, 92*, 732–744.
40. Iqbal, J., Iqbal, M. A., Ahmad, A., Khan, M., Qamar, A., & Han, K. (2019). Comparison of spectral efficiency techniques in device-to-device communication for 5G. *IEEE Access, 7*, 440–449.
41. Militano, L., Araniti, G., Condoluci, M., Farris, I., & Iera, A. (2015). Device-to-device communications for 5G internet of things. *EAI Endorsed Transactions on Internet of Things, 1*(1), 1–15.
42. Choi, Y. J., & Paul, R. (2019). Autonomous interface selection for multi-radio D2D communication. *IEEE Access, 7*, 090–108.
43. Fodor, G., Parkvall, S., Sorrentino, S., Wallentin, P., Lu, Q., & Brahmi, N. (2014). Device-to-device communications for national security and public safety. *IEEE Access, 2*, 1510–1520.
44. Lei, L., Zhong, Z., Lin, C., & Shen, X. (2012). Operator controlled device-to-device communications in LTE-advanced networks. *IEEE Wireless Communications, 19*(3), 96–104.
45. Bertram, T. (2019). Fahrerassistenzsysteme: Von der Assistenz zum au- tomatisierten Fahren. In *4. Internationale ATZ-Fachtagung Automatisiertes Fahren*. Springer Vieweg.
46. Feroz, A., Kavitha, N., Kasthuri, M., & Sudha, R. (2019). Vehicle to vehicle communication for collision avoidance. *International Journal of Emerging Technology and Innovative Engineering, 5*(7), 544–549.
47. Cao, Y., Jiang, T., Kaiwartya, O., Sun, H., Zhou, H., & Wang, R. (2019). Toward pre-empted EV charging recommendation through V2V-based reservation system. *EEE Transcations on Systems, Man, and Cybernetics: Systems, 51*, 1–14.
48. Mehrabi, M., You, D., Latzko, V., Salah, H., Reisslein, M., & Fitzek, F. H. P. (2019). Device-enhanced MEC: Multi-access edge computing (MEC) aided by end device computation and caching: A survey. *IEEE Access, 7*, 166079–166108.
49. Mehrabi, M., Shen, S., Latzko, V., Wang, Y., & Fitzek, F. H. P. (2020). Energy-aware cooperative offloading framework for inter-dependent and delay-sensitive tasks. In *IEEE global communications conference: Selected areas in communications: Cloud & fog/edge computing, networking and storage, Taipei*. IEEE.
50. Liu, F., Huang, Z., & Wang, L. (2019). Energy-efficient collaborative task computation offloading in cloud-assisted edge computing for IoT sensors. *Sensors, 19*(5), 1105.
51. Doppler, K. (2009). Device-to-device communication as an underlay to LTE-advanced networks. *IEEE Communications Magazine, 47*(12), 42–49.
52. Kim, D. H. (2014). Multi-device-to-multi-device communication in cellular network for efficient contents distribution. In *Proceedings of the IEEE international conference on consumer electronics, Las Vegas, NV, USA* (pp. 244–247). IEEE.
53. Mehrabi, M., Salah, H., & Fitzek, F. H. P. (2019). A survey on mobility management for MEC-enabled systems. In *IEEE 2nd 5G World Forum (5GWF), Dresden, Germany* (pp. 259–263). IEEE.
54. Spreitzer, R., Moonsamy, V., Korak, T., & Mangard, S. Systematic classification of side-channel attacks: A case study for mobile devices. *IEEE Communication Surveys and Tutorials, 20*(1), 465–488.

Conclusion: Enabling 6G Mobile Networks

Jonathan Rodriguez, Christos Verikoukis, John S. Vardakas, and Nikos Passas

The future society is heading toward an increasingly digitized world, which is connected and data-driven, where many services will be dependent on instant and virtually unlimited connectivity. As fifth-generation research reaches the twilight, the research community must go beyond 5G and look toward the 2030 connectivity landscape, namely, 6G. It is worthy to note that it is not clear exactly what 6G will be, but most certainly, it will consider immature technologies as part of the drive toward beyond 5G, but more specifically, it will be influenced by the way in which data is collected, processed, transmitted, and consumed within the wireless network.

5G technology was driven by the commercial operators to accommodate future capacity requirements for their customer base, as well as complemented by enhanced productivity demands from industry in the shape of IoT (Internet of Things). The technical success of 5G hinges on enabling technology that will deliver a much wider range of data services to a much broader variety of devices and users. However, 6G will be more encompassing in terms of communication requirements, being more society centric in terms of requirements; in addition to industry requirement, it is widely accepted that the 6G drive will be influenced by global policy on sustainability goals for an ageing and growing population, as well as addressing societal challenges. The aim is to deliver a 6G architecture that

J. Rodriguez
Instituto de Telecommunicações, Campus Universitário Santiago, Aveiro, Portugal

Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd, UK

C. Verikoukis
Centre Tecnològic de Telecomunicacions de Catalunyal, Parc Mediterrani de la Tecnologia (PMT), Castelldefels, Spain

J. S. Vardakas
Iquadrat Informática SL, Barcelona, Spain

N. Passas
University of Athens, Panepistimiopolis, Athens, Greece

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3

promotes digital inclusion and accessibility, as well as unlocking economic value and opportunities in rural communities.

Harnessing on the plethora of services offered by 5G technology, 6G aims to integrate an even richer set of services to its portfolio, which includes virtual and augmented reality, and even mixed reality, telepresence, and autonomous vehicles for ecological transport and logistics. This will be based on introducing new enabling technology that can target ambitious KPIs (key performance indicators) and that factor in 10–100 times more capability over 5G networks. This will require disruptive architectures that can build on 5G technology, to deliver market relevant solutions.

Indeed the 6G vision is wide-reaching to cater for many vertical sectors, and it is not the purpose of this book to cover all aspects on 6G, but to provide a primer on the "fundamentals of 6G Mobile networks" that is characterized by UDN deployment, optical-wireless convergence, and cloud-based services through softwarization; as such, this book is self-contained and was structured accordingly.

Targeting 4 main parts and 15 chapters, we aimed to provide an insight into the 6G odyssey starting from 5G as a baseline toward the latest developments on 6G worldwide.

Part (I) entitled the "5G and Beyond Mobile Landscape" presented the 5G and beyond landscape, providing the interested reader with insights into future emerging 6G use cases, enabling technologies, and system requirements. In one comprehensive chapter, we opened the discussion for several promising use cases for B5G/6G such as holographic telepresence, digital twin, autonomous vehicles, and distributed AI, among others, which can address communications challenges such as "digital transformation" and "fully connected society" by the year 2030. To this end, several enabling technologies were discussed for B5G/6G, noticeably on network virtualization, TeraHertz, and Visible Light Communications, among others. The chapter reviewed 6G system requirements and the foreseen architecture, along with ongoing 6G research activities around the globe, highlighting future research opportunities and challenges.

Part (II) entitled the "UDNs for B5G" highlights the UDN core and provided insights into technology challenges and solutions for enabling the hyper-dense deployment of small cells. B5G will harness massive-scale MIMO and distributed architectures to support the notion of cell-free MIMO, and therefore the optimal operating conditions are investigated here, giving valuable insights toward their future deployment. It was concluded that uplink cell-free massive MIMO systems with transmit and receive antennas in LoS (Line of Sight) are not considered favorable propagation conditions motivating the use of multistage receivers: on the contrary, they hold in the case of Rayleigh fading. To deal with the asynchronous transmission between network nodes in cell-free architectures, filter bank-based multicarrier (FBMC) waveforms have been proposed as a potential alternative to legacy 5G CP-OFDM (Cyclic Prefix-Orthogonal Frequency Division Multiplexing) due to better spectral efficiency and robustness to synchronization errors. The key drawback of FBMC systems is the high intrinsic interference caused by the loss of complex orthogonality between subcarriers. To address the interference problem, an iterative interference cancellation (IIC)-based bit-interleaved coded modulation with iterative decoding (BICM-ID) receiver was proposed, which is compared to the baseline CP-OFDM waveform. Moreover, central to 6G enabling technologies is RIS (Reconfigurable Intelligent Services) where the channel propagation conditions are modified through reflective surfaces, opening up new opportunities going beyond typical signal processing-enhancing techniques on legacy transceivers. We developed an overview on the latest development in this area for spurring future research on this highly evolving topic. Networking resources represent a significant cost to the operator in terms of capital (CAPEX) and operational (OPEX) expenditure. Therefore, how to attain the best use of the available resources is always on the operator's agenda. This section targeted optimal radio resource management techniques that assume cooperation within distributed antenna systems using game theory, which included co-multipoint transmissions. Under the umbrella of resource management, handovers were investigated for small-cell technologies. An uplink-based received signal strength measurements approach was proposed that provided a significant overhead gain in contrast to legacy downlink reference-based handover approaches. The handover protocol was engineered according to current standards implementations (4G/5G), making it a strong candidate for legacy and future emerging B5G. Information security for B5G and small cells is also a critical issue. Confidential information will be downloaded, uploaded, and processed via the network of MSCs and relayed using foreign nodes. Cryptographic security solutions are capable of solving data privacy challenges. Moreover, it was identified that certificate chaining and PD (Partially Distributed)-TTP (Trusted Third Party)-based solutions have inherent design flaws, whereas FD (Fully Distributed)-TTP-based solutions are solvable. Therefore, new decentralized trusted third-party approaches were proposed such as the so-called "Distributed Trusted Authority-based key management (DISTANT)" scheme, which can provide a candidate security solution for small-cell networks, in particular for unsupervised networks. The 5G and beyond radio transceiver front-end vision focuses on high integration level and energy efficiency in battery-powered devices. In particular, for optimizing the amount of talk time per battery charge in low-voltage operation, research has focused on further increasing the average efficiency of power amplifiers, being identified as the most power consuming RF module. This work investigated the possibility of using MMIC (Monolithic Microwave Integrated Circuit) Power Amplifier (PA) design for 5G handsets, where MMIC technology is employed to enable compact circuit configurations and the confinement of electromagnetic fields within the semiconductor materials, very suitable for above 6GHz frequencies. Considering the 5G user transceivers, there was also a need for high gain, compact size, high operation frequency (5G high-band), and reliability. In this context, reconfigurable filtennas and phased array antenna are investigated for legacy and beyond 5G networks. All these topics were elaborated in Chaps. 2, 3, 4, 5, and 6.

Part (III) entitled "PON technology for UDNs" outlined how passive optical networks (PONs) will play a deeper role in integrated optical-wireless fronthauling. The 5G market has shaped the design requirements for mobile networking toward even higher-capacity networks to cater for the foreseen demand in traffic. This has spurred a new viewpoint in backhaul networking involving the gradual migration from the existing WDM (Wave Division Multiplexing)-PONs to ultradense WDM-PONs within urban areas. A dynamic way to achieve these aforementioned requirements is by employing hybrid photonic wireless links operating in the lightly licensed millimeter-wave (mm-wave) bands that co-exist with UDWDM-PONs, raising new challenges in terms of seamless interoperability and signal detection. This section aims to answer these challenges. On one hand, the integration of optoelectronic integrated components within the 5G base stations is elaborated. On the other hand, optical channel impairments (high free space path-loss (FSPL), chromatic dispersion, and phase noise) will affect the radio detection and link performance. This section aimed to study and analyze techniques to reduce the degradation by revisiting the signal processing in the radio-optical transceiver link. The radio modulation format selection is key to receiver performance; however, there are limited studies available for investigating the impact of legacy and emerging modulation schemes on mmWave ARoF systems. In this context, modulation candidates are compared, which suggests that legacy OFDM might not tick all the boxes as it once did; moreover, in such a converged system, the analogue radio signal will be subject to chromatic dispersion in the standard single-mode fiber (SSMF), spurring the need for equalization to compensate dispersion in the fiber. Therefore, this section studied channel equalization at the radio receiver based on a simulated mmWave ARoF scenario and provides new insights in performance. Optical amplifiers and modulators are crucial devices in mmWave ARoF systems. The REAM (reflective electroabsorption modulator)-SOA (semiconductor optical amplifier) integrated into a single chip is investigated as an alternative to directly modulated lasers (DMLs) in the optical link, where EAM-based transmitters have the potential to provide better transmission performances because of the absence of adiabatic chirp. Moreover, the SOA is sought to increase signal propagation distances to envisage the 5G coverage requirements. Therefore, device and optical link performance is investigated in terms of key parameters such as extinction ratio, insertion losses, and gain. In particular, an experimental digital transmission is demonstrated by utilizing this device, achieving a bit rate of 50-Gb/s. As the 5G milestone approaches, there needs to be a concerted effort toward practical performance evaluation, deployment strategies, and optimization in order to fully capitalize on the 5G technology in order to attain the ambitious 5G KPIs. This section aimed to provide just that; new insights on the optical-wireless link performance for converged FiWi/mmWave 5G networks are given, characterized by a wide steering angle (90°) and multi-user support. The performance is evaluated for enhanced mobile broadband (eMBB) and dense fixed wireless access (FWA) hotspot scenarios. These topics were elaborated by Chaps. 7, 8, and 9.

Part (IV) entitled "Cloud based UDNs for beyond 5G" addresses the cloud-based pillar of the book. One of the core aspects of B5G platforms is enabling network virtualization, a networking paradigm borrowed from the cloud computing world that supports highly configurable and dynamic allocation of resources and overall system flexibility. This will open up new opportunities in terms of lowering the cost of ownership for mobile network operators, as well as enhancing the QoS (Quality of Service) for end users from lower latency and reliable service provisioning, to reducing the battery consumption in handset devices. This section provided design recommendations for the planning of virtual networking infrastructures and resource allocation (RA) design. In the first instance, a new cost-efficient planning strategy was proposed referred to as shared-path shared-compute planning (SPSCP), which assigns a primary and a backup RCC (Radio Cloud Centre) node to each RAU (Radio Access Unit). To reduce the overall cost of the network, the SPSCP strategy tries to maximize sharing of the backup connectivity and computing resources. Thereafter, the focus migrated from resource backup planning to the RA problem, i.e., in other words, how we decide on the optimal resource allocation to serve the subscriber group in terms of both computational and radio resources. In this context, a notable literature review was conducted on existing approaches for RA, highlighting the existing technology gaps and open research challenges. Finally, a hybrid resource allocation design aiming at improving energy efficiency (EE) on the C-RAN is given. Although C-RAN appears to be a promising access architecture, the efficient management of the resources in C-RAN to satisfy traffic demand is still a significant challenge due to user mobility and the dynamic nature of the networking environment. In this context, this section aimed to shed some light on the specific challenges associated with C-RAN management and potential solutions based on network slicing, functional split, and their self-automation through the use of AI (artificial intelligence). A joint slice-based C-RAN solution is proposed by exploring different functional split configurations between the central and distributed units, which was shown to enhance the fronthaul (FH) network infrastructure scalability, as well as provide enhanced QoS. Resource management raises significant challenges on the computational side in terms of scaling, allocation, migration, and optimization. This section elaborated on recent developments in this area. Softwarization introduces massive flexibility into managing networks effectively, where the future envisages a "complete virtualization" of the network that includes mobile devices (vehicular units, drones, and mobile users, among others) acting as an additional pool of networking resources. The question arises as to how we can exploit UDNs and virtualization for "enhanced delivery of broadband services." In this context, this section reviewed the latest developments on this topic, which includes presenting novel concepts on energy-efficient content distribution for virtual UDNs that relies on virtualization, long-range, and local area networking capability, as well as simulation tools for validating protocols based on virtual networking instances. The explosive growth in multimedia applications and content over the last decade has placed enormous demands on bandwidth and computationally limited backhaul networks, which is likely to continue as we head toward the 6G era. The ability to manage this content in terms of caching offers multiple benefits such as network offloading, service latency, and cost reduction that results in enhanced cellular network performance. However, to find the optimal content caching strategy is a challenge within itself, which is often modelled as an optimization problem to increase QoE-related parameters (such as service latency, throughput, cache hit probability, energy, etc.) under resource (such as cache size, computation, bandwidth, etc.)-limited networks. In this section, the paradigm was further extended, by exploring not only the inherent flexibility that the underlying virtual infrastructure can provide in terms of caching services, but also how we use machine learning to self-automate and enhance the caching updating policy. Finally, task offloading harnessing on MEC (mobile edge computing) capability is proposed. In this scenario, the authors extend the notion of the edge network to include mobile devices/helper nodes in the near vicinity to provide computation services. This is an emerging 6G paradigm that is referred to as "dew computing," which is exploited here toward offloading content tasks. These topics are all elaborated by Chaps. 10, 11, 12, 13, 14, and 15.

This book edition targets to provide a concerted technology roadmap toward 6G focused on the interoperability between the wireless and optical domain, including the benefits that are introduced through virtualization and software-defined radio. We aim to be at the forefront of beyond 5G technology by reflecting the integrated works of several major European collaborative projects (H2020-ETN-SECRET, H2020-ETN-5GSTEPFWD, H2020-ETN-SPOTLIGHT). We hope this book served to provide the readership unique insights into the 6G international research effort and served as a launch pad for you to create your own footprint in the 6G arena.

Index

J. Rodriguez et al. (eds.), *Enabling 6G Mobile Networks*,
https://doi.org/10.1007/978-3-030-74648-3