Wireless Networks

Mugen Peng
Zhongyuan Zhao
Yaohua Sun

Fog Radio Access Networks (F-RAN)

Architectures, Technologies, and Applications

Springer

Wireless Networks

Series Editor

Xuemin Sherman Shen
University of Waterloo
Waterloo, ON, Canada

The purpose of Springer's new Wireless Networks book series is to establish the state of the art and set the course for future research and development in wireless communication networks. The scope of this series includes not only all aspects of wireless networks (including cellular networks, WiFi, sensor networks, and vehicular networks), but related areas such as cloud computing and big data. The series serves as a central source of references for wireless networks research and development. It aims to publish thorough and cohesive overviews on specific topics in wireless networks, as well as works that are larger in scope than survey articles and that contain more detailed background information. The series also provides coverage of advanced and timely topics worthy of monographs, contributed volumes, textbooks and handbooks.

More information about this series at http://www.springer.com/series/14180

Mugen Peng • Zhongyuan Zhao • Yaohua Sun

Fog Radio Access Networks (F-RAN)

Architectures, Technologies, and Applications

Springer

Mugen Peng
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing, China

Zhongyuan Zhao
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing, China

Yaohua Sun
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing, China

ISSN 2366-1186 ISSN 2366-1445 (electronic)
Wireless Networks
ISBN 978-3-030-50734-3 ISBN 978-3-030-50735-0 (eBook)
https://doi.org/10.1007/978-3-030-50735-0

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

Preface

In the fifth generation (5G) radio access networks and beyond, a paradigm of fog computing-based radio access network (F-RAN) has emerged to meet the requirements of the explosively increasing high-speed applications and the massive number of Internet-of-Things (IoT) devices. Inherited from both heterogeneous networks (HetNets) and cloud computing-based radio access networks (C-RANs), F-RANs take full advantages of cloud computing, fog computing, and heterogenous cooperative processing. F-RANs can coordinate the severe interference via the adaptive distributed and centralized processing techniques and provide great flexibility to satisfy quality-of-service requirements of various intelligent applications and services, such as the ultralow-latency of about 1.0 ms required by self-driving based Internet of vehicles, and up to 99.999% reliability, and even 10^6 connections/km^2 density as required by intelligent manufacturing.

Since F-RAN first combines fog computing, cloud computing, and HetNets together and has been regarded as the key evolution path to the 5G beyond system, it has drawn a large number of attention from both academia and industry. As integrating with artificial intelligence (AI), non-orthogonal multiple access (NOMA), and other advanced emerging technologies, F-RANs have entered a new researching and development era to solve the challenges that 5G beyond system meets. Actually, it is well known that F-RAN is potentially an evolutionary path to the sixth generation (6G) mobile system. This is a cutting-edge technique of multiple disciplines, including AI, wireless networks, radio signal processing, fog computing, and cloud computing. The versions and classical application scenarios defined in 5G are hardly fulfilled in 2020, while they will be fully provided in 6G. In particular, 6G can meet requirements of enhanced mobile broadband, massive machine-type communications, and ultra-reliable and low-latency communications for rich IoT services, which will penetrate various industry and business applications. In terms of these rich IoT services, F-RAN can provide a unified framework for massive access of heterogonous IoT devices, which simplifies the control and management mechanisms. With respect to AI applied in F-RANs, sufficient computation resources in fog nodes can be provided to execute machine learning and deep learning algorithms, which means that both the cost and the training efficiencies can be significantly improved.

As the F-RAN moves from the theoretical research to real world applications industry and academia are working together towards the protocols defined in standards and algorithms of all air interface layers in products, so as to enable spectral-, energy-, and cost-efficient F-RANs to be widely used as a key solution for 5G beyond and even 6G systems.

This book is firstly intended to present a comprehensive overview framework of recent advances in F-RANs, from both the academia and industry perspectives. In particular, this book covers the architecture, performance analysis, physical-layer design, resource allocation, computation offload, and field trials. The recent academic research results of F-RANs, such as the analytical results of theoretical performance limits and the optimization theory-based resource allocation algorithms, have been introduced. Meanwhile, to promote the implementation of F-RANs, the latest standardization procedure and testbed design have been discussed as well. Finally, this book will be concluded by summarizing the existing open issues and future trends of F-RANs.

We sincerely hope that this book will serve as a powerful reference for engineers and students in the majors of electronic, computations, and communications, which results in motivating a large number of students and researchers to tackle these numerous open issues and challenges highlighted in F-RANs for 5G beyond and even 6G systems.

Beijing, China Mugen Peng
Beijing, China Zhongyuan Zhao
Beijing, China Yaohua Sun

Acknowledgements

First, we would like to express our sincere gratitude to all contributors, without whom this book would have never been published. Our contributors are as follows: Mugen Peng, Zhongyuan Zhao, Yaohua Sun, Shi Yan, Chenxi Liu, Bin Cao, Xiqing Liu, Hongyu Xiang, Kecheng Zhang, Binghong Liu, Zhendong Mao, Wenbin Wu, Tian Dang, Ling Qi, Yangcheng Zhou, Yuan Ai, Xinran Zhang, Xian Zhang, Bonan Yin, and Wenyun Chen. Meanwhile, we would like to thank Springer Press staff for their continuous encouragement and support.

Second, we would like to thank our colleagues and friends all over the world who promoted fog computing into cellular networks and Internet-of-Things (IoT) for 5G and 6G developments. This book was supported in part by the State Major Science and Technology Special Project (Grant No. 2018ZX03001025-004), the National Natural Science Foundation of China under No. 61925101, 61921003, 61831002, and 61901044, the Beijing Natural Science Foundation under No. JQ18016, and the National Program for Special Support of Eminent Professionals. Without the support of these funding, F-RANs are hard to formulate and the corresponding research works could be hardly achieved within a short time. Meanwhile, I would like to thank Prof. H. Vincent Poor and Dr. Chonggang Wang because they helped open the window of C-RANs when Prof. Mugen Peng was an academic visiting fellow at Princeton University.

Third, we also thank all of those who have played a part in the preparation of this book. Our coordinator, Hemalatha Velarasu, provided useful guidelines and monitored the whole process carefully. Our Senior Editor, Mary E. James, arranged the book review and fed back the review comments in a very short time. Also, special thanks are given to Production Editors, Brian Halm and Brinda Megasyamalan, as well as Assistant Editor, Zoe Kennedy.

Most importantly, we appreciate Series Editor, Prof. Xuemin (Sherman) Shen, for his great support and invaluable suggestions.

Last but not least, we would also like to sincerely thank the support of our family members for supporting our research work when we were away from home.

Contents

1 Brief Introduction of Fog Radio Access Networks 1
1.1 History and Evolution of RANs 1
1.1.1 HetNet Overview 3
1.1.2 C-RAN Overview 5
1.1.3 5G Overview 7
1.1.4 6G Prospect 9
1.2 Fog Computing 12
1.2.1 Fog Network Structure 13
1.2.2 Fog and Edge Computing 15
1.3 F-RANs for 5G and Beyond 16
1.4 Relative Standards for F-RANs 18
1.5 Summary 18
References 18

2 System Architecture of Fog Radio Access Networks 21
2.1 System Architecture of F-RANs 21
2.2 Network Slicing in F-RANs 24
2.3 Radio Interface Techniques in F-RANs 26
2.4 Application Cases for F-RANs 27
2.4.1 F-RAN Enabled V2X 28
2.4.2 F-RAN Enabled AR/VR 30
2.4.3 F-RAN Enabled Space Communications 31
2.4.4 F-RAN Enabled UAV Communications 33
2.5 F-RAN Evolution 35
2.5.1 System Architecture for AI-driven F-RANs 35
2.5.2 Principles of AI in F-RANs 36
2.5.3 Challenges and Future Works 38
2.6 Summary 39
References 40

3 Theoretical Performance Analysis of Fog Radio Access Networks ... 41
3.1 Ergodic Capacity of User Association in F-RANs ... 41
3.1.1 System Model ... 41
3.1.2 Performance Analysis of Ergodic Capacity ... 42
3.1.3 Numerical Results ... 48
3.2 Effective Capacity with Content Caching in F-RANs ... 49
3.2.1 System Model ... 49
3.2.2 Performance Analysis of Effective Capacity ... 52
3.2.3 Simulation Results ... 57
3.3 Summary ... 59
References ... 59

4 Cooperative Signal Processing in Fog Radio Access Networks ... 61
4.1 Cooperative Non-orthogonal Multiple Access in F-RANs ... 61
4.1.1 Background ... 61
4.1.2 System Model ... 62
4.1.3 Performance Analysis ... 64
4.1.4 Performance Analysis ... 67
4.2 Contract Based Cooperative Signal Processing ... 68
4.2.1 Background ... 68
4.2.2 System Model and Cooperative Signal Processing Framework ... 69
4.2.3 Optimal Contract Design Under Complete CSIs ... 72
4.2.4 Contract Design Under Practical Channel Estimation ... 75
4.2.5 Numerical Results ... 79
4.3 Summary ... 82
References ... 83

5 Flexible Network Management in Fog Radio Access Networks ... 85
5.1 The Access Slicing Paradigm ... 85
5.1.1 Background ... 85
5.1.2 A Hierarchical Management Architecture for Access Slicing ... 86
5.1.3 Key Techniques for Access Slicing in F-RANs ... 88
5.1.4 Challenges and Open Issues ... 91
5.2 Resource Management in Sliced F-RANs ... 91
5.2.1 Background ... 91
5.2.2 System Model ... 92
5.2.3 Problem Formulation and Lyapunov Optimization ... 95
5.2.4 Solution for Orthogonal and Multiplexed Subchannel Strategies ... 98
5.2.5 Simulation Results ... 101
5.3 Summary ... 104
References ... 104

6 Dynamic Resource Allocation in Fog Radio Access Networks 105
6.1 Centralized Cost-Aware Energy Efficiency Optimization in F-RANs 105
6.1.1 Background 105
6.1.2 System Model and Problem Formulation 106
6.1.3 Problem Solution and Algorithm Design 109
6.1.4 Simulation Results 113
6.2 Cooperative Game Based Interference Management 114
6.2.1 Background 114
6.2.2 System Model 117
6.2.3 Coalitional Game Formulation 118
6.2.4 Coalition Formation Algorithm 119
6.2.5 Simulation Results and Analysis 120
6.3 Deep Reinforcement Learning-Based Resource Management 121
6.3.1 Background 121
6.3.2 System Model 122
6.3.3 Problem Formulation and Decoupling 124
6.3.4 DRL Based Mode Selection and Resource Management . 125
6.3.5 Simulation Results and Analysis 127
6.4 Summary 129
References 130

7 Content Caching in Fog Radio Access Networks 133
7.1 Hierarchical Cooperative Content Caching Framework in F-RANs 133
7.2 Content Pushing and Delivering Schemes in F-RANs 134
7.2.1 System Model 134
7.2.2 The Conventional Content Pushing and Delivering Scheme 135
7.2.3 NOMA-Based Content Pushing and Delivering Scheme . 136
7.3 The Performance of Content Pushing and Delivering in F-RANs . 137
7.3.1 The Outage Probability of NOMA-Based Scheme 137
7.3.2 Further Discussion of Theorem 7.1 138
7.3.3 Asymptotic Analysis of Outage Probability 141
7.4 Joint Optimization of Cache Management and Resource Allocation 141
7.4.1 Problem Formulation 142
7.5 Computation Complexity Analysis and Simulation Results 149
7.5.1 Computation Complexity Analysis 149
7.5.2 Simulation Results 149
7.6 Summary 151
References 152

8 Computation Offloading in Fog Radio Access Networks 153
8.1 Performance Analysis of Hierarchical Cloud-Fog Computation Offloading 153
8.1.1 System Model... 153
8.1.2 Hierarchical Cloud-Fog Computation Offloading Framework 154
8.1.3 Computation Offloading Probability and Latency Analysis 156
8.1.4 Further Discussion of Offloading Strategy 161
8.2 Joint Optimization of Computation Offloading in F-RANs 162
8.2.1 Problem Formulation 162
8.2.2 Resource Management and Offloading Decision Optimization 166
8.3 Simulation Results 173
8.3.1 Simulation Results of Latency Performance 173
8.3.2 Simulation Results of Joint Optimization Algorithm 173
8.4 Summary.. 178
References... 178

9 Prototype Design of Fog Radio Access Networks 179
9.1 Design Basics... 179
9.1.1 Enabling Fog Computing.................................. 179
9.1.2 Network Controller 181
9.2 Useful Development Tools...................................... 182
9.2.1 SDR Software... 182
9.2.2 QT 186
9.2.3 Database Software 187
9.2.4 Python... 188
9.2.5 TensorFlow... 189
9.2.6 Docker... 189
9.3 An Example of F-RAN Prototypes 191
9.3.1 Hardware Platforms 191
9.3.2 Client Software 191
9.3.3 Featured Functionalities................................... 195
9.4 Performance Evaluation of F-RAN Prototypes 197
9.5 Summary.. 200
References... 201

10 Future Trends and Open Issues in Fog Radio Access Networks 203
10.1 Future Trends of F-RANs: Federated Learning-Based Paradigms. 206
10.1.1 The Conventional Federated Learning Paradigm 207
10.1.2 A Hierarchical Federated Learning Paradigm............ 208
10.1.3 Potential Applications of Federated Learning in F-RANs 209
10.2 Fundamentals of Federated Learning in F-RANs.................. 209
10.2.1 Loss Compensation.. 210
10.2.2 Model Compression 211
10.2.3 Performance of Loss Compensation and Model Compression 212
10.3 Key Enabling Techniques of Future Evolved F-RANs 214
10.3.1 Hierarchical Cloud and Fog Computing.................. 214
10.3.2 Advanced Transmission Techniques...................... 214
10.3.3 Resource Management and User Scheduling 215
10.3.4 Intelligent NFV 215
10.4 Open Issues 216
10.4.1 Massive Multiple Access.................................. 216
10.4.2 New Theory and Techniques of Deep Learning.......... 217
10.4.3 Security and Privacy Issue of Local Model Feedback 217
10.5 Summary... 217
References.. 218

Acronyms 219

Index.. 223

Chapter 1
Brief Introduction of Fog Radio Access Networks

In Sect. 1.1, we will firstly review the evolution path of radio access networks (RANs) from heterogeneous networks (HetNets) and cloud radio access networks (C-RANs) to the fifth generation (5G) and even the sixth generation (6G) mobile communication systems, and various new emerging services will be introduced, which show the necessity of presenting fog computing. Then, in Sect. 1.2, the key features of fog computing and the relationship between fog computing and mobile edge computing (MEC) will be summarized. Finally, the fog radio access network (F-RAN) for 5G and 6G will be presented in Sect. 1.3, and the applications as well as standardization activities with respect to F-RANs will be introduced in Sect. 1.4.

1.1 History and Evolution of RANs

Mobile communication systems based on the cellular technology have a very tremendous growth in the past 50 years, and the mobile communication generation generally depends on the changes in terms of system structure, mobility speed, radio transmit and networking technology, occupied frequency, data capacity, transmit latency, etc. The first generation (1G) is a basic voice analog phone system based on frequency division multiple access (FDMA) in the frequency band of 824–894 MHz with channel bandwidth of 30 KHz. The second generation (2G) uses a digital technology and supports voice and messaging services based on time division multiple access (TDMA) and narrow-band code division multiple access (CDMA). The third generation (3G) allows packet data to transmit the bit rate up to 14 Mbps operating at the frequency band of 2100 MHz with the frequency bandwidth of 15–20 MHz, which are mainly used for the high-speed mobile internet service and multimedia video communication. The fourth generation (4G) is capable of providing 10 Mbps–1 Gbps peak rate through orthogonal frequency division multiple access (OFDMA). Based on advanced techniques, such as

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0_1

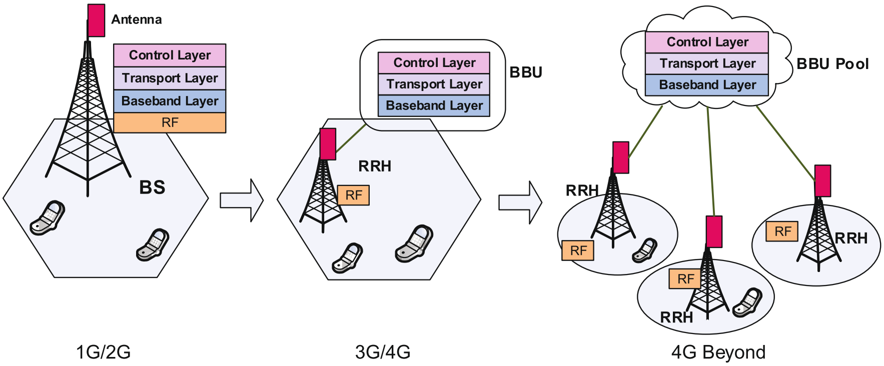


Fig. 1.1 The evolution of BS structure in cellular systems (Peng et al. 2015b)

multiple-input and multiple-output (MIMO), non-orthogonal multiple access (NOMA), and millimeter wave communications, 5G opens a new era in mobile communication technology, which can support ultra-high-speed data (peak capacity is up to 20 Gbps), ultralow latency (no more than 1 ms), massive connection (at least 1 million connected devices per square kilometer), and high mobility (up to 500 km/h move speed).

The 2G/3G cellular systems only provide a relatively low-speed (up to 100 kbps per user) data service over a large coverage area. To increase transmit data rate for packet service, the wireless local area networks (WLAN) have been presented to provide a high-speed data service (up to 11 Mbps with IEEE 802.11b standard and 54 Mbps with IEEE 802.11a standard) over a geographical small area (Peng and Wang 2009). Inspired by WLAN, the cellular network begins to evolve the functions of base station (BS), as shown in Fig. 1.1. In 1G/2G cellular systems, radio signal processing and radio resource management functionalities are integrated together inside a base station (BS), and all BSs are controlled in a centralized way. However, in 3G/4G cellular systems, the functions of the traditional BS are divided into remote radio head (RRH) and baseband unit (BBU). In this new BS architecture, the location of BBU can be far from RRH for a low site rental and the convenience of maintenance, and BBUs can be locally coordinated to suppress the inter-cell interference in a distributed way (Peng et al. 2015b).

The alternative approach to increasing the transmission bit rate is to increase the transmit power or deployment density of BSs, which creates server inter-cell interference to other serving user equipment (UEs) in adjacent cells (Hossain 2008). These traditional cellular systems are reaching their breaking points to provide high capacity due to the server interference, and the conventional cellular architectures that are devised to cater to seamless coverage and optimized for high bit rate for homogeneous traffic have been facing unprecedented challenges and problems. Different from the seamless coverage, these bursting packet traffics derived from the mobile internet mainly pursue the high data rate only in some special indoors or hot spots. As a result, there is an increasing trend to deploy small BSs in hot spots, such as relay stations (RSs), distributed antennas, and access points (APs) for picocells, femtocells, and small cells. These new small BSs are either operator-deployed or consumer-deployed, and they have taken big changes to the traditional cellular architecture, which forms a mix of small cells underlying the macrocells. To address these challenges and the corresponding changes, heterogeneous network (HetNet) comprising of marco base stations (MBSs) and small BSs has been presented as an emerging network paradigm evolution in 4G/5G systems (Vision 2013).

1.1.1 HetNet Overview

HetNet is an advanced networking technology that can cost-efficiently improve coverage and capacity. The traditional high power nodes (HPNs) are mainly used to complete the seamless coverage and provide high mobility, while low power nodes (LPNs), such as small BSs, RSs, APs for picocells, femtocells, and small cells, are mainly deployed to provide high data rates in some hot spots. By deploying additional LPNs underlying the HPN, and bringing LPNs close to UEs, HetNets can potentially improve spatial efficiency through improving the capacity of UEs in the cell edge, decrease the energy consumption due to short transmit distance, and keep good connectivity and high mobile capability (Quek et al. 2013).

There are two types of HetNets, i.e., inter-HetNets and intra-HetNets. The inter-HetNets are presented for the different heterogeneous radio access networks to interwork and complete cooperative functionalities. For example, the interworking between WLAN and 3G/4G systems is necessary for users to access the Internet flexibly. Inter-HetNet can facilitate the flexible utilizations of frequency bands across different RANs and leverage the frequency spectrum, which have been widely defined in standards and successfully used in real systems.

Through deploying a large number of LPNs with multi-tier layers in the coverage holes and hot spots, intra-HetNet is presented to increase the dense reuse of spectrum and reduce energy consumption for achieving high SE and EE in 4G and beyond systems. The strategies to reuse the frequency bands among HPNs and LPNs for intra-HetNets are generally categorized into overlay, underlay, and hybrid. As shown in Fig. 1.2a, the overlay strategy means that the frequency spectrums are orthogonally used by HPNs and LPNs, which suggests that the available spectrum band is not enough, but the cross-tier interference could be avoided under the low frequency utilization ratio. If users associated with HPNs occupy the spectral channels, the users associated with LPNs stop to transmit immediately if they also use the same frequency channels. The overlay strategy is effective in avoiding the inter-tier interference, but it requires accurate spectrum sensing and complex cognitive radio.

The underlay strategy shown in Fig. 1.2b is preferred from the operator's perspective, where HPNs and LPNs access the whole frequency band with the reuse way. Both HPNs and LPNs are assigned the same frequency bands under controlling the severe interference. With the advanced interference coordination techniques,

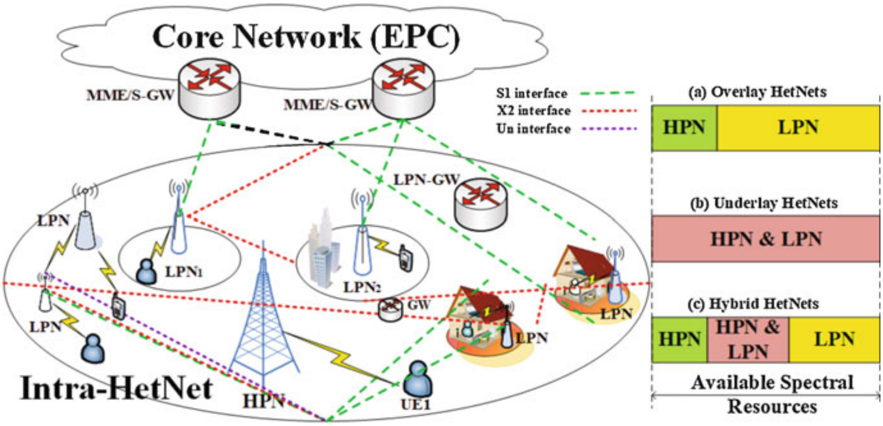


Fig. 1.2 Three frequency reuse strategies in intra-HetNets

the frequency band can be reused. In particular, the severe inter-tier interference is hardly coordinated due to the random deployment of LPNs.

As shown in Fig. 1.2c, in the hybrid overlay and underlay strategy, LPNs partially reuse the spectral resources of HPNs and thus results in the underlay structure. While the other portion of spectral bands are separately and orthogonally reserved for HPNs and LPNs, respectively. Considering the balance between performance gains and implementing complexity, the hybrid strategy is a good solution.

Since the frequency band is scarce, and the underlay HetNet strategy is the most promising to improve SE and EE, it has already attracted significant attentions and been defined in 3GPP standards. However, for the successful rollouts, it still comes with own challenges. As shown in Fig. 1.3, interference coordination and cancelation (ICC), radio resource allocation optimization (RRAO), cooperative radio resource management (CRRM), and self-organizing network (SON) are four key techniques. In terms of ICC, different types of interference should be tackled in the physical layer, including the inter-tier, inter-cell, and intra-cell interference. RRAO aims to assign the scarce physical radio blocks (PRBs) to different users for maximizing SE or EE with low complexity in aspects of the multi-dimensional resources. Since the resource assignment is strictly related to the radio channel status and the adopted ICC in physical layer, RRAO often is based on the cross-layer design mode. In addition, RRAO is responsible for scheduling the PRBs in one cell, while CRRM mainly tackles the radio resource management among multiple adjacent cells, which is often relevant to mobility management and soft frequency reuse. Finally, to reduce the configuration and optimization cost, SON is proposed to enable AI into the HetNet organization.

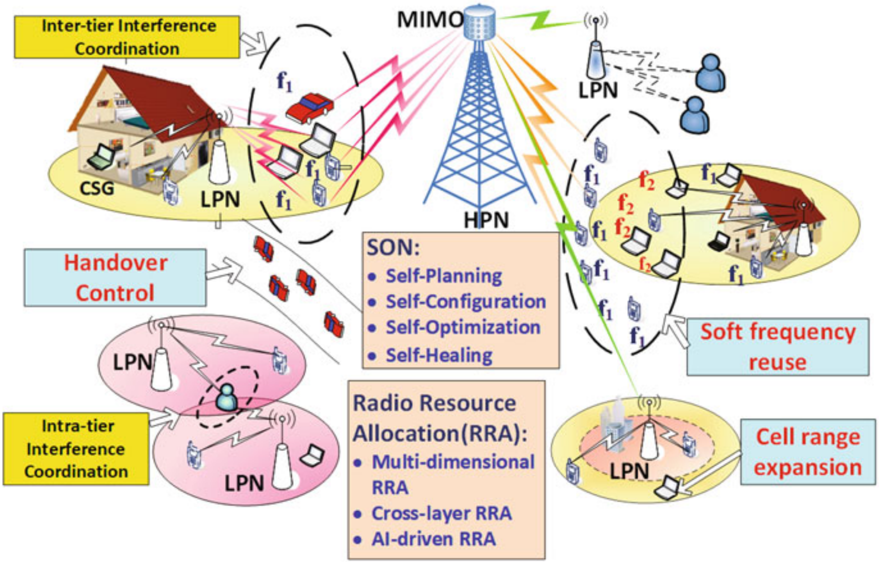


Fig. 1.3 System model and key techniques in HetNets

1.1.2 C-RAN Overview

Since most energy is consumed by BSs in the traditional cellular systems, it is appealing to migrate storage and computation into the "cloud" to create a "computing entity" so as to optimize the limited radio resource and save the energy consumption. To tackle the severe interference, C-RAN has been recognized as an advanced networking architecture, which was mainly proposed by China Mobile Company. The core idea is to move most signal processing and radio resource management functions from BSs into BBU pool. The BBU pool can take all functions of BBU together and curtail both capital and operating expenditures through cloud computing capabilities while providing high SE and EE (Peng et al. 2016b).

The history of C-RAN developments is shown in Fig. 1.4, and the concept of C-RAN is firstly proposed with the name of wireless network cloud (WNC) by IBM Company in 2010 to lower the networking cost and increase networking flexibility. However, the term C-RAN is formally proposed and fully exploited by China Mobile Company in 2011. After that, the industry started to research C-RANs. ZTE Company focused on tackling with the fiber scarcity and presented several solutions, including the colored fiber connection and optical transport network bearer (Simeone et al. 2016). To complete the large-scale processing in the BBU pool, an efficient and scalable GPP was presented by Intel. Followed by the virtualization techniques, Alcatel-Lucent Company presented the cloud BSs to decrease consuming the computation resources under guaranteeing the required

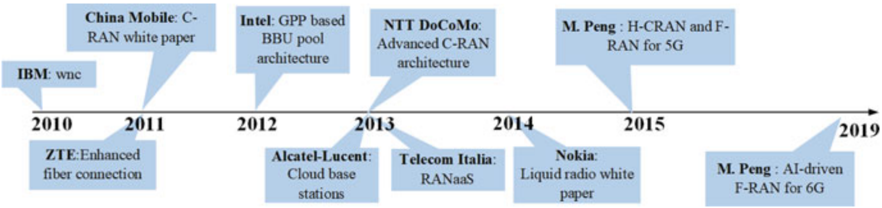

IBM: wnc
China Mobile: C-RAN white paper
Intel: GPP based BBU pool architecture
NTT DoCoMo: Advanced C-RAN architecture
M. Peng : H-CRAN and F-RAN for 5G
2010
2011
2012
2013
2014
2015
2019
ZTE:Enhanced fiber connection
Alcatel-Lucent: Cloud base stations
Telecom Italia: RANaaS
Nokia: Liquid radio white paper
M. Peng : AI-driven F-RAN for 6G

Fig. 1.4 Milestones of C-RAN development

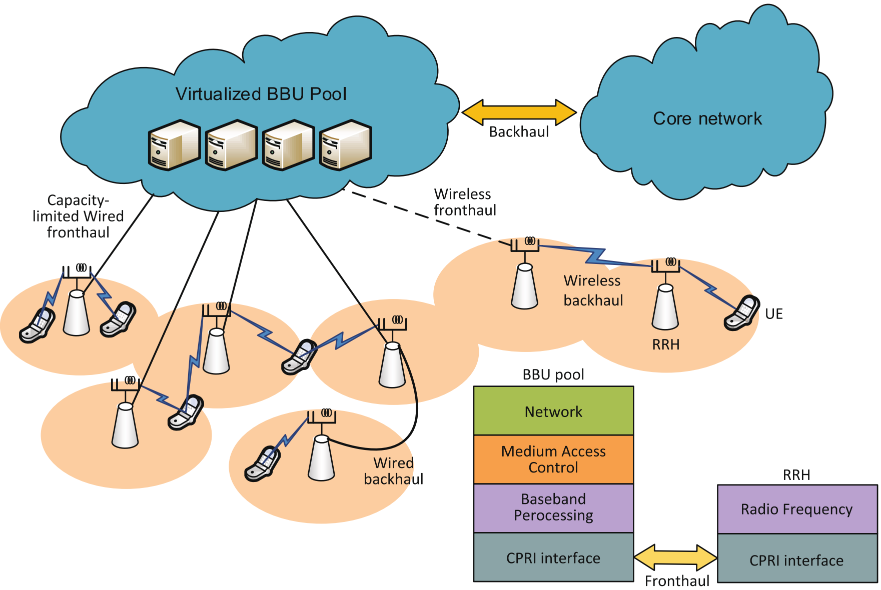


Fig. 1.5 System architecture of C-RANs (Peng et al. 2015a)

performances. In 2013, the centralized C-RAN for 4G has been discussed by NTT DoCoMo Company, and the RAN-as-a-Service (RANaaS) was emphasized by Telecom Italia Company. In 2014, the centralized C-RAN was discussed in the white paper of Liquid Radio released by Nokia Networks, and the heterogeneous C-RAN was firstly proposed by the authors of this book to tackle the challenges of C-RANs, which can promote C-RANs to evolve 6G.

As shown in Fig. 1.5, the core idea of C-RAN is to decouple functions of the traditional BSs into several RRHs and a centralized BBU pool. To support high capacity in hot spots, RRHs with radio frequency functions can be locally deployed. To suppress the severe inter-tier and inter-cell interferences, the virtualized BBU pool can take large-scale collaborative processing (LSCP), cooperative radio resource allocation (CRRA), and intelligent networking through introducing cloud computing technique. The bottleneck of C-RAN is the fronthaul that communicates with the BBU pool and RRHs. The fronthaul can be with the common public radio interface (CPRI) protocol, and it is often capacity and delay constrained (Peng et al. 2015a).

C-RANs have been advocated by both mobile operators and equipment vendors due to the potential significant benefits through introducing cloud computing into cellular networks, but they also come with their own challenges. The biggest problem is the constraints of capacity and transmit delay from fronthaul links. Meanwhile, the processing complexity and the corresponding delay in the BBU pool degrade the performance gains from LSCP and CRRA. Meanwhile, due to the non-ideal channel status information for all access links among RRHs and active UEs in the BBU pool, it is hard to make interference mitigate (Park et al. 2013). The evolution of C-RAN is necessary, but C-RAN's advantages should be kept in consideration.

1.1.3 5G Overview

Unlike 2G, 3G, and 4G, 5G is expected to fundamentally transform the role that telecommunications technology plays in society, which can enable further economic growth and pervasive digitalization of a hyper-connected society. Not only people can be connected to 5G via smartphones whenever needed, but also devices, machines, and even things can create the communicate society through 5G. As a result, 5G formulates the internet of everything and connects people, devices, machines, things, data, applications, transport systems, and cities via wireless networks, which can support a wide variety of applications and services (Chih-Lin et al. 2014).

There are three typical usage scenarios well defined in 3GPP for 5G, which includes:

- Enhanced mobile broadband (eMBB): This usage is to deal with hugely increased data rates, high user density, and high traffic capacity for hot spots scenarios as well as seamless coverage and high mobility scenarios with still improved used data rates.
- Massive machine-type communications (MMTC): This usage requires low power consumption and low data rates for the connected devices. It is mainly with IoT service.
- Ultra-reliable and low-latency communications (URLLC): This usage caters for safety-critical and mission-critical applications, such as the massive connections in the industry IoT.

As shown in Fig. 1.6, 5G is expected to provide 20 times the peak data rate, 10 times lower latency, and 3 times more spectral efficiency than 4G. 5G can transport a huge amount of data with ultra-high bit rate, reliably connect an extremely large number of devices, and process high volumes of data with minimal delay. 5G is expected to deliver significantly increased operational performance, i.e., increased spectral efficiency, higher data rates, low-latency, as well as superior user

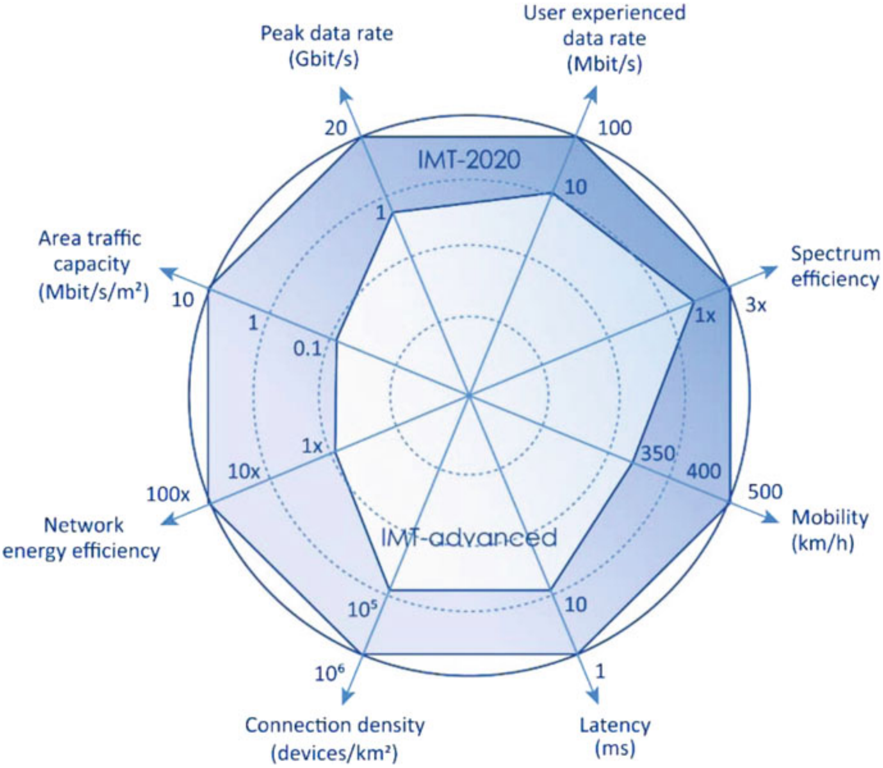


Fig. 1.6 Performance requirements comparisons between 4G and 5G (Chih-Lin et al. 2014)

experience. Meanwhile, 5G is cater not only for eMBB, but also cater for massive deployment of IoTs, which can offer acceptable levels of energy consumption, terminal cost, network deployment, configuration, and operation cost.

The increased capacity and peak data rates require more frequency spectrum and vastly more spectral efficient techniques in 5G than those in 4G. The spectrum bands allocated for 5G can be divided into three main categories: sub-1 GHz, 1–6 GHz, and above 6 GHz. Since the propagation properties of the signal benefits to create large coverage areas and deep in-building penetration, sub-1 GHz bands are often used to support traditional voice, real-time emerging services, and special services in high mobility, which can extend the coverage from urban to suburban and rural areas. The 1–6 GHz bands offer a reasonable mixture of coverage and capacity, while spectrum bands above 6 GHz provide significant capacity thanks to the very large bandwidth.

The additional spectrum mainly comes from frequency bands above 24 GHz, which poses a huge number of challenges from the intrinsic propagation characteristics of millimeter waves. To enhance transmit performance in the physical layer of 5G, new radio (NR) has been defined in 3GPP, which is mainly based on flexible multiple access and coding techniques. Until now, there are two main frequency bands that have been well defined, i.e., sub-6 GHz and the mmWave range (24–100 GHz) (Series 2015).

Unlike the traditional cellular system that requires both RAN and core network work in the same generation to be deployed, 5G is expected to integrate elements of different generations with different configurations. Standalone (SA) mode is defined for using only one RAN, while non-standalone (NSA) mode is defined for combining multiple RANs. For NSA, the 5G NR or the evolved 4G LTE radio cells and the core network can be operated alone, which suggests that the NR or evolved LTE radio cells can be used for both control and user planes. While SA is a simple solution for operators, to make user of service continuity, it is deployed as an independent network through normal inter-generation handover between 4G and 5G.

The initial phase of NSA in 5G mainly focuses on eMBB, which provides high bit rate complemented by moderate latency improvements and supports several classical use cases, such as AR/VR, UltraHD, 360-° streaming video. MMTC has been already developed as a part of NB-IoT technologies. A huge number of MMTC devices connect to the 5G BS, which makes it infeasible to allocate a priori resources to individual MMTC devices. As a result, the corresponding radio access mechanisms should be enhanced in 5G. The usage scenario of URLLC supports low-latency transmissions with small payloads and high reliability, whose bit rate is relatively low, and the main challenge is to ensure the transmit error rate should be typically lower than 10^{-5}.

To satisfy with requirements of the aforementioned three usage scenarios, especially allow them to coexist with a unified network architecture, network slicing is urgent, which jointly allocates the resources of communication, cloud computing, edge computing, fog computing, cloud storage, edge storage, and its aim is to guarantee the isolation under the required performance levels.

1.1.4 6G Prospect

Since the vision and requirements for 5G defined by International Telecommunication Union (ITU) were initially issued in 2014, the pace of 5G development was fast and smooth. In 2016, 5G standardization was formally launched in 3GPP, and then 5G technology trials were started to conduct by major official organizations of all over the world. In 2017, the first version of NSA was completely finished, and then the second version of SA standard was finalized in 2018. In 2019, several 5G pre-commercial networks have been developed widely in the world. Just like the emergence of smartphones stimulated 3G applications and triggered the demand for large-scale deployment of 4G, it is believed that some modes of IoT business will also stimulate the outbreak of 5G industry at some point in the 5G era, thereby stimulating the demand for the future 6G network (Letaief et al. 2019).

Table 1.1 Possible capabilities of 6G in comparison with 5G

Major Factors	5G	6G
Peak data rate	Up to 20 Gbps	> 100 Gbps
User experience data rate	1 Gbps	> 10 Gbps
Traffic density	10 Tbps/km^2	> 100 Tbps/km^2
Connection density	1 million/km^2	10 million/km^2
Delay	10 ms	< 1ms
Mobility	up to 350 km/h	up to 1000 km/h
Spectrum efficiency	3–5x relative to 4G	> 3x relative to 5G
Energy efficiency	1000x relative to 4G	> 10x relative to 5G
Coverage percent	Up to 70%	> 99%
Reliability	Up to 99.9%	> 99.999%
Positioning precision	1 m	Centimeter level
Receiver sensitivity	Up to − 120 dBm	< −130 dBm

With the open of the scale-up commercial deployment of 5G, more and more researchers and related organizations began to consider the evolution of 5G. At the 2018 Mobile World Congress, an official of the Federal Communications Commission looked ahead to 6G in public. Not only the USA, China also has launched 6G related work in March 2018. Consider that wireless communication systems upgrade to a new generation every 10 years, it can be foreseen that there is a certain consensus on starting 6G related research since 2020.

The goal of 6G is to meet the needs of the informatization society ten years later, so the 6G vision should focus on the needs that cannot be satisfied by 5G. The 6G vision requires massive connectivity, reliability, real-time, and throughput requirements, which are new and huge challenges to the existing 5G. As shown in Table 1.1, 6G is expected to upgrade and improve to achieve 10–100 times higher peak data rate, system capacity, spectrum efficiency, moving speed than 5G. Meanwhile, it will achieve lower delay, wider and deeper coverage, which enables to serve the interconnection of everything, fully support the development of intelligent life and industries. There are several key characteristics that have been general consent as follows (Chen et al. 2020):

- 6G is expected to be ubiquitous and integrated with broader and deeper coverage than 5G, including many kinds of communications, such as terrestrial land communication, space communication, air communication, sea and underwater communication. With AI driven intent-based networking and network slicing technologies, 6G can serve in various application scenarios, such as airspace, sky, land, and sea. In a word, 6G can realize a global ubiquitous mobile broadband communication system.
- 6G is expected to work on a higher frequency band and a wider bandwidth to achieve higher peak bit rate and average network capacity than 5G, such as mmWave communication, TeraHertz communication, visible light communication, and so on. Compared with 5G, 6G can provide a data rate up to 10–100

times, supporting the peak data rate with 1 Tbps and the user experienced data rate with 10 Gbps. In addition, 6G can achieve a flexible frequency sharing goal, which can further enhance the frequency reuse efficiency.

- 6G is expected to be a personalized intelligent and visualized network. Based on SDN, NFV, SDR, cloud computing, fog computing, and AI techniques, 6G will realize the intent-based, software-defined, flexible, and virtualized networking, which depends on the communication, computing, and communication cooperation. Meanwhile, fog computing will be promising to make cloud and edge computing adaptive to the application and networking status. As a result, the traditional centralized 5G will be evolved into the advanced phase of communication, computing, and communication cooperation. 6G should be data centralized, content centralized, user centralized, and fully service centralized.
- 6G is expected to have an endogenous security, and the function security is integrated designed. By introducing blockchain based trust and safety communication mechanisms, 6G will have the capability of self-awareness, self-analyzed, self-optimized, self-healing, and self-protect. Both the real-time dynamic analysis and the adaptive risk evaluation will be incorporated, which help realize the space cybersecurity.
- 6G is expected to merge communication, computation, sensing, and navigation functions together. To make sure the seamless coverage in sea and mountain, 6G will use satellite communication and make it cooperatively work with the terrestrial land communication. Meanwhile, the satellite navigation and positioning systems and even the radar sensing systems will be incorporated. Based on the open and software-defined networking architecture, 6G can make networking fast and self-intelligent development.
- 6G is expected to enhance its intelligence via collecting and deeply analyzing these massive configurations and running data. Meanwhile, 6G can realize everything intelligence and group collective intelligence, i.e., swarm intelligence.

As shown in Fig. 1.7, from the viewpoint of 3GPP standard organization, according to the standard scheduling, 3GPP Release 16 mainly for NR techniques will plan to be finalized in the early 2020, then research on 5G beyond systems will begin from Release 17 toward Release 19. According to the scheduling, the key technique research for 6G may be followed from Release 20 about in 2023. In the ITU standard organization, 5G standards are expected to be formally issued at the end of 2020, then the corresponding technology research on 6G may be started, in which the 6G vision and technology trend will be first considered. Meanwhile, from the viewpoint of industry, the 6G relative research has been started since 2018. The visions, performance requirements, and key technologies have been discussed by academics and industries all over the world since 2019. It is expected that these works will further undergo during 2024–2026, then the standard related works for 6G will be scheduled after 2026, and the final 6G standards will be finalized toward 2030.

The challenges of 6G include system coverage, peak capacity, average user data rate, transmit delay, movement speed, SE, and EE. 6G will develop the new air

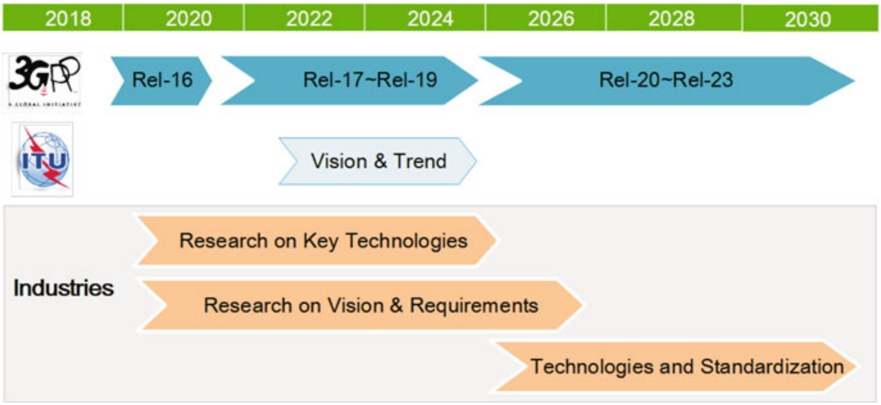

2018
2020
2022
2024
2026
2028
2030
3GPP
Rel-16
Rel-17~Rel-19
Rel-20~Rel-23
ITU
Vision & Trend
Industries
Research on Key Technologies
Research on Vision & Requirements
Technologies and Standardization

Fig. 1.7 Road map of 6G standards (Letaief et al. 2019)

interface like NR in 5G that enables multiple heterogeneous wireless transmission accesses. Meanwhile, it will make radio communication to merge computation, navigation, and sensing, all of which require cloud computing, fog computing, and AI technology together to empower 6G.

1.2 Fog Computing

Cloud computing is generally used to describe data centers available to massive users over the Internet. It can be used to save costs and help the customers focus on their core business instead of deploying massive computing storages with high IT skills. Thanks to the virtualization technique, cloud computing has a significant development since NFV and SDN has been presented. Unfortunately, cloud computing has some drawbacks, especially for the IoT services, including (Chiang and Zhang 2016):

- High transmit latency: more and more 5G and 6G applications require a low-latency, but cloud computing cannot guarantee the low transmit latency because the distance between client devices and data processing centers is relatively long;
- Low computing efficiency: Cloud computing needs an always-on connection to work properly, which suggests that a failure may reduce the reliability of the whole network. Meanwhile, interruptions are easy to occur in the cloud computing-based system and make customers suffer from a large number of unexpected outages;
- Security and privacy: The private data will be transferred to data processing centers through globally connected radio and wire channels alongside thousands of gigabytes of other users' huge data, and it is vulnerable to cyberattacks and data loss. Sensitive data should not be transmitted and processed in the cloud

server but in the edge devices because it is easy to be eavesdropped, corrupted, or even destroyed.

A possible solution to alleviate these aforementioned issues was presented by Cisco company in 2014 when fog computing is introduced as an extension of cloud computing capabilities to the network edge. Fog computing can be regarded as the extension of cloud computing, which consists of multiple edge nodes directly connected to physical devices. Fog computing and cloud computing are often interconnected, and their relationship is often confused. In nature, fog is closer to the earth than clouds. While in the technological world, fog is closer to end-users than cloud, which brings the capabilities of cloud computing down to the ground.

Fog computing is used to move the execution of computing tasks from the cloud server closer to the terminal sources through exploiting switch routers, access points, base stations, and gateways. The main difference between fog computing and cloud computing is that cloud computing is often centralized, while the fog computing is often distributed and decentralized. Fog computing has many benefits for IoTs, big data, and real-time analysis, whose main advantages mainly include:

- Low computing and communication latency: Fog computing is geographically closer to users and is able to provide more instant responses;
- Low requirements on communication bandwidth: A huge volume of data and information are stored at different edge devices, instead of sending them to the center server via the unideal channel;
- High security and privacy protection: The data in fog computing is processed by a huge number of edge devices in a complex distributed system. Thanks to the blockchain, the security and privacy protection is further enhanced in fog computing systems.
- Improved user experience: Due to the local processing and analysis, the instant responses and low time delay experience satisfy the users in fog computing systems. For example, the AR/VR users will have a significant improvement through fog computing.

When fog computing is applied into networks, fog network is formulated. Fog network will work in advanced architecture that depends on many clients and network edge devices to execute a large number of communication, sensing, configuration, control, measurement, storage, and management functions. More simply, fog network distributes computation functions closer to the edge devices generating the data.

1.2.1 Fog Network Structure

To drive industry and academic research into fog computing, including the system architecture, hardware testbed design, and a large number of inter-operability and composability deliverables, the OpenFog Consortium was founded in 2016

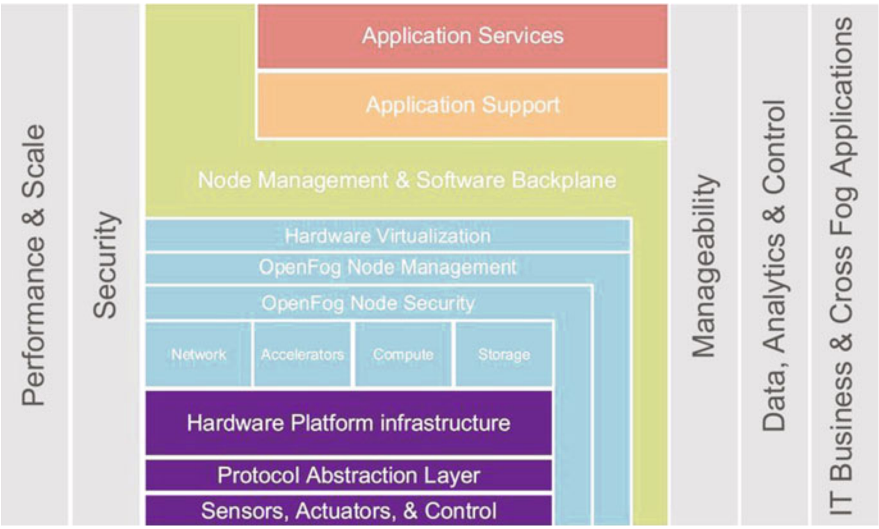

Performance & Scale
Security
Application Services
Application Support
Node Management & Software Backplane
Hardware Virtualization
OpenFog Node Management
OpenFog Node Security
Network
Accelerators
Compute
Storage
Hardware Platform infrastructure
Protocol Abstraction Layer
Sensors, Actuators, & Control
Manageability
Data, Analytics & Control
IT Business & Cross Fog Applications

Fig. 1.8 OpenFog reference architecture description with perspectives (Ope 2016)

(Ope 2016). In particular, the OpenFog Consortium issued the open fog computing architecture (OpenFog architecture) in the white paper in Feb. 2016 and announced the OpenFogReference architecture in Feb. 2017. As illustrated in Fig. 1.8, the structural aspects and perspectives of the reference architecture have been comprehensively defined, which is used as a common baseline for achieving a multi-vendor inter-operable fog computing ecosystem. Recently, the OpenFog Consortium tries to put its reference architecture into IEEE standards. The OpenFog architecture as the fog computing reference architecture has been adopted and issued as the IEEE 1934–2018 standard in Aug. 2018 (Park et al. 2013). Furthermore, the Industrial Internet Consortium (IIC) has announced to merge the OpenFog Consortium in Jan. 2019, which suggests that fog network will be discussed in the industrial IoT systems.

A typical OpenFogReference architecture has been shown in Fig. 1.9. Fog nodes can communicate with each other through wired or wireless channels and have communication, computing, networking, storing, sensing, control, and management functions. There are often three tiers in a typical fog network: cloud computing layer, fog computing layer, and terminal layer. However, more tiers or different tiers can be allowed for special applications. With the increased tiers, each tier's functions have to be sifted, and more accurate data has to be extracted and analyzed to enhance intelligence in each tier. When fog computing is used in RANs, the fog node is strictly relative to BSs and UEs, and it is often termed by new entity, such as fog AP (F-AP) and fog UE (F-UE).

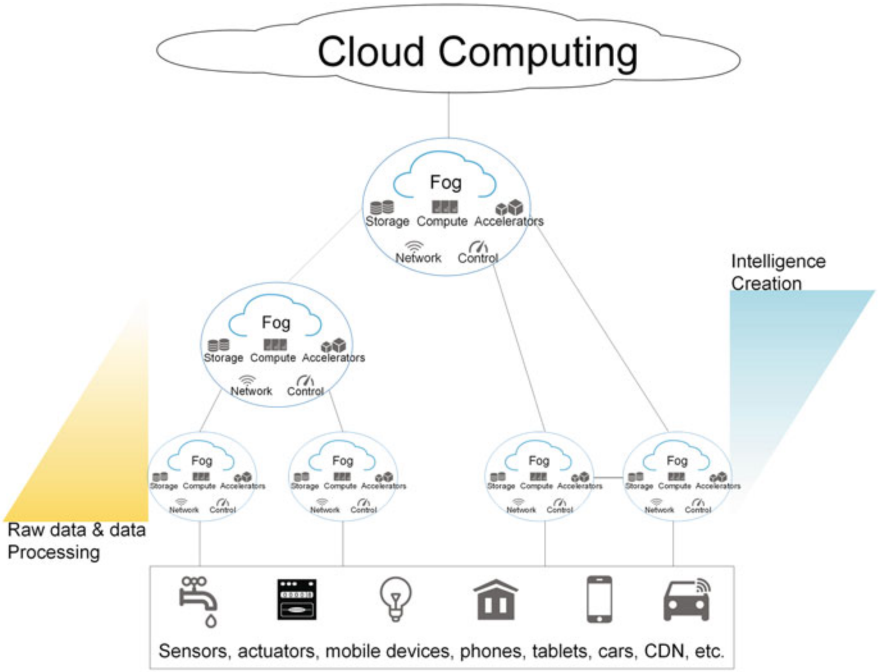

Cloud Computing
Fog
Storage Compute Accelerators
Network Control
Intelligence Creation
Raw data & data Processing
Sensors, actuators, mobile devices, phones, tablets, cars, CDN, etc.

Fig. 1.9 A typical hierarchical architecture based on fog computing

1.2.2 Fog and Edge Computing

The terms fog computing and edge computing are often used interchangeably in many publications, both of which have the similar functions. Fog computing, as defined by the OpenFog Consortium, is "a system-level horizontal architecture that distributes resources and services of computing, storage, control and networking anywhere along the continuum from cloud to things" (Chiang et al. 2017). On the other hand, the IIC defined edge computing as a "distributed computing that is performed near the edge, where the nearness is determined by the system requirements. The Edge is the boundary between the pertinent digital and physical entities, delineated by IoT devices" (Mao et al. 2017). The edge computing leverages on the processing resources, which have been already located in the network devices. However, fog computing leverages on the edge device's resources and facilitates the distribution of application logic in a cloud-to-thing continuum. As a result, fog computing is to shift typical cloud computing capabilities towards the network edge, which is totally different from edge computing.

Multi-access edge computing or mobile edge computing (MEC), as a special element in edge computing, is an advanced technique to meet ultralow-latency performance requirements as well as provide rich computing environments for intelligent applications and services closer to network edge devices in 5G (ETS

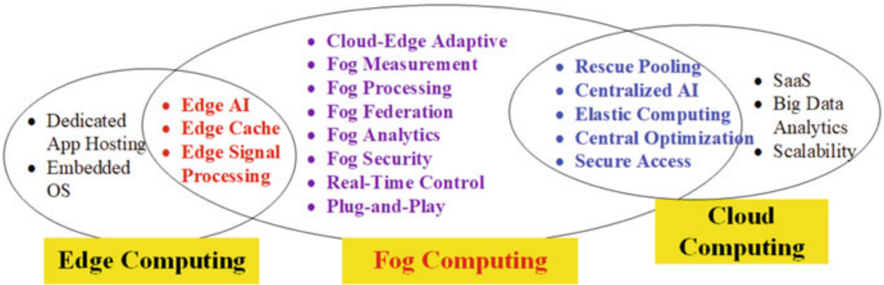


Fig. 1.10 Relationships among fog computing, edge computing, and cloud computing

2014). As shown in Fig. 1.10, fog computing is the combination of partial edge computing and cloud computing. In particular, from the viewpoint of RAN, fog computing can be regarded as a whole big picture, while edge computing and cloud computing are specific functions of fog computing. The biggest benefit of fog computing is to make edge computing and cloud computing adaptive to service, networking status, and performance requirements.

1.3 F-RANs for 5G and Beyond

Since the practical fronthaul is often capacity and time delay constrained in both C-RANs and heterogeneous C-RANs, which results in a significant decrease on SE, EE, and latency gains, the fog computing is incorporated, and fog computing-based RAN (F-RAN) has been firstly proposed to take full advantages of edge caching and AI at network edge in Peng et al. (2016a). It is noted that the proposed F-RAN will be changed into C-RAN if the edge cache, edge AI, and edge computing functions are not considered. In addition, if the cloud computing function is not considered, F-RANs will be changed in the pure MEC systems. In a word, F-RANs make cloud computing and edge computing adaptive to the network performance requirements, service, and network status. Through F-RANs, the user-centric networking is easy to achieve.

There are several apparent benefits from F-RANs, such as real-time signal processing and flexible networking at the network edge. F-RANs are adaptive to the dynamic traffic and radio environment and alleviate the capacity and transmit latency burdens on the fronthaul and BBU pool. Table 1.2 presents the comparisons between the aforementioned MEC and the proposed F-RANs. F-RANs are enhancements and evolutions of both C-RANs and HetNets, while MEC adds the computing capability in the edge devices. It is noted that MEC does not contradict with F-RANs but rather complement with them.

Table 1.2 Difference between MEC and F-RAN

	MEC	F-RAN
Motivation	Enable an open radio access network which can host third party innovative applications and content at the edge of the network	To overcome the disadvantages of the fronthaul constraints with limited capacity and long time delay
With C-RAN	Independent with C-RANs	Incorporate an enhancement and evolution of C-RANs
Key Technique	Computing offloading	Edge caching and AI
Deployment Scheme	Be compatible with 4G/5G RAN architectures	A novel system architecture evolved from HetNets and C-RANs by introducing fog computing

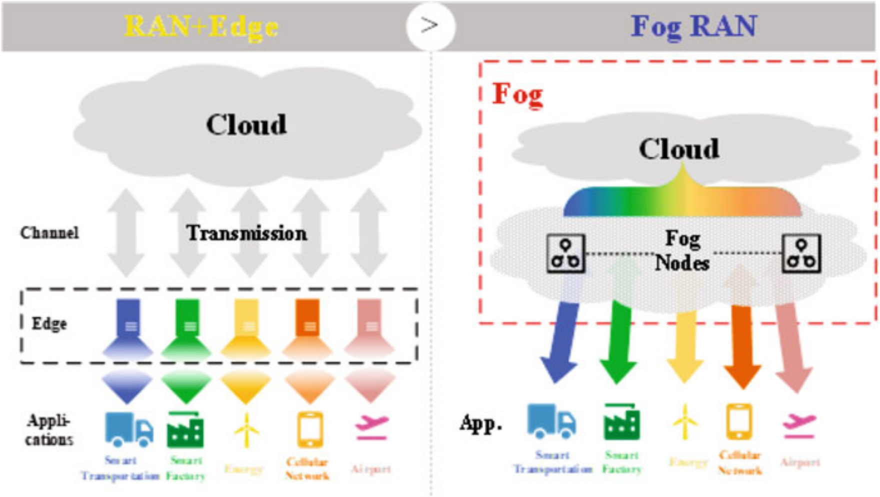


Fig. 1.11 Comparison between MEC and F-RAN

In 5G and beyond systems, the confusion among F-RAN and MEC arises from their joint goal of decentralized computing for a better end-user experience than in 4G. As shown in Fig. 1.11, through adding the edge computing layer, MEC can supply the edge cache and edge AI capabilities. While in F-RANs, the functions of both cloud computing and edge computing are incorporated, and C-RAN and MEC modes will be adaptively triggered to meet performance requirements of smart services. Furthermore, advanced edge signal processing and resource scheduling can be executed in F-APs.

1.4 Relative Standards for F-RANs

The standards of F-RANs mainly depend on the standard development of MEC. European Telecommunications Standards Institute (ETSI) launched the MEC industry specification group (ISG) to design the standards of MEC for RANs in late 2014, whose aims are to build open and standardized environments for seamless integration of applications from vendors, service providers, and third-parties. MEC ISG has expanded the initial scope to the other RANs besides 4G and 5G in 2017, hence the renaming mobile edge computing term is renamed as the new term “multi-access edge computing.”

In terms of MEC in 3GPP, to characterize extreme key performance indicator (KPI) for URLLC user scenario, two service requirements have been defined in 3GPP Release 15 TS 22.261 file (3GP 2019). Meanwhile, the efforts of MEC standards in 3GPP have been ongoing for Release 17. Moreover, 3GPP SA6 WG has worked on the technician study of the application architecture for enabling MEC in Release 17 (IEE 2018). It is noted that IEEE 1934 standard was largely developed by the OpenFog Consortium that has been introduced above (ETS 2016). IEEE 1934 is jointly developed by ARM, Cisco, Dell, Intel, Microsoft, and Princeton University, which defines the distributed resources and services framework for fog computing, in which the communication, storage, control, computing, networking, and management along with the cloud-to-things continuum have been widely addressed.

1.5 Summary

In this chapter, fog radio access network (F-RAN) as a promising candidate for the fifth generation (5G) and even the sixth generation (6G) has been well introduced, which can satisfy various performance demands by leveraging cloud computing, edge computing, and heterogenous networking. In particular, the RAN evolution path has been reviewed and the benefits brought by F-RANs have been highlighted. Meanwhile, we have elaborated the difference among fog computing, edge computing, and cloud computing and pointed out the key properties and techniques of F-RANs. Moreover, the related standardization progress related to fog computing and F-RANs has been introduced as well. This chapter has shown a technological comprehensive framework of F-RANs, which help readers to study the principles and key techniques of F-RANs in the following chapters.

References

(2014) Mobile-edge computing introductory technical white paper. ETSI, France, pp 1–36

(2016) Mobile edge computing (MEC): framework and reference architecture. ETSI, France, pp 1–18

(2016) OpenFog architecture overview white paper. https://www.openfogconsortium.org/wp-content/uploads/OpenFog-Architecture-Overview-WP-2-2016.pdf. OpenFog consortium architecture working group

(2018) IEEE standard for adoption of openfog reference architecture for fog computing. IEEE, New York. https://standards.ieee.org/standard/1934-2018.html

(2019) 3GPP SA2 study on enhancement of support for edge computing in 5GC. https://portal.3gpp.org/desktopmodules/WorkItem/WorkItemDetails.aspx?workitemId=830032

Chen S et al. (2020) Vision, requirements, and technology trend of 6G: how to tackle the challenges of system coverage, capacity, user data-rate and movement speed. IEEE Wirel Commun 27(2):218–228

Chiang M, Zhang T (2016) Fog and IoT: an overview of research opportunities. IEEE Int Things J 3(6):854–864

Chiang M et al. (2017) Clarifying fog computing and networking: 10 questions and answers. IEEE Commun Mag 55(4):18–20

Chih-Lin I et al. (2014) Toward green and soft: A 5G perspective. IEEE Commun Mag 52(2):66–73

Hossain E (2008) Heterogeneous wireless access networks. Springer, New York

Letaief KB et al. (2019) The roadmap to 6G-AI empowered wireless networks. IEEE Commun Mag 57(8):84–90

Mao Y et al. (2017) A survey on mobile edge computing: the communication perspective. IEEE Commun Surv Tutorials 19(4):2322–2358

Park S et al. (2013) Robust and efficient distributed compression for cloud radio access networks. IEEE Trans Veh Tech 62(2):692–703

Peng M, Wang W (2009) Technologies and standards for TD-SCDMA evolutions to IMT-Advanced. IEEE Commun Mag 47(1):50–58

Peng M et al. (2015a) Fronthaul-constrained cloud radio access networks: insights and challenges. IEEE Wirel Commun 22(2):152–160

Peng M et al. (2015b) Recent advances in underlay heterogeneous networks: interference control, resource allocation, and self-organization. IEEE Commun Surv Tutorials 17(2):700–729

Peng M et al. (2016a) Fog computing based radio access networks: issues and challenges. IEEE Netw 30(4):46–53

Peng M et al. (2016b) Recent advances in cloud radio access networks: system architectures, key techniques, and open issues. IEEE Commun Surv Tutorials 18(3):2282–2308

Quek T et al. (2013) Small cell networks: Deployment, PHY techniques, and resource allocation. Cambridge University Press, New York

Series M (2015) IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond. In: Recommendation ITU, pp 2083–2090

Simeone O et al. (2016) Cloud radio access network: virtualizing wireless access for dense heterogeneous systems. J Commun Netw 18(2):135–149

Vision I (2013) Framework and overall objectives of the future development of IMT for 2020 and beyond. In: Working document toward preliminary draft new recommendation ITU-R M

Chapter 2
System Architecture of Fog Radio Access Networks

In Sect. 2.1, the system architecture of F-RANs is presented, including the network elements, transmission and access modes, and the coordination mechanisms between the cloud and fog computing layers. In Sect. 2.2, networking slicing in F-RANs is introduced and key radio interface techniques are summarized in Sect. 2.3. In Sect. 2.4, several use cases of F-RANs are elaborated, followed by the discussion on future evolution path of network architecture with respect to F-RANs in Sect. 2.5.

2.1 System Architecture of F-RANs

F-RANs can be regarded as an evolution system from C-RAN and H-CRAN. In C-RAN and H-CRAN system architectures, all collaboration radio signal processing (CRSP) functions are taken out in the BBU pool, and the application contents are pre-stored centrally at the cloud application server, which makes a large number of user equipment (UEs) exchange their information quickly with the BBU pool and the cloud application server. Compared with C-RANs, the centralized control functions in H-CRANs are moved from the BBU pool to the high power node (HPN). Meanwhile, UEs with high mobility communicate with the HPN without executing handover procedure, which results in alleviating the capacity burdens on the fronthaul and decreases the unnecessary handover to keep the network stable.

The long transmit latency and the heavy capacity burdens on the fronthaul are two key challenges in both C-RANs and H-CRANs. An efficient solution is to decrease the transmitted content to the BBU pool, which suggests that a part of contents can be obtained at the local BSs, and some radio signals can be processed at the local BSs or other edge devices. In particular, to avoid all contents are achieved directly from the centralized cloud application server, some popular contents could

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0_2

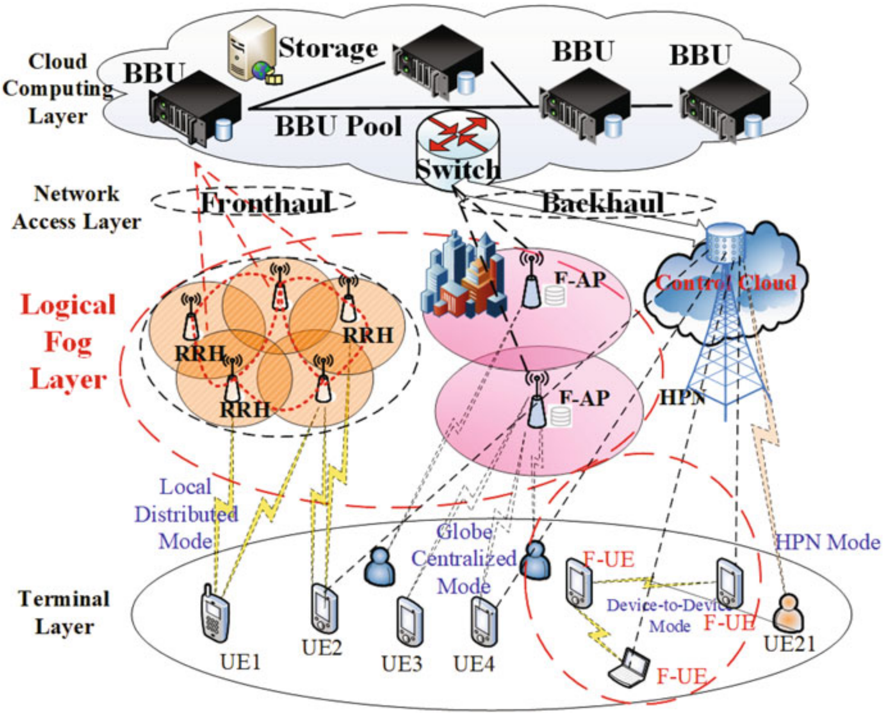


Fig. 2.1 Four layers in F-RAN

be delivered from the edge caching of adjacent F-APs or "intelligent and powerful" UEs (denoted by F-UEs) (Peng et al. 2016).

As shown in Fig. 2.1, to fulfill functions of F-RANs, the additional layer, termed by logical fog layer, is added to execute distributed communication and storage functions between the network layer and terminal layer. Global centralized communication and storage cloud layer, the centralized control cloud layer, the distributed logical communication cloud layer, and the distributed logical storage cloud layer are defined, which can make F-RANs take full advantages of fog computing, cloud computing, heterogeneous networking, edge cache, and edge AI.

The system architecture of F-RANs for 5G as shown in Fig. 2.2 can be taken as one practical example that implements four layers defined in Fig. 2.1, which comprises terminal layer, network access layer, and cloud computing layer. Different from mobile or multi-access edge computing (MEC) for 5G, the C-RAN mode working as BBU pool and RRH is included in F-RANs. The F-AP can be regarded as DU defined for 5G in 3GPP, and BBU pool can be tackled as CU. Two F-UEs can communicate directly through D2D technique. F-UEs communicate with the F-APs through wireless communication channels, while the communication between F-APs and the baseband units (BBU) pools is conducted by fronthaul links. The centralized storage and communication scheduling are in charge of the macro

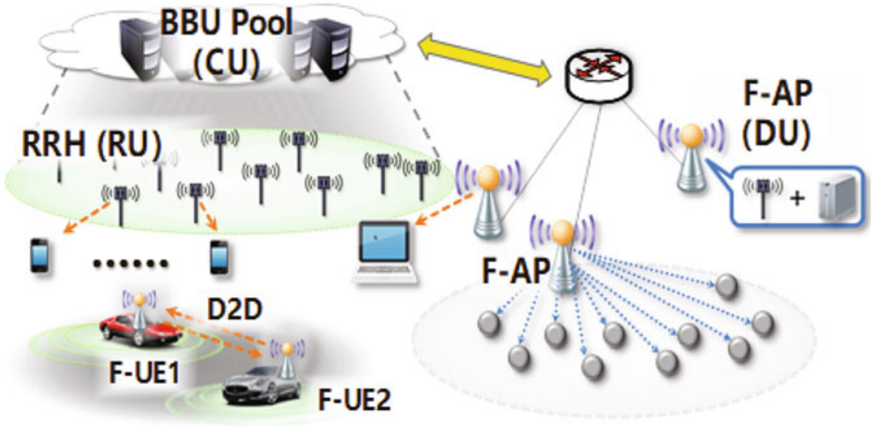


Fig. 2.2 F-RAN Architecture for 5G

remote radio head (MRRH), which is interfaced with BBU pools via backhaul links. F-APs therein are the evolutionary products of conventional radio remote heads (RRHs) in C-RANs, which nevertheless not only have the function of remote transmission but also have the capabilities of data processing and radio resource management. In addition, different from the centralized content storage scheme in C-RANs, the caching capabilities of both the F-APs and the F-UEs in F-RANs are better utilized than ever before due to the rapid upgrades of terminal technologies, which make a substantial part of popular content items ubiquitous and much nearer to the F-UEs. Specifically, mobile F-UEs can download his/her desired contents from its attached F-APs or the neighbor F-UEs located within its device-to-device (D2D) communication region instead of constructing complicated communication links with the core networks. As a consequence, the flexible services with less power consumption, low access latency, and high scalability can be provided for F-UEs. The heavy burden of both backhaul and fronthaul links is also alleviated as a byproduct (Peng and Zhang 2016).

F-APs and F-UEs as two key network elements are newly defined to enhance the functions of traditional RRHs and UEs, respectively. Different from traditional BS and small cell APs, F-AP not only acts as a BS, but also has functions of edge cache and edge AI. Through local signal processing, distributed radio resource coordination, and pre-store the popular content, F-AP can alleviate the capacity burdens on the fronthaul links, decrease the queuing and transmitting latency, and suppress the interference efficiently. F-UEs mean the special UEs with D2D and edge cache functions.

2.2 Network Slicing in F-RANs

Driven by the emerging applications of mobile Internet and Internet of Things (IoTs), 5G is expected to satisfy diverse use cases and business models. Among various services in 5G, the extreme mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable MTC (uMTC) as three typical service types are specified. The legacy 3G and 4G cellular network architectures are originally designed to improve the transmission rate, extend the coverage, and enhance the mobility for mobile broadband consumers, without considering the characteristics of mMTC and uMTC. 5G has presented network slicing as a key technique to meet the diverse use cases and business models with a cost-efficient, flexible, and soft-defined way.

Through network slicing technique, the unified network architecture is divided into multiple isolated slice instances, and each instance has appropriate networking functions with the corresponding advanced communication techniques for a specific use case or business model. Thanks to SDN and NFV, network slice instances and the isolation among them can be conveniently realized. It is noted that most existing network slicing works mainly focus on the core network (CN) aspect, while the slicing on the RAN aspect will become more and more important.

Accurately, the CN-based network slicing has several key challenges. Network slicing as an end-to-end solution has to cover the full characteristics of RANs. Especially in some use cases like uMTC demanding for an ultralow-latency, the CN-based slicing without considering the slicing of RAN is not easy to meet the performance requirements. Second, the CN-based network slice instance is mainly business and service driven, which does not take the characteristics of RANs into full account. For example, when the radio resources in RANs are not enough, the requested network slicing coming from CN may not work efficiently.

The hierarchical architecture of RAN-based network slicing for F-RANs can provide a large number of user cases and business models with the required QoS. There are two slicing layers for RAN-based network slicing to fulfill the functions of orchestration, control and data functions. As shown in Fig. 2.3 (Xiang et al. 2017), slice instance layer and centralized orchestration layer are newly defined. The slice instance layer consists of various slice instances for each user case or business model. Through slice isolation, each slice instance can operate as a septated logical RAN with specified control/data plane. For the centralized orchestration layer, the RAN-based slice orchestration is used to handle the centralized orchestration for dynamic provisioning in different slices and manage the communication and computing resources among several RAN-based slice instances.

Figure 2.4 shows the hierarchical architecture of RAN-based network slicing. In the centralized orchestration layer, the multi-dimensional information acquired from edge devices, terminals, and sensors are collected and analyzed and perform the centralized multi-dimensional resources management for all slice instances. Thanks to the NOMA technique, the radio resource for different slice instance can be orthogonal or shared. The slice instance layer is responsible for the slice-

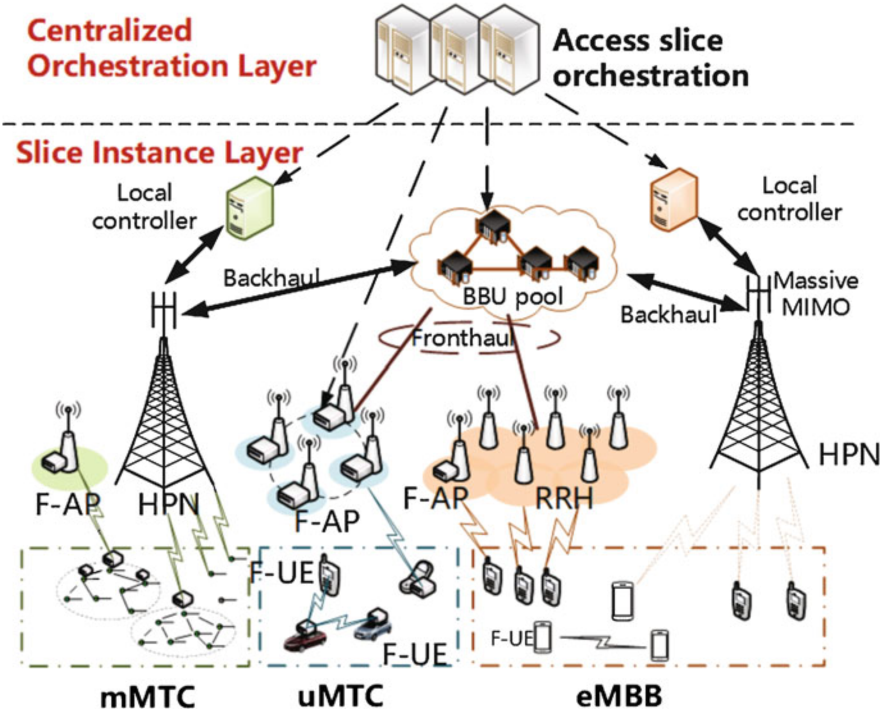


Fig. 2.3 Network slicing in F-RANs

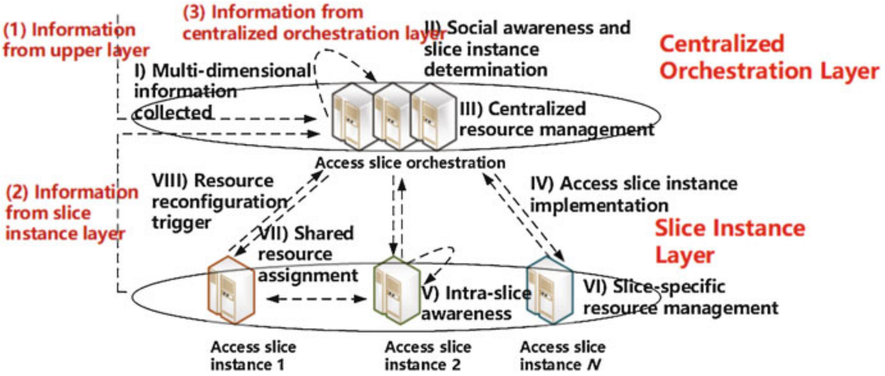

(1) Information from upper layer
(3) Information from centralized orchestration layer
II) Social awareness and slice instance determination
Centralized Orchestration Layer
I) Multi-dimensional information collected
III) Centralized resource management
Access slice orchestration
(2) Information from slice instance layer
VIII) Resource reconfiguration trigger
VII) Shared resource assignment
IV) Access slice instance implementation
Slice Instance Layer
V) Intra-slice awareness
VI) Slice-specific resource management
Access slice instance 1
Access slice instance 2
Access slice instance N

Fig. 2.4 Hierarchical Layer for F-RANs

specific configuration, analysis, operation, and management. When the pre-assigned radio or computing resources in a certain instance cannot guarantee the performance requirement, some share resources among different RAN-based slice instance can be rescheduled.

2.3 Radio Interface Techniques in F-RANs

Non-orthogonal multiple access (NOMA) is an effective approach to improving SE in 5G system (Gu et al. 2018). Compared with the conventional orthogonal multiple access (OMA) techniques, such as FDMA, TDMA, CDMA, and even OFDMA, NOMA can transmit several data streams with a same radio resource block simultaneously, which can achieve significant performance gains in terms of capacity, latency, and connectivity (Ding et al. 2018).

The key idea behind NOMA is to ensure that multiple users are served simultaneously within the same given time/frequency resource block (RB), utilizing superposition coding (SC) techniques at the transmitter and SIC at the receiver. To further improve the system capacity, decrease the queuing delay, and increase the connection number, the NOMA is anticipated to be used as the key technique in the physical layer of F-RANs. In NOMA-enabled F-RANs, the contents downloading and the tasks offloading between F-UEs and F-APs can be conducted, which fulfills the stringent performance and delay requirements of content transmission and task computing with finite battery and computing capacity (Zhao et al. 2017).

Figure 2.5 illustrates an example of system model for the NOMA-enabled F-RANs. The power domain NOMA with SIC has received numerous attention due to its capability of providing enhanced SE and EE (Qi et al. 2019). The main principle of NOMA is simultaneously accommodating multiple F-UEs with diverse power levels in a single RB via SC, which makes the access of massive F-UEs possible (Yan et al. 2020).

In traditional C-RANs, UEs can only access the RRH from the cloud application server to obtain the desired content, which suggests that only the global C-RAN mode takes effects. However, due to the existence of edge caching in the neighbor UEs for F-RANs, the desired UE can communicate with the neighbor F-UEs through the D2D communication. Otherwise, it can be served by accessing F-APs or RRHs according to the radio channel status and cache conditions around (Sun et al. 2019b). As well known, there are mainly four classical transmission modes in F-RANs: D2D, F-AP, C-RAN, and HPN. The D2D mode will be enabled and triggered when the communication distance between these two adjacent F-UEs is sufficiently short and the transmitting F-UE has cached the content that the receiving F-UE needs (Dang and Peng 2019). F-AP mode will be enabled and triggered when the D2D mode fails while the neighbor F-AP can. Moreover, if F-AP cannot provide the desired UE's required to transmit capacity, however, there are more than one RRH, which can provide UEs with the satisfied capacity, the desired UE will use the C-RAN mode. Otherwise, the traditional HPN mode will be used for the desired UE (Yan et al. 2017).

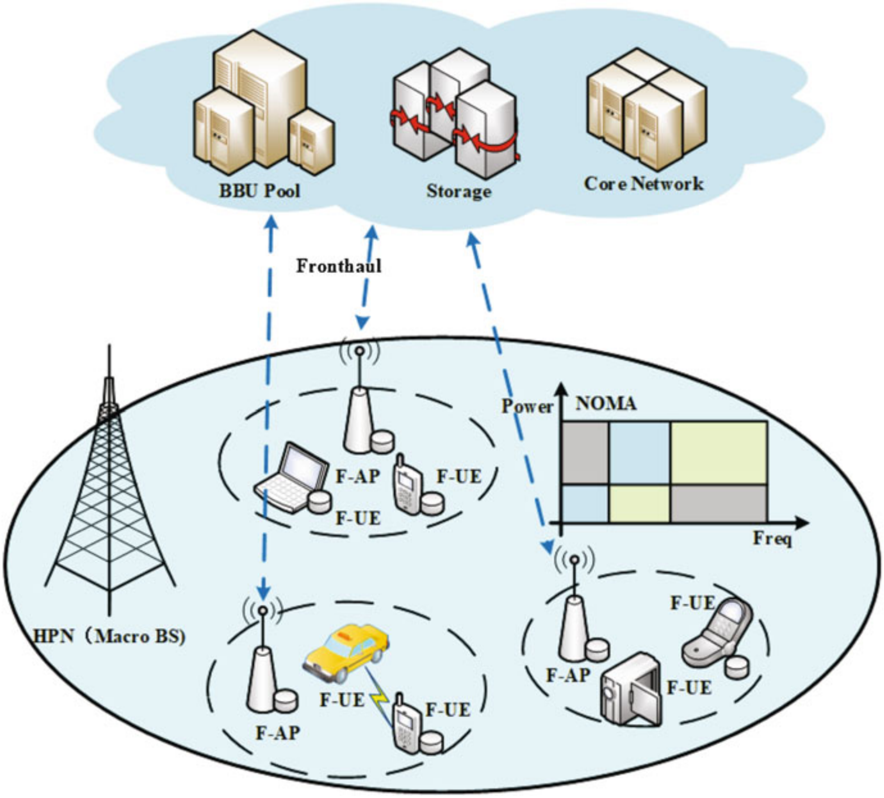


Fig. 2.5 Principle of NOMA-enabled F-RANs

With adaptive model selection, F-RANs can bring higher SE and lower latency compared to using C-RAN mode only, and meanwhile F-RANs are also capable of improving EE. Note that the global C-RAN mode can suppress all inter-RRH interference in the ideal status, however, it is often constrained by the capacity-limited fronthaul and high complexity due to LSCP, which results in not good SE and EE in practical. As a result, in the real F-RANs, the association mode is key and should be carefully designed, in which many key factors should be jointly optimized, including the desired content in the edge cache, the desired SINR, and the energy consumption.

2.4 Application Cases for F-RANs

As advanced technical solutions of 5G, F-RANs will have promising applications in autonomous vehicles, smart city infrastructure, traffic management, industrial automation, augmented reality (AR)/ virtual reality (VR), drones, etc. In this section, vehicular-to-everything (V2X), AR/VR, space communications, and unmanned aerial vehicle (UAV) as four typical use cases will be briefly introduced.

2.4.1 F-RAN Enabled V2X

A large number of cities are or will deploy intelligent transportation systems (ITS), which supports the connected vehicle technology. As one of the most anticipated smart city services, ITS is a promising technology to enhance driving safety and efficiency, provided that vehicles as well as transportation infrastructure have the abilities of sensing, connectivity, processing, and autonomy. V2X, also known as C2X (car-to-everything), is promoted as a solution to traffic safety. The formation of V2X, evolved from typical vehicular communication forms, vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I), to more generalized vehicles-to-any entities, including pedestrians (vehicle-to-pedestrians, V2P) and cloud (vehicle-to-cloud, V2C). Under the support of flexible communication scenarios, V2X can perform a relatively efficient and real-time data exchange. However, the capability of transmitting a huge data collected from massive sensors to a remote cloud will lead to serious delay problems in the ITS, which is not acceptable as the delayed decisions might cause a long transmit delay, traffic congestion, and even serious accidents (Chen et al. 2017).

The traditional communication standards of V2X include DSRC technology based on IEEE 802.11p and cellular-V2X technology evolved from D2D communication, both of which have their merits and demerits. DSRC has a stable, low-latency beacon support for safety-related communication, but its low scalability is reflected in the case of unable to provide the required time sensitive characteristics in dense road scenarios. Compared with DSRC, cellular-V2X has more available radio frequency (RF) resources and better support of infrastructure, but it seems difficult to meet the strict delay requirement of safety-related messages. Higher synchronization requirements and excessive occupation of conventional cellular communication resources are also its potential weakness.

With the increase of both communications nodes and advanced service demands, both DSRC and cellular-V2X seem unable to provide a stable communication network with ultralow-latency and large capacity for future emerging V2X applications. Therefore, the introduction of new technologies seems to be an irresistible tendency and a possible solution for V2X performance upgrade, which provides a promising land of ultralow-latency, high data rate, and high-density networks (Zhang et al. 2019). Meanwhile, the extensive and comprehensive information collection in V2X is also achieved by the sharing of messages through single wireless link between vehicles and other nodes including infrastructures, which becomes the main assumption of information collection since it is a cost-efficient way. However, high-frequency data exchange could lead to high-occupancy of RF resources.

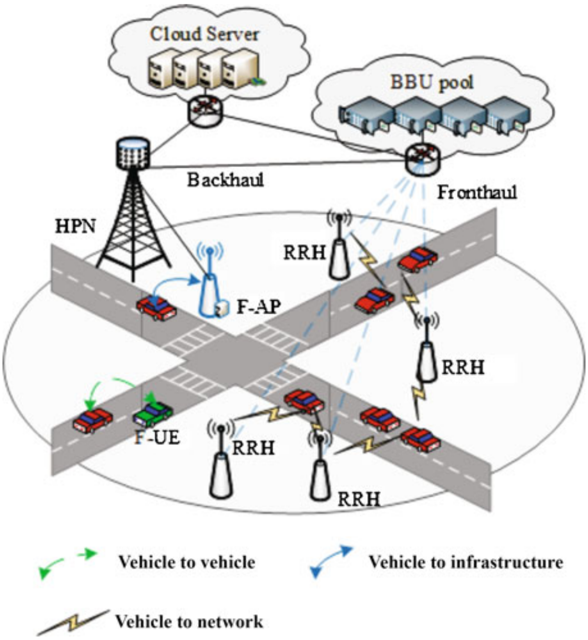


Fig. 2.6 Application of F-RANs into cellular-V2X

To efficiently support cellular V2X, F-RAN is promising. If roadside units or BSs have prefixed locations and strong computing and storage abilities compared with vehicles, which have a direct way of environment sensing instead of relying on wireless link status information through the dissemination of vehicles, the QoS for ITS can be surely improved. As shown in Fig. 2.6, the aforementioned side units or base stations, uniformly denoted by F-AP in F-RANs, can meet the strict performance requirements of V2V, V2I, V2P, and V2C.

There are three levels for the F-RAN-based cellular-V2X: cloud level, fog level, and sensor level. The cloud level provides city-level monitoring and centralized control from a remote location and makes global decisions (such as global traffic management and large-scale traffic light control). The fog level comprises widely deployed F-APs, which are mainly used to perform collaborative radio signal processing, resource management, and local scheduling. As generally deployed across different areas, F-APs collect a large number of data that are sent or produced by vehicles, then process and analyze these measurement data, and report the processed data to the cloud server. Meanwhile, F-APs can act as intermediate devices that connect the cloud level and vehicles. The sensor level residing at vehicles is responsible for processing the massive data produced by vehicles and downloading essential information, such as video information and accidents to road users. Note that the sensor is responsible for processing the massive data produced by vehicles and uploading essential information, such as the planned route and accidents (Liao et al. 2019).

2.4.2 F-RAN Enabled AR/VR

The rapid development of wireless technology has enabled the emergence of VR/AR, which is anticipated to reach 254.4 Petabytes per month by 2022. This huge amount of VR/AR traffic poses many challenges on the 5G and beyond wireless networks, such as huge capacity and very low-latency. Against this backdrop, it is challenging to deliver immersing VR experience for the users. Motivated by the benefits of F-RANs, the AR/VR delivery over F-RANs is promising (Bastug et al. 2017).

As illustrated in Fig. 2.7, the F-RAN-based mobile VR delivery system consists of one cloud server, one HPN, *M* F-APs, *N* VR-UEs, and multiple RRHs. In this system model, in order to reduce the transmission latency, the cloud server extracts 360° VR videos into monocular videos (MVs) and stereoscopic videos (SVs). The F-APs and VR-UEs are equipped with caching and computation capabilities and thereby can judiciously cache some MVs and SVs. The VR delivery procedure works as follows: (1) Tracking: the VR-UEs first track the viewpoint and request it through the uplink; (2) Extraction: the VR-UEs/F-APs extract the MV of the viewpoint if it is cached, otherwise requesting the MV from the cloud server. (3) Projection: the VR-UEs/F-APs project the MV into SV, and the VR video is transmitted to VR-UEs through the downlink; (4) Rendering: the VR-UEs render the SV to 360° VR videos (Dang and Peng 2019).

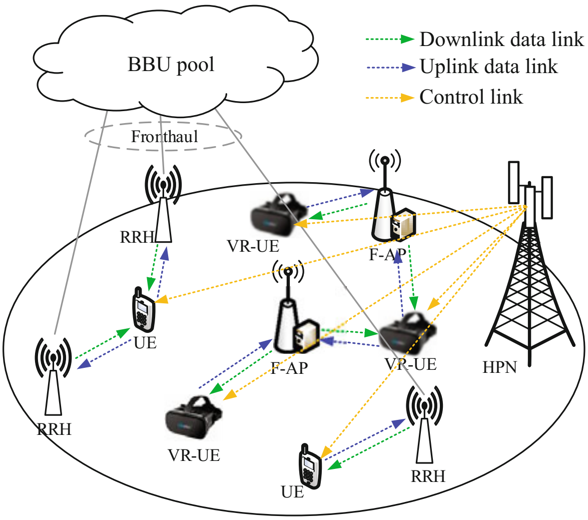


Fig. 2.7 The VR delivery model in F-RANs

2.4.3 F-RAN Enabled Space Communications

The overall architecture of F-RANs applied into the space communication is illustrated in Fig. 2.8. There are two kinds of F-RANs according to locations: space and air F-RAN (SAF-RAN, including SF-RAN and AF-RAN) and terrestrial F-RAN (TF-RAN). As a result, there are two layers: the space or air layer and the terrestrial layer. The UEs from the remote and dense areas are usually connected to the space or air fog access point (SAF-AP) in SAF-RANs and the terrestrial for access point (TF-AP) in TF-RANs, respectively, while users moving from place to place may perform handover from space to terrestrial connections and vice versa (Guidotti et al. 2019).

In the space or air layer, any satellite or air platform station can act as a SAF-AP and then performs dynamic resource management to allocate resources among tasks. The accessed SAF-AP may distribute its storage or computation tasks among its neighboring SAF-APs for cooperative processing. The transmitted data can be forwarded via inter-satellite links and satellite feeder links to the cloud server in the ground or the satellite-based core network when necessary.

In the terrestrial layer, the access points or nodes in TF-RANs are divided into two categories: RRHs with the C-RAN mode and TF-AP. UE can be served by more than one RRHs at the same time, whereas it is allowed to be served by only one TF-AP or SAF-AP where there is no coverage of TF-RANs. Also, different from RRHs, both TF-AP and SAF-AP can possess agent-based capabilities, i.e., edge AI enabled

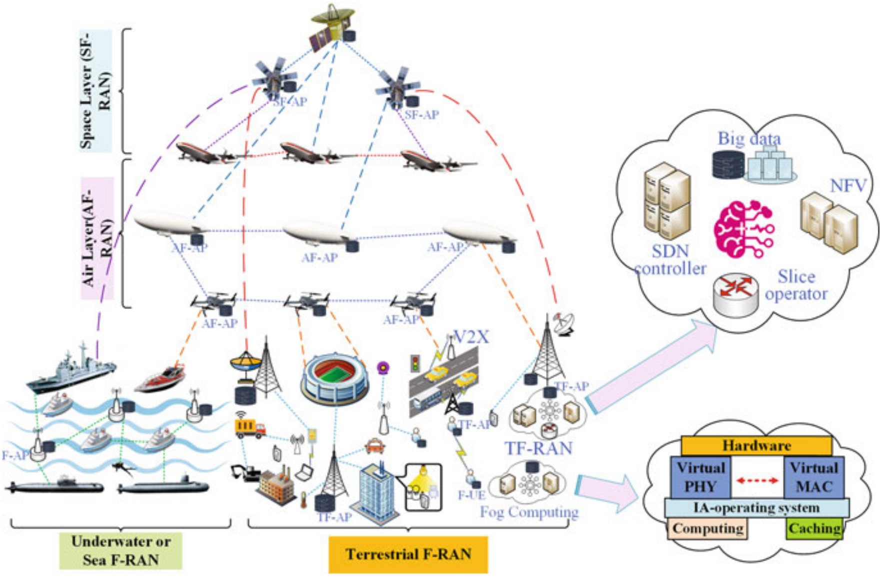


Fig. 2.8 Application of F-RANs into pace communications

local processing, cooperative resource management, and distributed storing. These consistencies among space, air, and terrestrial F-APs make it possible for UEs to seamlessly switch from one to the other.

The TF-AP connects to the cloud server via wired link and exchange data with each other when necessary. The cloud server can be abstracted into a few core functionalities such as network controller, service manager, operation manager, safety manager, and cloud gateway. Further, as the cloud server should handle connections with both SAF-APs and TF-APs, network slicing is an attractive technique to effectively manage different F-RANs and perform resource allocation among them.

Note that the SAF-AP in the space level often depends on the low earth orbit (LEO) satellite, and the SAF-AP in the air level often depends on the low-altitude UAV. Geosynchronous orbit (GEO) satellite is stationary relative to the earth and serves only a fixed coverage, and the steerable beams are restricted in both dynamic ranges and performance. First of all, it is not capable of serving all the remote areas on the earth once launched. Additionally, it has the best performance around the sub-satellite point whereas it does not guarantee the same when the users are located near the edge of its coverage. Second, the high round-trip delay of the GEO satellite, e.g., about 500 ms, makes it deployed at this orbit difficult to preserve the attractive characteristics as in TF-RANs. Third, to achieve the expected service performance, GEO satellites have suffered heavy burden incurred by overcoming the negative impact of far distance and complicated space environment, which makes GEO satellites not be capable of further implementing functions of SF-AP.

Non-geostationary spacecraft such as LEO satellite provides an alternative for SAF-RAN to its GEO counterparts. Unlike the fixed orbit, LEO satellite coverage changes as it revolves around the earth. It can cover the whole earth surface within certain period so that any remote area would be served. In addition, LEO satellites are cost-effective compared to GEOs. Meanwhile, there is no need to wait until the whole constellation is deployed. This is consistent with the general phased LEO constellation deployment. The lower round-trip delay of a LEO satellite is around 20 ms, which makes it a suitable candidate for RAN in the ground. LEO satellites are designed to satisfy various requirements, which can not only handle physical layer tasks, but also be capable with switch and routing processing in upper layers. Application scenarios such as remote sensing in remote areas require the satellites to carry out storage, computation, and pattern recognition tasks before forwarding the reduced dataset back to the cloud server through the satellite gateway, which suggests that the SF-AP based LEO satellite is promising in space communications (Di et al. 2019).

Although promising and attractive, SF-AP based LEO satellites face challenges especially when AI techniques are employed. First, LEO satellites periodically cover all areas of the earth surface. Thus, different from the machine learning-based GEO resource allocation, which only takes into account the user distribution characteristics in a restricted area, the LEO resource allocation should consider the distribution of UEs over all global surface for its training. Though reliable traffic models have been developed with factors as terminal types, busy hours, and the gross domestic product per inhabitant of the population covered by the beam, it is still necessary to design the learning convergence and its effectiveness for the online performance. Second, the dynamic nature of the LEO-user channel, e.g., free space loss and rain attenuation, further decreases channel gains.

2.4.4 F-RAN Enabled UAV Communications

UAV communications have attracted a large number of attentions because they have several advantages, including flexibility in grouping, no communication environment restriction, large coverage, high mobility, and low cost. Despite there are many promising benefits, UAV communications are also faced with several challenges, such as the low bit rate due to bad radio channel, randomly UAV networking topologies, severe interferences, limited wireless backhaul links. In order to overcome these limitations and maximize the advantages of UAV communications, F-RAN can be included, where fog computing by leveraging the use of UAVs in the air level is appealing and expected to be enormous in 5G and beyond.

Due to the ability of on-demand and swift deployment, the air-to-ground fog computing provided by F-AP enabled UAV will find many promising uses in F-RANs. F-AP enabled UAV enjoys the LoS-dominated UAV-ground channel and the UAV's controllable high mobility in 3D spaces that can be exploited for the enhancement of cooperative task executions in F-RANs. The probabilities of LoS channels between the UAV and the ground F-APs are in general very big due to the high altitude of UAVs, and thus UAV-ground communications are significantly less affected as compared to the traditional cellular communications. Meanwhile, the swift 3D deployment and even dynamic movement of UAVs becomes feasible so that they can adjust their locations and trajectories to maintain favorable LoS channels (Yan et al. 2019).

Figure 2.9 shows the deployment of a cluster of rotary-wing F-AP enabled UAVs that provide services for the coverage area. These F-AP enabled UAVs can achieve on-demand computation service provision for base stations. In addition to extending the coverage, the F-AP enabled UAVs can work as aerial helper to provide air-to-ground fog computing services for terrestrial devices in the emergency situations, e.g., earthquake, typhoon, where the terrestrial devices having no reach of central cloud nor terrestrial F-AP. The F-AP enabled UAV is also suitable to provide on-demand communication and computation services to low-power IoT devices such as sensors and tags in IoT systems, e.g., intelligent farming.

When F-RANs are applied into UAVs, the resource allocation is much complicated compared to the case of TF-RANs, as UAV-ground links change dramatically due to the high mobility of the UAV. On the timely changing of the UAV trajectory, the resource allocation should be frequently re-calculated, which will cause enormous computing cost. As a result, the resource allocation methods will cause intensive computation cost making it impractical. In contrast, DRL is a good choice to solve such resource allocation issue.

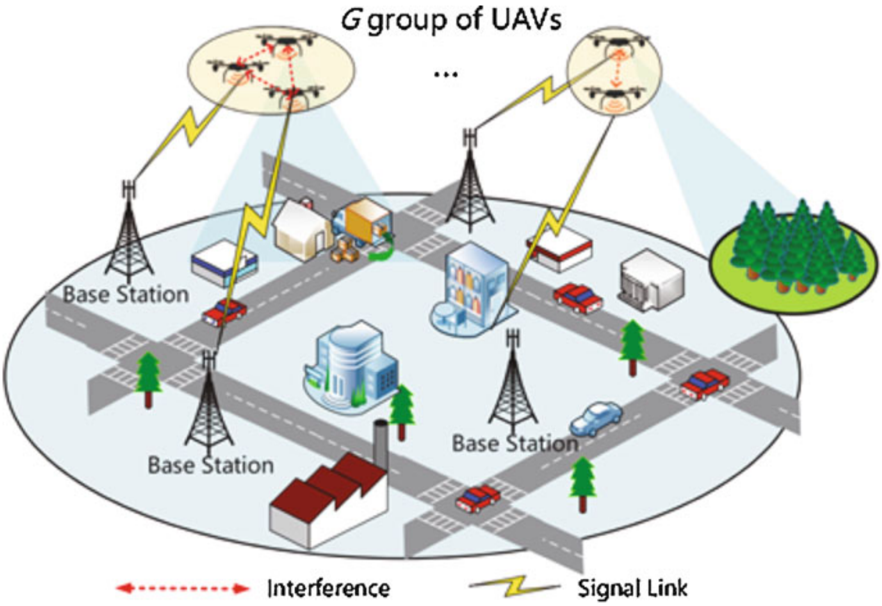


Fig. 2.9 Application of F-RANs into UAV communications

In practice, each F-AP enabled UAV has to make good use of the scarce radio frequency resource, i.e., communication resource, and its computation capacities to realize the resource-efficient fog computing. Speaking of the radio frequency resource, UAV-based F-RAN supported by 4G/5G, IEEE-802.11 or IEEE-802.15 technique, takes each OFDMA based sub-channel as the atom resource block and rationally distributes these fixed number of sub-channels to each UAV-GT link for satisfying data rate. In addition, the allocation of communication resource also needs to control the power of each sub-channel. Basically, the downlink power of one F-AP enabled UAV is limited, i.e., the overall power allocated to all the sub-channels from the F-AP enabled UAV is fixed at a time. Therefore, F-AP enabled UAV should adopt power control strategies to distribute the limited downlink power to sub-channels to support satisfying data rate of UAV-GT links. The power control should also consider eliminating the co-channel interferences to benefit the UAV-GT links. For example, if there is UAV-GT link having sub-channels interfering other UAV-GT links, the power of the interfering sub-channels should be decreased, vice versa (Li et al. 2019).

For the computation resource, F-AP enabled UAV has to allocate its limited computation resources, i.e., capacities, to each UAV-GT pair for the task execution. Similar to the communication resources, F-AP enabled UAV takes its computation resource as numbers of resource blocks and allocates them to UAV-GT pairs in a designed strategy. Overall, the allocations of the communication and computation resources jointly determine the latency of each GT task, which is composed of the time spent on data transmission in UAV-GT link and the time spent on executing the task by the F-AP enabled UAV.

2.5 F-RAN Evolution

With the emergence of data-hungry multimedia applications in F-RANs such as AR/VR, the resulting data is increasingly huge with more complex features. In addition, the progressively complicated architecture of modern wireless networks is becoming more complicated as well as the coexistence of various forms of resources (such as spectrum, time slots, antennas, transmit power, computing power, etc.) poses difficulties in network resource management. To better support high-speed data applications and massive connections of IoT devices, artificial intelligence (AI) approaches start to emerge in F-RANs as a promising candidate to meet the increasingly strict requirements of network services in terms of ultra-low-latency, high reliability and density, and intelligent decision.

2.5.1 System Architecture for AI-driven F-RANs

It can be envisioned that with the increasing diversity of the F-RAN applications, the multimedia data becomes huge, heterogeneous, and high-dimensional. Thus, transmitting raw data to the fog node/cloud directly causes high communication overhead and utilizing raw data for network optimization directly may lead to low efficiency and high computing overhead issues. Unfortunately, the traditional F-RANs without considering AI techniques face the following four severe challenges: lack of unified and collaborative management of fog resources, lack of deep perception of network, services, users, and contents, lack of hierarchical cooperation mechanism for the cloud-fog resource allocation problem, and lack of machine learning algorithms suitable for F-RANs. As shown in Fig. 2.10, the concerned application and system architecture for the AI-driven F-RANs can potentially lead to efficient, rapid, trustworthy management operations (An et al. 2019).

To promote AI applied in F-RANs, the first step to design the architecture of AI-driven F-RANs is to realize unified management of multi-dimensional heterogeneous fog resources (e.g., computing, storage, and network resources) by using virtualization technologies. Then, based on the acquired perceptual information (dynamic demands of network services, usage status of fog resources, user behaviors, content popularity), AI assisted F-RAN efficiently allocates cloud-fog resources according to differentiated service requirements. Finally, by designing appropriate big data algorithms, this aforementioned system architecture can achieve intelligent allocation and optimal deployment of fog resources while meeting the dynamic and differentiated service requirements.

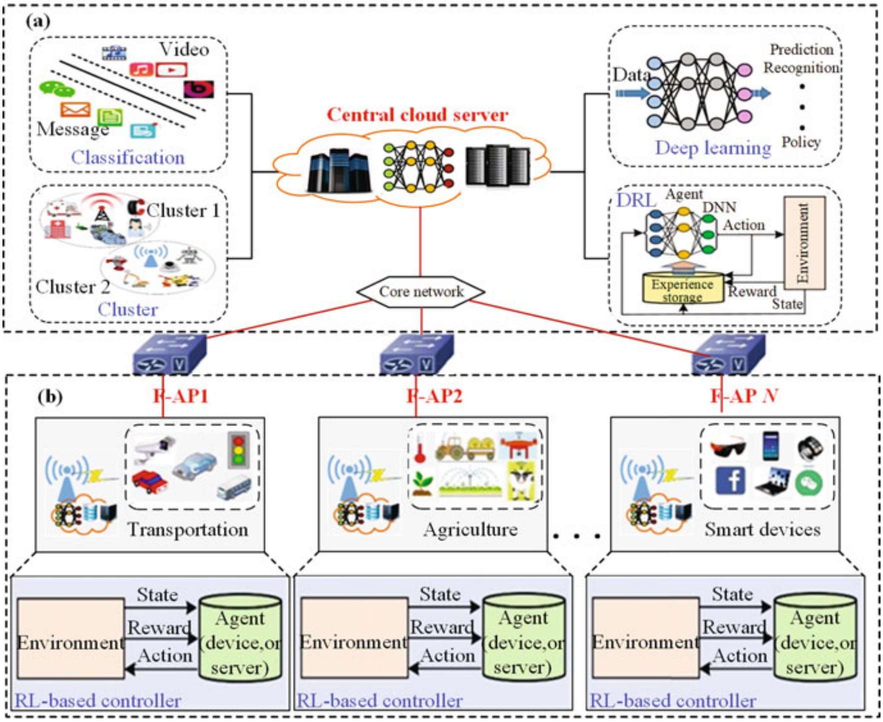


Fig. 2.10 System model for AI-driven F-RANs

2.5.2 Principles of AI in F-RANs

Machine learning (ML) is capable of solving complex problems without explicit programming, and AI is concentrated on empowering machines with intelligence imitating the human mind. ML can be categorized into supervised, unsupervised, and reinforcement learning. Supervised learning aims to determine a general rule that maps inputs to outputs based on the training samples whose inputs are paired with the outputs. In contrast to supervised learning, unsupervised learning aims to infer the underlying structure of data without any label, and hence it is suitable for exploratory analysis, such as the clustering of fog nodes.

Recently, some popular tools of modern ML, such as deep neural networks (DNNs) and deep reinforcement learning (DRL) algorithms, have led to impressive advances in different practical areas, e.g., natural language processing, image recognition, and recommendation systems. For example, the DNNs with supervised learning can be applied to power control and beamforming design of fog nodes in F-RANs. Also, these tools can be tailored to the issues in F-RANs.

On the other hand, AI has evolved to multi-disciplinary techniques including optimization theory, ML, and game theory. Among these techniques, ML plays an indispensable role. ML equips a computer with a “brain” to learn environment from high-dimensional raw input data and make intelligent decisions. For instance, ML is capable of extracting features from a great volume of raw data from using DNNs for further analysis. More attractively, it can tolerant incomplete or even erroneous input raw data. Furthermore, ML can imitate human brains to learn from the interaction with the dynamic environment and make control decisions using DRL algorithms according to the feedback of network states and system rewards from the environment.

DRL is particularly advantageous in handling complicated control problems, and it can be used to solve the complicated resource allocation problems in F-RANs. In Sun et al. (2019a), we present a DRL framework for a dynamic F-RAN to allow the network controller in the cloud to properly select communication modes for users and control the on–off states of processors.

The advancement of DNNs sheds light on data processing. Specifically, convolutional neural networks (CNNs) have a strong ability of extracting spatial features from input signals using convolutional operations. Different from neural feedforward networks, recurrent neural networks (RNNs) have their internal states, which allow RNNs to exhibit temporal dynamic behavior and are suitable for processing the sequence data like handwriting, speech, vehicle trajectory, etc.

For most of the existing learning-based caching schemes, a central server gathers users’ data for the purpose of training. However, some of users’ data include sensitive and private information, such as age, gender, locations. Uploading these information to a central server presents a risk of breaching users’ privacy. Federated learning (FL) as an alternative distributed machine learning approach proposes to train a high quality model without gathering the data from users. FL leverages users’ local data and computation capacity of their devices to perform distributed training. In the federated setting, multiple communication rounds are run to achieve a high quality model. In each communication round, the F-AP distributes the current model to selected mobile users. These users compute an update to the current model independently (e.g., taking several iterations of gradient descent) by using their local data. Next, the model update from each selected user is uploaded to the associated F-AP where all updates are aggregated to generate a new shared global model by the federated average method. Typically, the federated average method is a weighted average, where the weight of a user depends on the user’s data size. That means a user with more data accounts for more contribution to the shared global model. A federated learning communication round for an F-AP and its associated users often consists of the following four steps:

1. The selected mobile users who are plugged-in and have good network connection participate in a federated communication round. Each of them downloads the current global model from the associated F-AP.
2. Using the local dataset, each user computes the updates for the downloaded global model.
3. The calculated model updates are uploaded to the F-AP from each user.
4. The F-AP aggregates all model updates to construct an improved global model.

Compared to the traditional centralized learning approach, the distinct advantage of FL can reduce privacy and security risks as uploading model updates is much more secure than uploading users' data.

2.5.3 Challenges and Future Works

To improve the flexibility and performances of F-RANs, AI has been proposed in the physical layer, medium access control layer, networking layer, and even application layer in F-RANs. Although initial studies and researches have been conducted on the applications of AI into F-RANs, several challenges and future works will be introduced and discussed in this section to facilitate further research in this area (Elsayed and Kantarci 2019).

A. AI-Driven Backhaul and Fronthaul

In F-RANs, backhaul and fronthaul are very important. For example, fronthaul has a great impact on the capacity and latency performances of C-RAN mode, and backhaul determines the cooperation performance among F-APs, RRHs, and small cell APs. There are various backhaul and fronthaul solutions in practical, such as wired communications like fiber and cable, wireless communications like the sub-6 GHz band, mmW, and Terahertz wave. Each solution will consume different amount of energy with different allowable frequency bandwidth, and how to configure and optimize these solutions are challenging. In this case, AI can be used to select, configure, and optimize these different solutions for backhaul and fronthaul. Consequently, the AI-driven backhaul and fronthaul techniques should be researched and exploited to further improve performances of F-RANs in the future.

B. AI-Driven Architecture and Infrastructure Design

In F-RANs, communication, computing, and cache are strictly coupled, and it is hard to optimize them simultaneously. Thanks to AI, it shows a new way to optimize them in F-RANs. The traditional system architecture cannot be efficiently suitable for F-RANs because they have to be adaptive to different application scenarios with diverse performance requirements. In particular, powerful servers embedded with GPUs could be deployed at the edge devices of F-RANs to implement AI based data analysis, radio signal processing, multi-dimensional resource management, and routing. Since AI mainly depends on the collected measurement data, it is necessary to store and analyze the data at the network edge, which results in achieving in-time data analysis. Furthermore, on the basis of SDN, NFV and AI, F-RANs could be flexible and configurable.

C. AI-Driven Network Slicing

Network slicing as an advanced technique in 5G is to re-organize the flexible system architectures and reschedule appropriate computing, caching, backhaul, fronthaul, and radio access communication resources on-demand to meet the performance requirements of different slice instances. In F-RANs, AI can enhance network slicing to work efficiently and have good performances. In particular, network slicing in F-RANs can benefit from AI in several aspects under slice isolation constraints: (1) AI helps learn the real-time mapping information from the application demands to resource management strategies, which results in constructing a suitable network slice instance quickly and efficiently. (2) Through using some advanced AI schemes, such as transfer learning and federated learning, the knowledge about mapping and resource management strategies in one application scenario can be useful in another application scenario, which speeds AI algorithms up. The advanced AI schemes that can be efficiently used in the network slicing of F-RANs should be further researched and exploited in the future (Xiang et al. 2020).

D. Standard Promotions and Applications

To promote AI applied in F-RANs and improve performances of AI-driven F-RANs, it is necessary to define the corresponding standards. What impacts on the protocol and interfaces when AI used in F-RANs? What structures and standards should be enhanced, added, updated, and even deleted? Meanwhile, it is urgent to present common standard related challenges in F-RANs to make researchers focus on the special AI algorithms that can be efficiently used in F-RANs and define the corresponding standards to take fair comparisons among different AI algorithms. Meanwhile, several hardware testbed and trail test related standards for AI-driven F-RANs should be researched, which are the basements to promote Ai-driven F-RANs commercial rollout successfully.

2.6 Summary

In this chapter, the traditional fog radio access network (F-RAN) architecture has been presented, which features a hierarchical architecture. Moreover, network slicing in F-RANs has been identified as a cost-efficient way to achieve flexible networking, and three key techniques for the radio interface have been summarized, which include non-orthogonal multiple access, interference control, and transmission mode selection. Meanwhile, several use cases of F-RANs have been elaborated, namely F-RAN enabled vehicles-to-anything communication, F-RAN enabled space communication, etc. Finally, we have envisioned that F-RANs will integrate with artificial intelligence to fully unleash its potential and the architecture, principles, and open issues have been discussed.

References

An J et al. (2019) EIF: toward an elastic IoT fog framework for AI services. IEEE Commun Mag 57(5):28–33

Bastug E et al. (2017) Toward interconnected virtual reality: opportunities, challenges, and enablers. IEEE Commun Mag 55(6):110–117

Chen S et al. (2017) Vehicle-to-everything (V2X) services supported by LTE-based systems and 5G. IEEE Commun Stand Mag 1(2):70–76

Dang T, Peng M (2019) Joint radio communication, caching and computing design for mobile virtual reality delivery in fog radio access networks. IEEE J Sel Areas Commun 37(7):1594–1607

Di B et al. (2019) Ultra-dense LEO: Integration of satellite access networks into 5G and beyond. IEEE Wirel Commun 26(2):62–69

Ding Z et al. (2018) Embracing non-orthogonal multiple access in future wireless networks. Front Inf Technol Electron Eng 19(3):322–339

Elsayed M, Kantarci ME (2019) Ai-enabled future wireless networks: challenges, opportunities, and open issues. IEEE Veh Technol Mag 14(3):70–77

Gu X et al. (2018) Outage probability analysis of non-orthogonal multiple access in cloud radio access networks. IEEE Commun Lett 22(1):149–152

Guidotti A et al. (2019) Architectures and key technical challenges for 5G systems incorporating satellites. IEEE Trans Veh Technol 68(3):2624–2639

Li Y et al. (2019) Resource allocation for optimizing energy efficiency in noma-based fog uav wireless networks. IEEE Netw. https://doi.org/10.1109/MNET.001.1900231

Liao S et al. (2019) Fog-enabled vehicle as a service for computing geographical migration in smart cities. IEEE Access 7:8726–8736

Peng M, Zhang K (2016) Recent advances in fog radio access networks: performance analysis and radio resource allocation. IEEE Access 4:5003–5009

Peng M et al. (2016) Fog computing based radio access networks: issues and challenges. IEEE Netw 30(4):46–53

Qi L et al. (2019) Advanced user association in non-orthogonal multiple access based fog radio access networks. IEEE Trans Commun 67(12):8408–8421

Sun Y et al. (2019a) Application of machine learning in wireless networks: key techniques and open issues. IEEE Commun Surv Tutorials 21(4):3072–3108

Sun Y et al. (2019b) A game-theoretic approach to cache and radio resource management in fog radio access networks. IEEE Trans Veh Technol 68(10):10,145–10,159

Xiang H et al. (2017) Network slicing in fog radio access networks: issues and challenges. IEEE Commun Mag 55(12):110–116

Xiang H et al. (2020) A realization of fog-ran slicing via deep reinforcement learning. IEEE Trans Wireless Commun 19(4):2515–2527. https://doi.org/10.1109/TWC.2020.2965927

Yan S et al. (2017) An evolutionary game for user access mode selection in fog radio access networks. IEEE Access 5:2200–2210

Yan S et al. (2019) A game theory approach for joint access selection and resource allocation in uav assisted IoT communication networks. IEEE IoT J 6(2):1663–1674

Yan S et al. (2020) Joint user access mode selection and content popularity prediction in non-orthogonal multiple access based f-rans. IEEE Trans Commun 68(1):654–666

Zhao Z et al. (2017) A non-orthogonal multiple access (noma)-based multicast scheme in wireless content caching networks. IEEE J Sel Areas Commun 35(12):2723–2735

Zhang X et al. (2019) Deep reinforcement learning based mode selection and resource allocation for cellular V2X communications. IEEE IoT J. https://doi.org/10.1109/JIOT.2019.2962715

Chapter 3
Theoretical Performance Analysis of Fog Radio Access Networks

The rest of this chapter is organized as follows: in Sect. 3.1, the ergodic capacity of user association in F-RANs is studied. Then, the effective capacity with content caching in F-RANs is analyzed in Sect. 3.2. Finally, the concluding marks of this chapter are provided by Sect. 3.3.

3.1 Ergodic Capacity of User Association in F-RANs

Unlike the conventional cell networks, amorphous coverage is provided by F-RANs. Therefore, the interference circumstance is difficult to be captured. It is challenging to evaluate the performance of interference coordination strategies, especially the collaborated user association schemes, and the transmission performance of F-RANs cannot be guaranteed. In this part, the ergodic capacity of different user association schemes are studied by establishing a stochastic geometry-based model, which can provide some insights for networking of F-RANs.

3.1.1 System Model

Consider an uplink transmission scenario in F-RANs, which consists of multiple F-APs and a single user U. Each F-AP is equipped with K antennas, while the user is equipped with a single antenna. To characterize the amorphous coverage of F-RANs, the locations of F-APs are modeled as a two-dimensional PPP Φ with intensity γ in a disc region $\mathbb{D}^2$, whose radius is denoted as R. The number of F-APs located in $\mathbb{D}^2$, which is denoted as M_R, follows Poisson distribution, i.e., $\Pr\left(M_R = m\right) = \left({\mu_D}^{M_R} / \left(M_R\right)!\right) e^{-\mu_D}$, and $\mu_D = \pi R^2 \gamma$.

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0_3

To achieve full diversity gains, the maximal ratio combining (MRC) transmission strategy is considered, and the corresponding received signal-to-noise-ratio (SNR) can be expressed as follows when U associated with the i-th F-AP:

$$\varphi_i = \frac{\rho_U d_i^{-\beta} G_i}{\omega^2}, \tag{3.1}$$

where d_i denotes the distance between the user U and the i-th potential associated F-AP, β denotes the value of path loss exponent, the transmit power of U is denoted as ρ_U, ω^2 is the power of noise, G_i is the combined flat channel gain, i.e., $G_i = \sum_{l=1}^{K} |g_{il}|^2$, and g_{il} denotes the flat channel gain between U and the l-th antenna of the i-th F-AP. In particular, Rayleigh flat channel fading model is employed, i.e., $g_{il} \sim CN(0, 1)$.

In this part, the following two user association strategies are considered:

- Single nearest F-AP association strategy: Since the channel condition is mainly determined by path loss, U associates with the nearest F-AP to guarantee the transmission reliability.
- M-nearest F-AP association strategy: U associates with multiple F-APs simultaneously, and the associated F-APs are selected based on the distance, i.e., U associates with M-nearest F-APs, and the MRC is employed among the associated F-APs.

Apparently, the transmission reliability can be improved by increasing the number of associated F-APs. However, the global CSI of associated F-APs are required to implement collaborated signal processing, which cause extra signaling overhead. Therefore, there exist a tradeoff between the signaling cost and transmission performance. To provide some insights of user association in F-RANs, the ergodic capacity of two aforementioned user association strategies is analyzed in the next part.

3.1.2 Performance Analysis of Ergodic Capacity

In this part, tractable expressions of ergodic capacity are provided to evaluate the performance of different user association strategies. Then some approximation results, i.e., the asymptotic analytical results and closed-form upper bound, are derived, which can provide some insights with respect to the impact of networking schemes.

Ergodic Capacity of Single Nearest F-AP Association Strategy

To derive the analytical results, the ergodic capacity can be expressed as follows:

$$\zeta_{1R} = \int_0^\infty p_{\varphi_{1R}}(\varphi)\log_2(1+\varphi)\,\mathrm{d}\varphi, \tag{3.2}$$

where $p_{\varphi_{1R}}$ denotes the PDF of received SNR φ_{1R}. To derive $p_{\varphi_{1R}}$, the outage probability of single nearest F-AP association strategy is firstly provided, which is defined as the occurrence of the event that the received SNR is smaller than a given threshold, i.e.,

$$\vartheta_{out\_1R} = \Pr[\varphi < \Psi], \tag{3.3}$$

where Ψ denotes the given threshold. A closed-form expression of outage probability with respect to the single nearest F-AP association strategy can be given as the following lemma.

Lemma *The outage probability of single nearest F-AP association strategy in our considered F-RAN transmission scenario can be expressed as*

$$\vartheta_{out\_1R} = \int_0^\infty \frac{\Upsilon\left(K, \frac{d^\beta\Psi}{P}\right)}{(K-1)!} e^{-\gamma\pi d^2} 2\pi\gamma d\,\mathrm{d}d, \tag{3.4}$$

where $\Upsilon(a,b)$ *denotes the lower incomplete gamma function defined by XX, and* $P = \frac{\rho_U}{\omega^2}$.

Proof First, the distribution of flat channel fading is provided. As introduced previously, the MRC strategy is employed, and thus the combined channel gains can be treated as a summation of K random variables, which follow identically independent exponential distribution. The corresponding PDF can be expressed as

$$p_{G_i}(x) = \frac{x^{K-1}e^{-x}}{(K-1)!}. \tag{3.5}$$

Next, we focus on the distance d_i between U and the nearest F-AP. Based on our established system model, the PDF of d_i can be expressed as follows (Andrews et al. 2011):

$$p_d(d) = e^{-\gamma\pi d^2} 2\pi\gamma d, d > 0 \tag{3.6}$$

Based on (3.5) and (3.6), the outage probability can be derived as

$$\begin{aligned}\vartheta_{out\_1R} &= \Pr[\varphi < \Psi] = \mathrm{E}\left[\Pr\left[PGd^{-\beta} < \Psi\right]\middle| d\right] \\ &= \int_0^\infty \frac{\Upsilon\left(K, \frac{d^\beta\Psi}{P}\right)}{(K-1)!} e^{-\gamma\pi d^2} 2\pi\gamma d\,\mathrm{d}d. \end{aligned}\tag{3.7}$$

And the proof has been finished.

Based on (3.2) and (3.4), the asymptotic analytical results of ergodic capacity in the high SNR region can be provided by the following theorem.

Theorem *In the high SNR region, the ergodic capacity of nearest F-AP user association strategy can be approximated as*

$$\zeta_{1R} = \frac{\sum\limits_{i=1}^{K-1} \frac{1}{i} + \frac{\beta}{2}\left[\ln\left(\pi\gamma\right) + C\right] - C + \ln\left(\rho/\omega^2\right)}{\ln(2)}, \tag{3.8}$$

where C denotes Euler's constant.

Proof Recalling (3.2), the PDF of $p_{\varphi_{1R}}\left(\varphi\right)$ can be derived as follows based on (3.4):

$$p_{\varphi_{1R}}\left(\varphi\right) = \frac{\partial\left(\int_0^\infty \Pr\left[G < \frac{d_1^\beta \Psi}{P}\right] e^{-\gamma\pi d_1^2} 2\pi\gamma d_1 \mathrm{d}d_1\right)}{\partial\Psi}. \tag{3.9}$$

In the high SNR region, the ergodic capacity can be approximated as follows by substituting (3.9) into (3.2):

$$\begin{aligned}\zeta_{1R} &\approx \int_0^\infty \int_0^\infty \frac{a^K \left(\varphi\right)^{K-1} e^{-ax}}{(K-1)!} e^{-\gamma\pi d^2} 2\pi\gamma d \mathrm{d}d \log_2\left(\varphi\right) \mathrm{d}\varphi \\ &= \frac{1}{\ln(2)} \int_0^\infty \left[\sum_{i=1}^{K-1} \frac{1}{i} - C - \ln\left(\frac{d^\beta}{P}\right)\right] e^{-\gamma\pi d^2} 2\pi\gamma d \mathrm{d}d \\ &= \frac{\sum\limits_{i=1}^{K-1} \frac{1}{i} + \frac{\beta}{2}\left[\ln\left(\pi\gamma\right) + C\right] - C + \ln\left(\rho/\omega^2\right)}{\ln(2)},\end{aligned} \tag{3.10}$$

where the approximation (a) in (3.10) can be obtained based on $\log_2\left(1+\varphi\right) \sim \log_2\left(\varphi\right)$ in the high SNR region. The proof has been finished.

As shown in (3.8), it indicates that the ergodic capacity is non-linearly increasing with respect to the number of antennas equipped by each F-AP K, the density of F-APs γ, and the transmit power ρ_U, when the single nearest F-AP association strategy is employed.

Ergodic Capacity of *M*-Nearest F-AP Association Strategy

When U associates with M F-APs, the received SNR can be expressed as follows when the MRC strategy is employed:

$$\varphi_M = \sum_{i=1}^{M} \frac{\rho G_i d_i^{-\beta}}{\omega^2}. \tag{3.11}$$

To obtain some tractable results, we consider two different cases of M, and the corresponding derivation is provided as follows.

(1) *When* $M = 2$: By following the derivation paradigm of single nearest F-AP association strategy, the outage probability is firstly provided by the following lemma.

Lemma *When U associates with 2-nearest F-APs, the outage probability is expressed as*

$$\vartheta_{out\_2R} \approx \int_{\left(\frac{2PK}{\Psi}\right)^{\frac{1}{\beta}}}^{\infty} 2\pi^2\gamma^2 d_2 \left[d_2^2 - \left(\frac{PKd_2^{\beta}}{\Psi d_2^{\beta} - PK} \right)^{\frac{2}{\beta}} \right] e^{-\pi\gamma d_2^2} dd_2. \tag{3.12}$$

Proof The outage probability can be expressed as follows when $M = 2$:

$$\vartheta_{out\_2R} = \Pr\left[Pd_1^{-\beta}G_1 + Pd_2^{-\beta}G_2 < \Psi \right], \tag{3.13}$$

where d_1 and d_2 denote distance between U and the nearest and the second nearest F-APs, respectively. To derive the outage probability, we first study the joint CDF of d_1 and d_2, which can be denoted as $\Pr(d_1, d_2)$. It is equivalent to derived the probability that there is no more than one F-APs located in a ring from the radius d_1 to d_2, which can be derived as

$$\begin{aligned}\Pr(d_1, d_2) = \Pr\left(\text{null } \in \odot d_1, \text{only one F} - \text{AP } \in \phi_{d_1 d_2}\right)\\ \cup \Pr\left(\text{null } \in \odot d_1, \text{null} \in \phi_{d_1 d_2}\right),\end{aligned} \tag{3.14}$$

where $\odot d_1$ denotes a disc centered at the origin with a given radius d_1, and $\phi_{d_1 d_2}$ is a ring that is concentric with $\odot d_1$, and its radius is from d_1 to d_2. Due to the implementation of PPP-based model, the joint CDF can be written as

$$\begin{aligned}&\Pr(d_1, d_2)\\ &= \left(e^{-\gamma\pi\left(d_2^2 - d_1^2\right)} + \left(\gamma\pi\left(d_2^2 - d_1^2\right)\right) e^{-\gamma\pi\left(d_2^2 - d_1^2\right)} \right) e^{-\gamma\pi\left(d_1^2\right)}.\end{aligned} \tag{3.15}$$

Based on (3.15), the joint PDF can be expressed as

$$p(d_1, d_2) = 4\pi^2\gamma^2 d_1 d_2 e^{-\pi\gamma d_2^2}. \tag{3.16}$$

By substituting (3.16) into (3.13), ϑ_{out_2R} can be expressed as

$$\vartheta_{out\_2R} = \int_{\left(\frac{\mathrm{E}\{PG_1+PG_2\}}{2}\right)^{\frac{1}{\beta}}}^{\infty} \int_{\left(\frac{\mathrm{E}\{PG_1\}d_2^{\beta}}{\Psi d_2^{\beta}-\mathrm{E}\{PG_1\}}\right)^{\frac{1}{\beta}}}^{d_2} 4\pi^2\gamma^2 d_1 d_2 e^{-\pi\gamma d_2^2} dd_1 dd_2. \tag{3.17}$$

Then (3.12) can be derived by taking expectation of G_i, and the proof has been finished.

Based on (3.12), the PDF of received SNR can be approximated as follows when $M = 2$:

$$p_{\varphi 2R}(\varphi) = \int_{\left(\frac{2KP}{\Psi}\right)^{\frac{2}{\beta}}}^{\infty} \frac{2\pi^2\gamma^2(KP)^{\frac{2}{\beta}}}{\beta} \left(\Psi - (KP)t^{-\frac{\beta}{2}}\right)^{-\frac{2}{\beta}-1} e^{-\pi\gamma t}\mathrm{d}t. \tag{3.18}$$

Then the ergodic capacity can be derived as follows based on (3.2):

$$\begin{aligned} \zeta_{2R} = {} & \int_0^{\infty}\int_{\left(\frac{2KP}{W}\right)^{\frac{2}{\beta}}}^{\infty} \frac{2\pi^2\gamma^2(KP)^{\frac{2}{\beta}}}{\beta} \left(W - (KP)t^{-\frac{\beta}{2}}\right)^{-\frac{2}{\beta}-1} \\ & \times e^{-\pi\gamma t}\log(1+W)\,\mathrm{d}t\mathrm{d}W. \end{aligned} \tag{3.19}$$

Furthermore, when $\beta = 4$, a simplified closed-form expression can be obtained, which can be expressed as

$$\begin{aligned} \zeta_{2R}^{\beta=4} &= \int_0^{\infty} \frac{2\pi^2\gamma^2(KP)^{1/2}e^{-\pi\gamma t}}{4\ln 2} \\ &\quad \times \left\{ 2\left[-\frac{\ln W}{\sqrt{W-(KP)t^{-2}}}\Bigg|_{\left(\frac{2KP}{t^2}\right)}^{\infty} + \frac{2}{\sqrt{(KP)t^{-2}}} \right.\right. \\ &\qquad \left.\left. \times \ \arctan\left[\frac{\sqrt{W-(KP)t^{-2}}}{\sqrt{(KP)t^{-2}}}\right]\Bigg|_{\left(\frac{2KP}{t^2}\right)}^{\infty} \right]\right\} \mathrm{d}t \\ &= \frac{\ln(2PK)+\pi/2-2+2C+2\ln(\pi\gamma)}{\ln 2}. \end{aligned} \tag{3.20}$$

As shown in which (3.20), the ergodic capacity can be improved by increasing the density of F-APs and the number of antennas of each F-AP when $\beta = 4$.

(2) *When* $M > 2$: To generalize Case *(1)* the key step is to capture the joint distribution of summation with respect to path loss of different F-APs, i.e., $\sum_{i=1}^{M} d_i^{-\beta}$. First, we focus on the expectation of path loss with respect to a single F-AP. Based on our established two-dimensional PPP model, the PDF of $x = \pi\gamma d_i^2$ can be given as $p(x) = \frac{x^{i-1}e^{-x}}{(i-1)!}$, and then the expectation of $d_i^{-\beta}$ is derived as

$$\mathrm{E}\left\{d_i^{-\beta}\right\} = (\pi\gamma)^{\frac{\beta}{2}} \int_0^\infty x^{-\frac{\beta}{2}} p(x)\mathrm{d}x = (\pi\gamma)^{\frac{\beta}{2}} \frac{\Gamma\left(i-\frac{\beta}{2}\right)}{\Gamma(i)}. \tag{3.21}$$

Please note that $\Gamma\left(i-\frac{\beta}{2}\right)$ is finite if and only if $i<\beta/2$. Based on (3.21), the outage probability can be given as follows when $M \geq \lfloor \beta/2 \rfloor + 1$:

$$\begin{aligned} \vartheta_{out\_MR} &= \Pr\left[\sum_{i=1}^{\lfloor \beta/2 \rfloor} PG_i d_i^{-\beta} + \sum_{i=\lfloor \beta/2 \rfloor+1}^{M} PG_i d_i^{-\beta} < \Psi\right] \\ &\approx \Pr\left[\sum_{i=1}^{\lfloor \beta/2 \rfloor} PG_i d_i^{-\beta} + PK \sum_{i=\lfloor \beta/2 \rfloor+1}^{M} (\pi\gamma)^2 \frac{\Gamma\left(i-\frac{\beta}{2}\right)}{\Gamma(i)} < \Psi\right], \end{aligned} \tag{3.22}$$

where $\lfloor\cdot\rfloor$ is the floor function.

Then, we focus on a special case when $\beta=4$, and (3.22) can be further derived as

$$\begin{aligned} \vartheta_{out\_MR}^{\beta=4} &\approx \Pr\left[\sum_{i=1}^{2} PG_i d_i^{-4} + \underbrace{PK(\pi\gamma)^2 \frac{M-2}{M-1}}_{S} < \Psi\right] \\ &= \int_{\left(\frac{2PK}{\Psi-S}\right)^{\frac{1}{4}}}^{\infty} 2\pi^2\gamma^2 d_2 \left[d_2^2 - \sqrt{\frac{PKd_2^4}{(\Psi-S)\,d_2^4 - PK}}\right] e^{-\pi\gamma d_2^2} dd_2. \end{aligned} \tag{3.23}$$

By substituting (3.23) into (3.2), the ergodic capacity can be derived as follows when U associates with M nearest F-APs, $M>2$:

$$\begin{aligned} \zeta_{MR}^{\beta=4} =& \int_0^\infty \pi^2\gamma^2 e^{-\pi\gamma t}\sqrt{KPt} \\ &\times \left[\ln\left(2KP+St^2\right) - \ln\left(t^2\right) + \frac{2}{\sqrt{KP+St^2}} \arctan\left(\sqrt{KP+St^2}\right)\right] \mathrm{d}t. \end{aligned} \tag{3.24}$$

In particular, a closed-form PDF of SNR can be derived as follows when $M\to\infty$ and $\beta=4$:

$$p_\infty(\varphi) = \frac{\pi\gamma\sqrt{KP}}{2\Psi^{3/2}} \exp\left(-\frac{KP\pi^3\gamma^4}{4\Psi}\right), \tag{3.25}$$

Based on (3.25), a tractable upper bound of ergodic capacity can be expressed as

$$\begin{aligned}\zeta^{Upper} &= \int_0^\infty \frac{\pi\gamma\sqrt{KP}}{2\Psi^{3/2}} \exp\left(-\frac{KP\pi^3\gamma^4}{4\Psi}\right) \log_2\left(1+\Psi\right) \mathrm{d}\Psi \\ &\approx \frac{C - \sum\limits_{j=0}^{\infty} \frac{1}{(j+1)(2j+1)} + \ln \frac{KP\pi^3\gamma^4}{4}}{\ln 2}.\end{aligned} \tag{3.26}$$

Equation (3.26) indicates that ζ^{Upper} is mainly determined by K, ρ_U, and γ, and γ is the leading term since it is with the highest exponent.

3.1.3 Numerical Results

In this part, the simulation results of ergodic capacity are provided to verify the accuracy of derived closed-form expressions, and evaluate the impact of γ, K and ρ_U. In particular, the number of antennas equipped with each F-AP is set as $K=4$, the path loss exponent is $\beta=4$, the radius of the disc is set as $R=$ 600 m, and the intensity of F-APs is $\gamma=10^{-4}$. The power density of noise is − 174 dBm/Hz, and the bandwidth is set as 100 MHz.

To evaluate the performance of different user association strategies, the ergodic capacity is plotted with different settings of the number of associated F-APs M in Fig. 3.1. As shown in the figure, the ergodic capacity increases monotonically as the transmit power increases, since the impact of co-channel interference can be mitigated by enlarging the scale of cooperative processing. Moreover, the Monte Carlo simulation results coincide with the numerical results based on derived analytical results, which can verify the accuracy of analytical results. The simulation results also show that the ergodic capacity can be improved by associating more F-APs. In particular, compared with the single F-AP association strategy, the performance can be improved significantly when U associates with 2 F-APs. However, the performance gain becomes saturated as M increases, and finally approaches the performance limits.

In Fig. 3.2, the ergodic capacity is plotted as a function with respect to the number of antennas equipped with each F-AP K, where the number of associated F-APs is set $M=$ 1, 2, 4, and 8, respectively. Similar to Fig. 3.1, the ergodic capacity keeps increasing as K increases. Moreover, the performance gain is not significant when U associates with more than 4 F-APs. In particular, when $K=4$, the ergodic capacity can be improved by 0.58 bps/Hz when the number of associated F-APs is increased from 1 to 2, while the performance gain is 0.28 bps/Hz when the number of associated F-APs is increased from 2 to 4. Moreover, the ergodic capacity is 9.21 bps/Hz when $K=8$ and $M=2$, and it is 8.06 bps/Hz when $K=2$ and $M=8$. All these results show that the ergodic capacity can be increased by enlarging K.

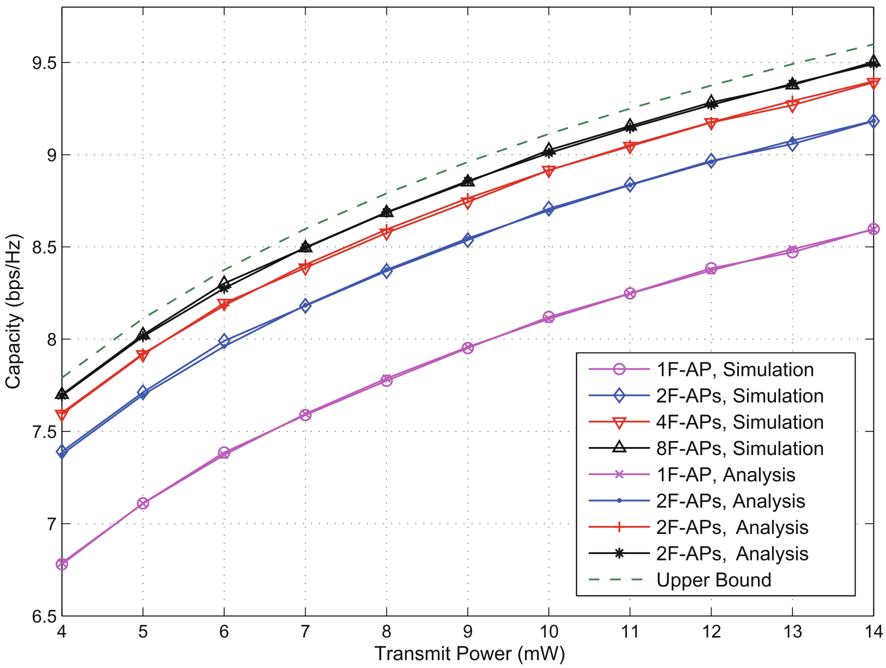


Fig. 3.1 Ergodic capacity versus transmit power ρ_U

3.2 Effective Capacity with Content Caching in F-RANs

Channel capacity cannot characterize the QoS of wireless services, since latency caused by responding and handling the service requests are not considered. Therefore, effective capacity is proposed by Wu et al. (2003), which can be defined as follows: considering a given requirement of QoS, effective capacity is the maximum arrival rate bore by a wireless channel. In this part, the effective capacity is studied to evaluate the performance gain of edge caching in F-RANs.

3.2.1 System Model

Consider a content delivery scenario of F-RANs, where each user requires a content object. As shown in Fig. 3.3, two categories of caches are considered here. A cloud cache is included at the cloud computing layer. In the fog-computing layer, an edge cache is equipped with each F-AP, which is named as F-AP cache. The F-APs connect with the cloud computing center via wired fronthaul. Since the F-AP caches locate more closely to the users than the cloud cache, the latency caused by fronthaul transportation can be avoided if the content objects can be obtained locally.

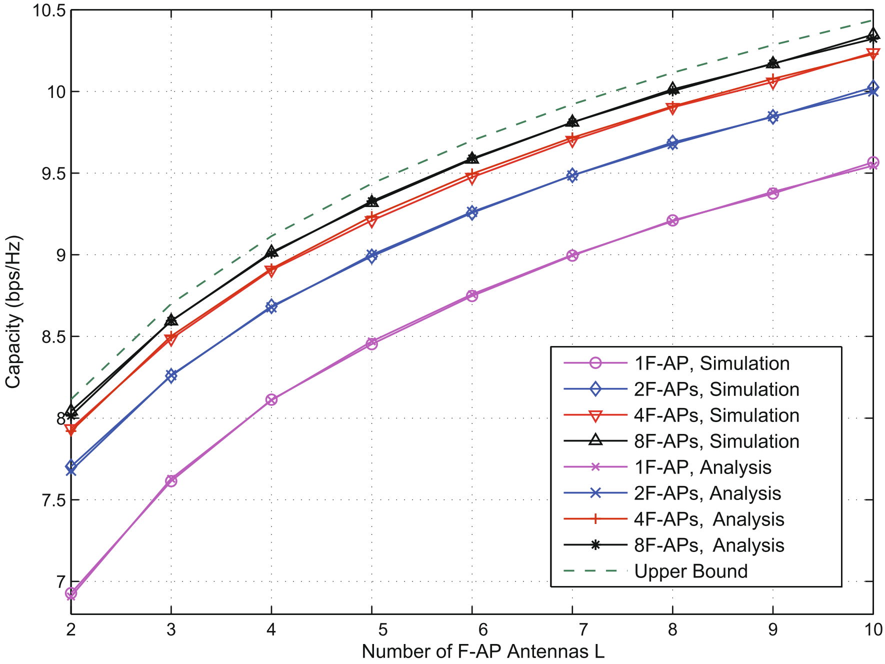


Fig. 3.2 Ergodic capacity versus antenna number *K*

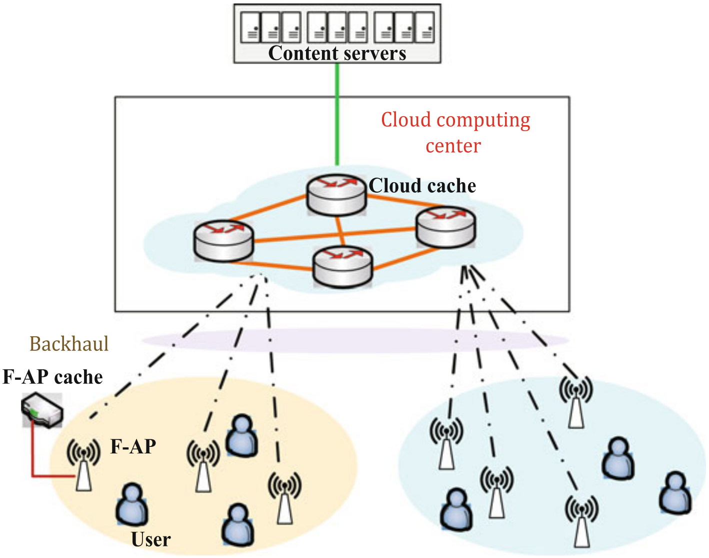


Fig. 3.3 System model of content caching in F-RANs

In F-RANs, each F-AP serves multiple users that required different content objects. The multicast strategy is employed to improve the spectrum efficiency and avoid fronthaul redundancy. When the F-AP receive the requests from the users, they can be responded locally if the requested content object is stored by its own F-AP cache. Otherwise, the F-AP should obtain the content objects from the cloud cache through the fronthaul. Without loss of generality, we focus on a typical F-AP F_T, and the required content object is transmitted to the user via the wireless channel, and the observation at a typical user u_T can be expressed as

$$r_T = \sqrt{P} g_m r_m^{-\alpha/2} x_m + \sum_{j \neq m} \sqrt{P} g_j r_j^{-\alpha/2} x_j + \gamma_0, \tag{3.27}$$

where x_m denotes the required content object for u_T from its associated F-AP R_m, both Rayleigh fading and path loss are considered to model the wireless channel, i.e., g_m denotes the flat Rayleigh fading, r_m denotes the distance between R_m and u_T, the path loss exponent is denoted as α, x_j, g_j, and r_j are defined similarly for an interfere F-AP R_j, $j \neq m$, $g_m, g_j \sim \mathscr{CN}(0, 1)$, γ_0 denotes the additive Gaussian noise, and the average transmit SNR is ω. Based on (3.27), the channel capacity with unit bandwidth can be written as

$$\kappa = \eta \log\left(1 + \omega\right), \text{ where } \omega = \frac{P \underset{m}{\overset{-\alpha}{r}} |g_m|^2}{\sum\limits_{j \neq m} P \underset{j}{\overset{-\alpha}{r}} \left|g_j\right|^2 + \varepsilon^2}, \tag{3.28}$$

where η denotes the spectral efficiency, and it is inversely proportional to the number of occupied radio resource for content transmissions in $\mathscr{C}_T$.

The key idea of effective capacity is to analyze the maximum latency that can be supported by a given transmission capacity. To capture the latency behavior of content delivery, the decay rate of a queue with stochastic length L can be expressed as

$$\phi = \lim_{l \to \infty} \frac{\log \Pr\{L > l\}}{l}. \tag{3.29}$$

When the threshold $l_{\max}$ is large, the queue length violation probability can be estimated as $\Pr\{L > l_{\max}\} \approx e^{-\phi l_{\max}}$ due to the large deviation theory. Hence, ϕ can be defined as the QoS exponent, i.e., it implies a slack QoS requirement when the value of ϕ is small, while a strict QoS requirement is indicated when ϕ is large. If the arrival rate is fixed, i.e., $a(t) = a$, it can be approximated as $\Pr\{R > r_{\max}\} \leq c\sqrt{\Pr\{L > l_{\max}\}}$ when $l_{\max}$ is large, where c is a content, and $l_{\max} = a r_{\max}$.

As introduced in Wu et al. (2003), the effective capacity is defined as follows:

$$E\left(\phi\right)=-\lim_{t\rightarrow\infty}\frac{1}{\phi t}\log E\left\{e^{-\phi O(t)}\right\},\tag{3.30}$$

where $O(t)=\sum_{0=t_0<t_1<\cdots<t_n=t}\int_{t_{i-1}}^{t_i}r\left(\tau\right)\mathrm{d}\tau$ is the transmitted service in a given time interval (0, t]. When the wireless channels are assumed to be block fading channels, the channel coefficient can be treated a constant, and the effective capacity can be expressed as follows:

$$E(\phi)=-\tfrac{1}{\phi T}\log\mathbb{E}\left\{e^{-\eta\phi T_{\mathcal{K}}}\right\}\overset{\text{(a)}}{=}-\tfrac{1}{\phi\bar{T}}\ln\mathbb{E}\left\{(1+\omega)^{-\eta\phi\bar{T}}\right\},\tag{3.31}$$

where $\bar{T}=T/\ln 2$, and (a) in (3.31) is derived by substituting (3.28) into (3.30).

3.2.2 Performance Analysis of Effective Capacity

In this part, a tractable expression of effective capacity is derived to evaluate the impact of edge caching in F-RANs. It can be assumed that all the provided content objects are kept by the cloud content cache $\mathscr{U}_{\mathcal{K}}$, which is modeled as a set $\Lambda_{\mathcal{K}}=\{O_1,\ldots,O_L\}$, and $O_1,\ldots,O_L$ are with the same size. Meanwhile, only a part content objects of $\Lambda_{\mathcal{K}}$ are stored by the edge cache of F_T, which can be denoted as $\Lambda_T=\left\{O_{T_1},\ldots,O_{T_{\mathcal{K}}}\right\}\subseteq\Lambda_{\mathcal{K}}$.

To capture the interference circumstance of F-RANs, the locations of F-APs are modeled as a homogenous PPP Φ_{R}, and its density can be denoted as ρ_{R}. Moreover, the locations of users can be characterized by a homogenous marked PPP $\Psi_{\mathrm{u}}(M_n)$ with density ρ_{u}. The mark M_n denotes the index of content object required by the n-th user U_n.

Effective Capacity of U_i Associated with F_T

To evaluate the QoS performance of content caching in F-RANs, the effective capacity of a specific user is firstly studied. To improve the utility of edge caching, the intra-cluster coordination can be supported, which means that the cached content objects can be shared in the cluster. Without loss of generality, we focus on a typical cluster $\mathscr{C}_T$ that includes F_T, and U_i is an associated user of F_T. The effective capacity of U_i can be provided by the following theorem.

Theorem *Assuming that* U_i*, its effective capacity can be expressed as follows when it associates with* F_T:

$$E_{i,m}\left(\phi_j,r_m\right)=-\frac{1}{\phi_j\bar{T}}\ln\left(\mathscr{G}\left(\phi_j,r_m\right)\right),\tag{3.32}$$

where ϕ_j is the QoS exponent of O_j, r_m denotes the distance between the studied typical user, and F_T, $\mathscr{G}\left(\phi_j, r_m\right)$ is defined as

$$G\left(\phi_j, r_m\right) \\ = \sum_{n=1}^{N}\left(e^{-2\pi A(\alpha)\omega_n^{\frac{2}{\alpha}}\rho_{\mathrm{R}}r_m^2-\frac{\omega_n r_m^{\alpha}\varepsilon^2}{P}} - e^{-2\pi A(\alpha)\omega_{n+1}^{\frac{2}{\alpha}}\rho_{\mathrm{R}}r_m^2-\frac{\omega_{n+1}r_m^{\alpha}\varepsilon^2}{P}}\right)\left(1+\bar{\omega}_n\right)^{-\eta\phi_j\bar{T}}, \tag{3.33}$$

and η denotes the spectrum efficiency, ω_n is the a given threshold of SINR, i.e., $[\omega_n, \omega_{n+1})$ denotes the n-th interval of SINR, $0 \leqslant \omega_1 < \cdots < \omega_n < \cdots < \omega_{N+1}$, $A\left(\alpha\right) = \frac{1}{\alpha}\Gamma\left(\frac{2}{\alpha}\right)\Gamma\left(1-\frac{2}{\alpha}\right)$, and $\Gamma(x)$ is the gamma function.

Proof Based on (3.31), we first focus on the expectation of $Y = (1+\omega)^{-\eta\phi_j\bar{T}}$. In particular, the range of Y is divided into disjoint intervals, and the corresponding endpoints are denoted as $\omega_1, \ldots, \omega_n, \ldots, \omega_{N+1}$. Then, $\mathbb{E}\{Y\}$ can be approximated as

$$\mathbb{E}\{Y\} \approx \sum_{n=1}^{N}\left(\Pr\{\omega < \omega_{n+1}\} - \Pr\{\omega < \omega_n\}\right)\left(1+\bar{\omega}_n\right)^{-\eta\phi_j\bar{T}}, \tag{3.34}$$

where $\bar{\omega}_n$ denotes the representative value of ω with respect to the n-th interval. As shown in (3.34), the outage probability $\Pr\{\omega < \omega_n\}$ should be derived to obtain (3.32). It is equivalent to derive the probability that U_i fails to access F_T, which can be written as

$$\begin{aligned}\Pr\left\{\omega < \omega_n\right\} &= \mathbb{E}_{\Phi_{\mathrm{R}}, g_m, \mathrm{R}_j\in\Phi_{\mathrm{R}}/\mathrm{R}_m}\left\{\Pr\left\{|g_m|^2 < \frac{\omega_n r_m^{\alpha}}{P}\left(\mathscr{P}+\varepsilon^2\right)\right\}\right\} \\ &= 1 - \underbrace{\mathbb{E}_{\Phi_{\mathrm{R}}, \mathrm{R}_j\in\Phi_{\mathrm{R}}/\mathrm{R}_m}\left\{e^{-\frac{\omega_i r_m^{\alpha}}{P}\left(\mathscr{P}+\varepsilon^2\right)}\right\}}_{\mathscr{T}_1},\end{aligned} \tag{3.35}$$

where $P = \sum_{\mathrm{R}_j\in\Phi_{\mathrm{R}}/\mathrm{R}_m} P r_j^{-\alpha}\left|g_j\right|^2$. The last equation in (3.35) can be obtained due to the fact that $|g_m|^2$ follows exponential distribution. Next, T_1 in (3.35) can be further derived as

$$\begin{aligned}\mathscr{T}_1 &= e^{-\frac{\omega_n r_m^{\alpha}\varepsilon^2}{P}}\mathbb{E}_{\Phi_{\mathrm{R}}}\left\{\prod_{\mathrm{R}_j\in\Phi_{\mathrm{R}}/\mathrm{R}_m}\mathbb{E}_{g_j}\left\{e^{-\omega_i r_m^{\alpha} r_j^{-\alpha}|g_j|^2}\right\}\right\} \\ &= e^{-\frac{\omega_n r_m^{\alpha}\varepsilon^2}{P}}\mathbb{E}_{\Phi_{\mathrm{R}}}\left\{\prod_{\mathrm{R}_j\in\Phi_l/\mathrm{R}_m}\left(\int_0^{\infty} e^{-(\omega_n r_m^{\alpha} r_j^{-\alpha}+1)x}\mathrm{d}x\right)\right\} \\ &= e^{-\frac{\omega_n r_m^{\alpha}\varepsilon^2}{P}}\mathbb{E}_{\Phi_{\mathrm{R}}}\left\{\prod_{\mathrm{R}_j\in\Phi_{\mathrm{R}}/\mathrm{R}_m}\frac{1}{1+\omega_n r_m^{\alpha} r_j^{-\alpha}}\right\}.\end{aligned} \tag{3.36}$$

Equation (3.36) can be derived since flat Rayleigh fading is assumed to be identically and independent distributed, and it can be expressed as follows due to the probability generating functional of PPP:

$$\mathscr{T}_1 = \prod_{l=1}^{L} e^{-\frac{\omega_i r_m^\alpha \varepsilon^2}{P}} \mathscr{Q}_1, \tag{3.37}$$

and

$$\mathscr{Q}_1 = \exp\left[-2\pi\rho_{\mathrm{R}}\int_0^\infty \left(1 - \frac{1}{1+\omega_n r_m^\alpha r_j^{-\alpha}}\right) r_j \mathrm{d}r_j\right]. \tag{3.38}$$

Then, an explicit expression of $\mathscr{Q}_1$ can be obtained by substituting $y = (\omega_n{}^{-1/\alpha}/r_m)r_j$ into (3.38):

$$\mathscr{Q}_1 = \exp\left(-2\pi\rho_{\mathrm{R}}\omega_n{}^{\frac{2}{\alpha}} r_m^2 \int_0^\infty \frac{y}{y^\alpha+1}\mathrm{d}y\right) = e^{-2\pi\rho_{\mathrm{R}}A(\alpha)\omega_n{}^{\frac{2}{\alpha}} r_m^2}. \tag{3.39}$$

And $\Pr\{\omega < \omega_n\}$ can be derived as follows:

$$\Pr\{\omega < \omega_n\} = 1 - e^{-2\pi A(\alpha)\omega_n^{\frac{2}{\alpha}}\rho_R r_m^2 - \frac{\omega_n r_m^\alpha \varepsilon^2}{P}}, \tag{3.40}$$

Eq. (3.32) can be obtained based on (3.40) and (3.34), and the proof has been finished.

Expected Effective Capacity of a Cluster $\mathscr{C}_T$

To evaluate the guaranteed QoS performance of F-RANs, the expected effective capacity of a specific cluster is studied. First, the hit ratio of local edge cache of F_T is defined to evaluate its utility, which is the probability that the required content objects are kept by the local edge cache of F_T:

$$\zeta_{\text{hit}}^T = \frac{\text{Number of requests served by } F_T \text{ locally}}{\text{Total Number of user requests}} = \sum_{O_k \in \mathscr{O}_T} \zeta_k, \tag{3.41}$$

where ζ_k denotes the probability that the users require content object O_k, $\mathscr{O}_T$ denotes that set of content objects that are stored by the edge cache of F_T. Similarly, the hit ratio that the required content objects are stored by the edge caches of other F-APs in the same cluster, but not the edge cache of F_T, can be expressed as

$$
\begin{aligned}
\zeta_{\text{hit}}^{\kappa} &= \frac{\text{Number of requests served by other edge caches in } \mathscr{C}_T, \text{ but not } F_T}{\text{Total Number of user requests}} \\
&= \sum_{O_k \in \mathscr{O}_\kappa / \mathscr{O}_T} \zeta_k,
\end{aligned} \tag{3.42}
$$

where $\mathscr{O}_\kappa$ denotes the set of content objects that are kept by the edge caches in $\mathscr{C}_T$.

Then, the average effective capacity of a specific cluster $\mathscr{C}_T$ with respect to O_k can be expressed as

$$
\bar{E}_k = \zeta_{\text{hit}}^{T} \bar{E}\left(\phi_k^{\text{T}}\right) + \zeta_{\text{hit}}^{\kappa} \bar{E}\left(\phi_k^{\text{C}}\right) + \zeta^{K} \bar{E}\left(\phi_k^{\text{K}}\right), \tag{3.43}
$$

where $\bar{E}\left(\phi_k^{\text{T}}\right)$ denotes the average effective capacity of O_k when it is kept by the local cache of F_T, $\bar{E}\left(\phi_k^{\text{C}}\right)$ is defined similarly when O_k is kept by other local caches in $\mathscr{C}_T$, but not the edge cache of F_T, $\bar{E}\left(\phi_k^{\text{K}}\right)$ is defined for the case when O_k is only stored by the cloud cache, and $\zeta_{\text{hit}}^{T} + \zeta_{\text{hit}}^{\kappa} + \zeta^{K} = 1$.

Based on (3.43), the average effective capacity of $\mathscr{C}_T$ can be defined as $\bar{E}_T = \sum_{k=1}^{K} \zeta_k \bar{E}_k$. Moreover, each F-AP serves only one content object, and thus all the F-APs in $\mathscr{C}_T$ can be divided into K disjoint subsets, which can be denoted as $\Phi_1, \ldots, \Phi_K$. Φ_k can be treated as a random thinning process of Φ_{R} (Haenggi 2012), and it is a homogenous PPP with a given density ρ_k, $\sum_{k=1}^{K} \rho_k = 1$. To ensure the serving performance, each user associates its nearest RRH, which can serve its required content object, to maximize the received signal power. Then, a tractable expression of average effective capacity of $\mathscr{C}_T$ is provided as follows.

Corollary *When the user accesses the nearest F-APs that allows it to access its required content object, the average effective capacity of $\mathscr{C}_T$ can be expressed as*

$$
\bar{E}_T = \zeta_{\text{hit}}^{T} \sum_{k=1}^{K} \bar{E}\left(\phi_k^{\text{T}}\right) + \zeta_{\text{hit}}^{\kappa} \sum_{k=1}^{K} \bar{E}\left(\phi_k^{\text{C}}\right) + \zeta^{K} \sum_{k=1}^{K} \bar{E}\left(\phi_k^{\text{K}}\right), \tag{3.44}
$$

where $\bar{E}\left(\phi_k^{\text{T}}\right)$, $\bar{E}\left(\phi_k^{\text{C}}\right)$, and $\bar{E}\left(\phi_k^{\text{C}}\right)$ are written as

$$
\bar{E}\left(\phi_k\right) = \zeta_k \sum_{n=1}^{N} \left[K_k\left(\omega_n\right) - K_k\left(\omega_{n+1}\right)\right]\left(1 + \bar{\omega}_n\right)^{-\eta \phi_k \bar{T}}, \ \phi_k = \phi_k^{\text{T}}, \ \phi_k^{\text{C}}, \text{ and } \phi_k^{\text{K}}, \tag{3.45}
$$

ζ_k is the popularity of content object O_k, and $\mathscr{K}_k\left(\omega_n\right)$ is defined as follows:

$$
\mathscr{K}_k\left(\omega_n\right) = 1 - 2\pi\rho_k \int_0^{\infty} r_m e^{-\left(2\pi A(\alpha)\omega_n^{\frac{2}{\alpha}}\left(\rho_{\text{R}} - \rho_k\right) + \pi\rho_k u\left(\omega_n, \alpha\right) + \pi\rho_k\right) r_m^2} e^{-\frac{\omega_n r_m^{\alpha} \varepsilon^2}{P}} \mathrm{d}r_m, \tag{3.46}
$$

where $u(\omega_n, \alpha) = \omega_n^{2/\alpha} \int_{\omega_n^{-2/\alpha}}^{\infty} \left(1 + x^{\alpha/2}\right)^{-1} \mathrm{d}x$. *In an interference limited scenario, where the impact of noise can be overlooked, an explicit expression of* $\mathscr{K}_k(\omega_n)$ *can be derived as*

$$\mathscr{K}_k(\omega_n) = 1 - \frac{1}{2A(\alpha)\,\omega_n^{2/\alpha}(l_k - 1) + u(\omega_n, \alpha) + 1}, \tag{3.47}$$

where $l_k = \rho_{\mathrm{R}}/\rho_k$.

Proof Due to (3.43), $\bar{E}(\phi_k)$ given by (3.45) can be defined as $\bar{E}(\phi_k) = \zeta_k \mathbb{E}_{r_m}\{E_{i,m}(\phi_k, r_m)\}$. In (3.45), $\mathscr{K}_k(\omega_n)$ is the outage probability of delivering O_k, which can be defined as the probability that the received SINR is lower than a given threshold ω_n. To derive a tractable expression of $K_k(\omega_n)$, we should take expectation of r_m to traverse all the possible conditions of the locations of F-APs. Based on the nearest serving F-AP association strategy, $\mathscr{T}_1$ in (3.37) can be written as

$$\mathscr{T}_1 = e^{-\frac{\omega_i r_m^{\alpha} \varepsilon^2}{P}} \mathscr{Q}_2 \prod_{k \neq k} \mathscr{Q}_1, \tag{3.48}$$

where $\mathscr{Q}_1$ follows the notation given by (3.39), and $\mathscr{Q}_2$ can be further derived as

$$\mathscr{Q}_2 = \exp\left[-2\pi\rho_{\mathrm{R}} \int_{r_j}^{\infty} \left(1 - \frac{1}{1 + \omega_n r_m^{\alpha} r_j^{-\alpha}}\right) r_j \mathrm{d}r_j\right] \overset{(a)}{=} e^{-\pi\rho_k u(\omega_n, \alpha) r_m^2}, \tag{3.49}$$

where (a) in (3.49) is derived by replacing x as $\omega_n^{-2/\alpha} r_j^2 / r_m^2$.

Next, the PDF of r_m can be written as $f(r_m) = 2\pi\rho_i r_m e^{-\pi\rho_i r_m^2}$. Recalling (3.39), (3.48), and (3.49), $\mathscr{K}_k(\omega_n)$ can be written as

$$\begin{aligned} &\mathscr{K}_k(\omega_n) \\ &\quad = \Pr\{\omega < \omega_n\} \\ &\quad = 1 - 2\pi\rho_k \textstyle\int_0^{\infty} r_m e^{-\left(2\pi A(\alpha)\omega_n^{\frac{2}{\alpha}}(\rho_{\mathrm{R}} - \rho_k) + \pi\rho_k u(\omega_n, \alpha) + \pi\rho_k\right) r_m^2} e^{-\frac{\omega_n r_m^{\alpha} \varepsilon^2}{P}} \mathrm{d}r_m. \end{aligned} \tag{3.50}$$

When we focus on an interference limited scenario, the impact of noise is overlooked, i.e., $\varepsilon^2 = 0$, and $\mathscr{K}_k(\omega_n)$ can be expressed as follows:

$$\begin{aligned} \mathscr{K}_k(\omega_n) &= 1 - 2\pi\rho_k \textstyle\int_0^{\infty} r_m e^{-\left(2\pi A(\alpha)\omega_n^{\frac{2}{\alpha}}(\rho_{\mathrm{R}} - \rho_k) + \pi\rho_k u(\omega_n, \alpha) + \pi\rho_k\right) r_m^2} \mathrm{d}r_m \\ &= 1 - \frac{1}{2A(\alpha)\omega_n^{2/\alpha}(l_k - 1) + u(\omega_n, \alpha) + 1}. \end{aligned} \tag{3.51}$$

And the proof has been finished.

Expected Effective Capacity of a Cluster $\mathscr{C}_T$

We focus on a special case when $\zeta_{\text{hit}}^{T} = \zeta_{\text{hit}}^{\kappa} = 0$, and $\bar{E}_T = \sum_{k=1}^{K} \bar{E}\left(\phi_k^{\text{C}}\right)$ is the effective capacity without edge caching in F-RANs. Since edge caching can provide better QoS experience, the QoS exponents in (3.44) satisfy the inequality that $\phi_k^{\text{T}} \leqslant \phi_k^{\text{C}} \leqslant \phi_k^{\text{K}}$. Therefore, the performance gains achieved by edge caching in F-RANs can be obtained as follows:

$$\begin{aligned}\Delta \bar{E}_T &= \bar{E}_T - \bar{E}_T\Big|_{\zeta_{\text{hit}}^{T}=\zeta_{\text{hit}}^{\kappa}=0}\\ &= \zeta_{\text{hit}}^{T}\left[\sum_{k=1}^{K}\left(\bar{E}\left(\phi_k^{\text{T}}\right) - \bar{E}\left(\phi_k^{\text{K}}\right)\right)\right] + \zeta_{\text{hit}}^{\kappa}\left[\sum_{k=1}^{K}\left(\bar{E}\left(\phi_k^{\text{C}}\right) - \bar{E}\left(\phi_k^{\text{K}}\right)\right)\right].\end{aligned} \tag{3.52}$$

Equation (3.52) indicates that the performance gains increase linearly with respect to the hit ratio of edge caching ζ_{hit}^{T} and $\zeta_{\text{hit}}^{\kappa}$. In particular, if the edge cache of F_T is large enough to keep all the required content objects, i.e., $\zeta_{\text{hit}}^{T} = 1$, it can achieve the theoretical performance limit. However, the volume of required content objects is so large that cannot be kept by a single edge cache. To further improve the utility of edge cache, an applicable method is to encourage intra-cluster coordinations, which can reduce the latency of content objects by avoiding accessing them from data center via fronthaul links.

Moreover, the effective capacity can also be improved by optimizing the networking and radio resource management strategies. For example, to guarantee the QoS of popular content objects, more F-APs should be used for their delivery, and the radio resource units with better channel conditions should be allocated.

3.2.3 Simulation Results

The simulation setting follows that in Zhao et al. (2016). In particular, the block fading channel model is implemented, and the length of channel coherence duration time is $T = 1$ ms. The locations of network nodes follow homogenous PPP models, and the path loss exponent is $\alpha = 4$.

The effective capacity of a typical user is provided in Fig. 3.4. It shows that the numerical results based on the theoretical results coincide with the Monte Carlo results, and thus the validity of our derivations can be demonstrated. Moreover, as the QoS exponent increases, which means that the requirement of the QoS guarantee is strict, the effective capacity decreases. Finally, the simulation results show that the performance of effective capacity is decided by the density of F-APs. In particular, the effective capacity can be improved by reducing the density of interfere F-APs, which can mitigate the impact of co-channel interference.

Figure 3.5 plots the relationship between the average effective capacity and the data rate of fronthaul links, where the QoS exponents are set as $\phi_k^{\text{T}} = 0.05$,

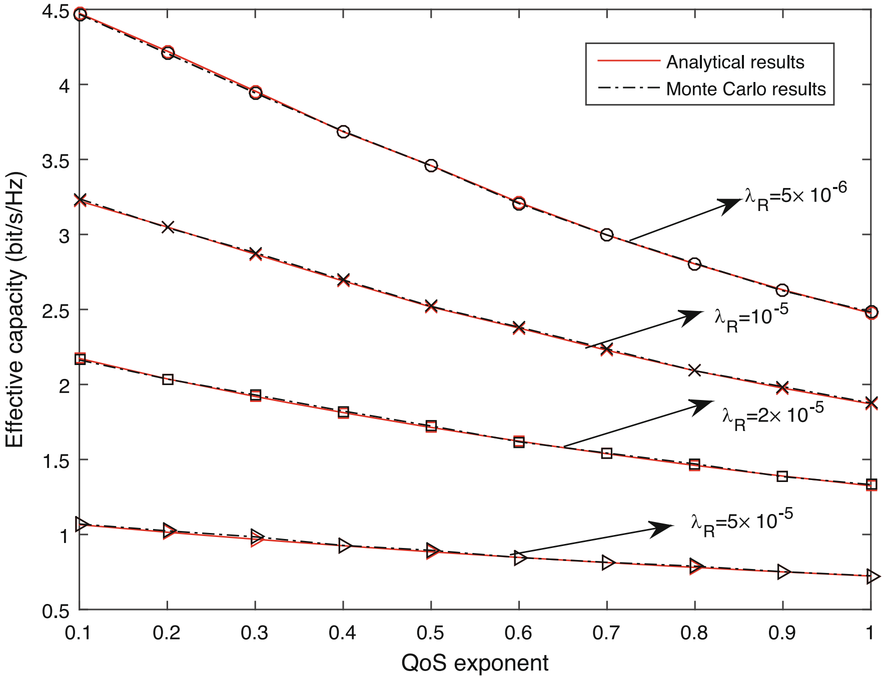


Fig. 3.4 Effective capacity of a typical user ($r_m = 50$ m)

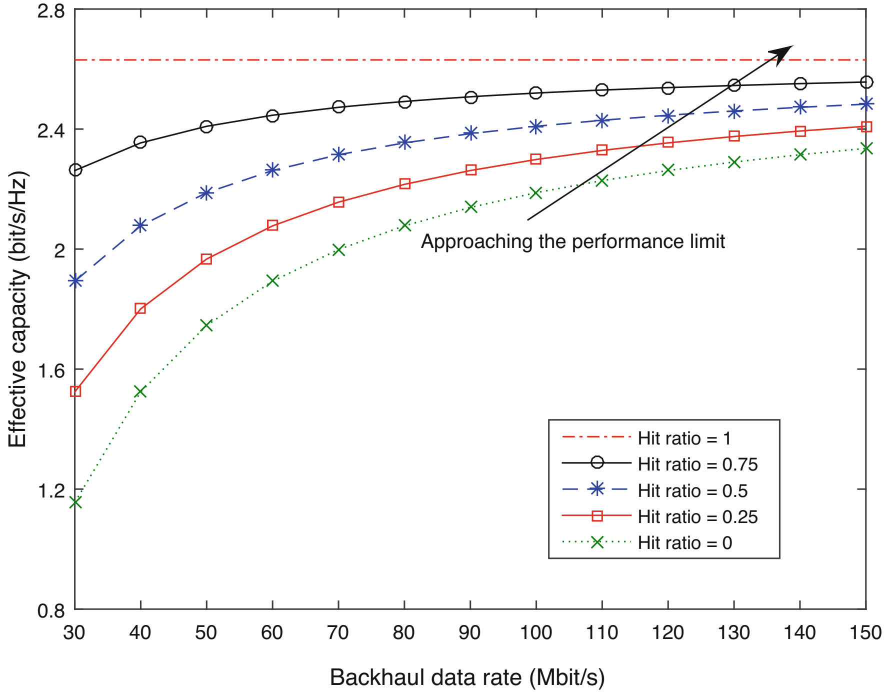


Fig. 3.5 Average effective capacity vs. backhaul data rate

$\phi_k^{\mathrm{C}} = 0.15$, and $\phi_k^{\mathrm{K}} = 0.35$, respectively. Moreover, the size of each content object is set as $B_K = 10$ Mbits. The simulation results show that the improvement of fronthaul data rate can support better QoS experience, and thus the average effective capacity increases. Moreover, when the hit ratio is enlarged, the performance of effective capacity is improved by responding more user requests locally. Finally, the simulation results indicate that it approaches the theoretical limit as the data rate of fronthaul and the size of edge caches can be enlarged, and thus the performance gains of edge caching in F-RANs can be verified.

3.3 Summary

In this chapter, the theoretic performance of F-RANs is studied. The transmission capacity and the QoS are analyzed to evaluate the performance of edge signal processing and edge caching, respectively. Both the analytical and numerical results show that the performance of F-RANs can be improved by encouraging edge processing, which can mitigate the loadings of fronthaul and reduce the cost and latency caused by the conventional centralized processing schemes.

References

Andrews J, et al (2011) A tractable approach to coverage and rate in cellular networks. IEEE Trans Commun 59:3122–3134

Haenggi M (2012) Stochastic geometry for wireless networks. Cambridge University Press, Cambridge

Wu D, et al (2003) Effective capacity: a wireless link model for support of quality of service. IEEE Trans Wireless Commun 2:630–643

Zhao Z, et al (2016) Cluster content caching: an energy-efficient approach to improve quality of service in cloud radio access networks. IEEE J Sel Areas Commun 34:1207–1221

Chapter 4
Cooperative Signal Processing in Fog Radio Access Networks

In Sect. 4.1, cooperative NOMA is studied to achieve higher diversity gain for all users. In Sect. 4.2, the cooperation between RRHs and the macrocell is proposed as an effective way to mitigate inter-tier interference.

4.1 Cooperative Non-orthogonal Multiple Access in F-RANs

4.1.1 Background

In non-orthogonal multiple access (NOMA) scenarios, multiple users can transmit at the same time, code, and frequency but with different power levels (Saito et al. 2013), and users with better channel conditions can acquire information of other users with the help of successive interference cancellation (Cover et al. 1991). Actually, this can be taken as useful prior information to boost system performance, which however is not exploited in Choi (2014) and Ding et al. (2014). In this chapter, prior information in a downlink NOMA system is fully utilized, which is the basis of cooperative NOMA transmission. Specifically, the transmission scheme is composed of two phases, namely downlink transmission phase and cooperative phase. In the latter, users with better connections to the base station are taken as relays that broadcast the superposition of decoded messages to other users, which facilitates enhancing the reception reliability. By analyzing the outage probability and diversity order achieved by this cooperative NOMA scheme, it can be found that cooperative NOMA can achieve the maximum diversity gain for all users.

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0_4

4.1.2 System Model

A downlink NOMA transmission scenario is considered, where there is one base station and K users. In the following, the cooperative NOMA scheme will be elaborated, which includes two phases, namely direct transmission phase and cooperative phase.

Downlink NOMA Transmission Phase

In this phase, the superposition of K messages, i.e., $\sum_{m=1}^{K} p_m s_m$, is sent to users by the base station according to the NOMA principle, in which s_m and p_m are the message and power coefficient of the m-th user, respectively. For the k-th user, it observes $y_{1,k} = \sum_{m=1}^{K} h_k p_m s_m + n_k$, with h_k being the channel coefficient between it and the base station. We use n_k to denote additive Gaussian noise. It is assumed that the users are ordered according to channel quality, i.e.,

$$|h_1|^2 \leq \cdots \leq |h_K|^2. \tag{4.1}$$

The use of NOMA implies $|p_1|^2 \geq \cdots \geq |p_K|^2$, with $\sum_{m=1}^{K} p_m^2 = 1$. When NOMA transmission is finished, K-th user employs successive detection and we have

$$SINR_{K,k} = \frac{|h_K|^2 |p_k|^2}{\sum_{m=k+1}^{K} \left|h_K^H p_m\right|^2 + \frac{1}{\rho}}, \tag{4.2}$$

where ρ is the transmit SNR. When the messages of all the other users are successfully decoded, the K-th user then decodes its own information and the SNR is given by $SNR_{K,K} = \rho |h_K|^2 |p_K|^2$. Note that only if $\log\left(1 + SINR_{K,k}\right) > R_k, \quad \forall 1 \leq k \leq K$, can the K-th user decode its own information, and R_k is the target data rate of user k.

Cooperative Phase

This phase contains $(K-1)$ time slots. During the first time slot, user K broadcasts $\sum_{m=1}^{K-1} q_{K,m} s_m$ with $\sum_{m=1}^{K-1} q_{K,m}^2 = 1$. The received signal of k-th user is given by

$$y_{2,k} = \sum_{m=1}^{K-1} g_{K,k} q_{K,m} s_m + n_{2,k}, \tag{4.3}$$

where $k < K$, $g_{K,k}$ is the channel coefficient between the K-th user and the k-th user.

With maximum ratio combining, the received signal from both phases are combined at the $(K-1)$-th user. When decoding the message of user k $(k<(K-1))$, the corresponding SINR is expressed as

$$SINR_{K-1,k} = \frac{|h_{K-1}|^2 p_k^2}{|h_{K-1}|^2 \sum_{m=k+1}^{K} p_m^2 + \frac{1}{\rho}} + \frac{\left|g_{K,K-1}\right|^2 q_{K,k}^2}{\left|g_{K,K-1}\right|^2 \sum_{m=k+1}^{K-1} q_{K,m}^2 + \frac{1}{\rho}}. \tag{4.4}$$

Once the other users' messages are decoded, the $(K-1)$-th user is able to decode its own information and the SINR is given by

$$SINR_{K-1,K-1} = \frac{|h_{K-1}|^2 p_{K-1}^2}{|h_{K-1}|^2 p_K^2 + \frac{1}{\rho}} + \left|g_{K,K-1}\right|^2 q_{K,K-1}^2. \tag{4.5}$$

For a general n-th time slot with $1 \le n \le (K-1)$, the combination of the $(K-n)$ messages are broadcasted by the $(K-n+1)$-th user and the k-th user satisfying $k<(K-n+1)$ receives

$$y_{2,k} = \sum_{m=1}^{K-n} g_{K-n+1,k}^{H} q_{K-n+1,m} s_m + n_{n+1,k}. \tag{4.6}$$

With the received signals in both phases, user $(K-n)$ is able to decode the message of user k with $1 \le k < (K-n)$ with the following SINR:

$$\begin{aligned} SINR_{K-n,k} = & \frac{|h_{K-n}|^2 p_k^2}{|h_{K-n}|^2 \sum_{m=k+1}^{K} p_m^2 + \frac{1}{\rho}} \\ & + \sum_{i=1}^{n} \frac{\left|g_{K-i+1,K-n}\right|^2 q_{K-i+1,k}^2}{\left|g_{K-i+1,K-n}\right|^2 \sum_{m=k+1}^{K-i} q_{K-i+1,m}^2 + \frac{1}{\rho}}. \end{aligned} \tag{4.7}$$

Moreover, it can decode its own information with the following SINR:

$$\begin{aligned} SINR_{K-n,K-n} = & \frac{|h_{K-n}|^2 p_{K-n}^2}{|h_{K-n}|^2 \sum_{m=K-n+1}^{K} p_m^2 + \frac{1}{\rho}} \\ & + \sum_{i=1}^{n-1} \frac{\left|g_{K-i+1,K-n}\right|^2 q_{K-i+1,K-n}^2}{\left|g_{K-i+1,K-n}\right|^2 \sum_{m=K-n+1}^{K-i} q_{K-i+1,m}^2 + \frac{1}{\rho}} \\ & + \rho \left|g_{K-n+1,K-n}\right|^2 q_{K-n+1,K-n}^2. \end{aligned} \tag{4.8}$$

Note that when there is no cooperative phase, the SINR of user $K-n$ is calculated as $\frac{|h_{K-n}|^2 p_{K-n}^2}{|h_{K-n}|^2 \sum_{m=K-n+1}^{K} p_m^2 + \frac{1}{\rho}}$.

4.1.3 Performance Analysis

When reliable detection can be achieved by the $(n-1)$ users with best channel conditions, the outage probability of user $(K-n)$ is given by

$$\mathrm{P}_o^{K-n} = \mathrm{P}\left(\mathrm{SINR}_{K-n,k} < \epsilon_k, \forall k \in \{1, \cdots, K-n\}\right), \tag{4.9}$$

in which $\epsilon_k = 2^{R_k} - 1$.

For notational simplicity, define $a_{k,i}^{K-n}=q^2_{K-i+1,k}$ and $b_{k,i}^{K-n}=\sum_{m=k+1}^{K-i} q^2_{K-i+1,m}$, with $1 \le k \le (K-n)$ and $1 \le i \le n$. Two special case are $a_{K-n,n}^{K-n} = q^2_{K-n+1,K-n}$ and $b_{K-n,n}^{K-n} = 0$. Moreover, for $1 \le k \le (K-n)$, define $a_{k,0}^{K-n} = p_k^2$ and $b_{k,0}^{K-n} = \sum_{m=k+1}^{K} p_m^2$. In the below proposition, the analysis result about the diversity order of the proposed NOMA scheme is elaborated.

Proposition 4.1 *Under the condition that reliable detection can be conducted at the* $(n-1)$ *users with best channel conditions. Then, the proposed two-phase NOMA scheme ensures the* $(K-n)$*-th ordered user achieves a diversity order of K if* $\epsilon_k < \frac{a_{k,i}^{K-n}}{b_{k,i}^{K-n}}$*, for* $1 \le k \le (K-n)$ *and* $0 \le i \le n$.

Proof For notational simplicity, define $z_{k,i}^{K-n}=\frac{|g_{K-i+1,K-n}|^2 q^2_{K-i+1,k}}{|g_{K-i+1,K-n}|^2 \sum_{m=k+1}^{K-l} q^2_{K-i+1,m} + \frac{1}{\rho}}$, where $1 \le k \le (K-n)$ and $1 \le i \le n$, except $z_{K-n,n}^{K-n}=\rho\left|g_{K-n+1,K-n}\right|^2 q^2_{K-n+1,K-n}$. In addition, define $z_{k,0}^{K-n} = \frac{|h_{K-n}|^2 p_k^2}{|h_{K-n}|^2 \sum_{m=k+1}^{K} p_m^2 + \frac{1}{\rho}}$. For $1 \le k \le (K-n)$, the SINRs can be expressed as follows:

$$SINR_{K-n,k} = z_{k,0}^{K-n} + \sum_{i=1}^{n} z_{k,i}^{K-n}. \tag{4.10}$$

Then, the outage probability is

$$\begin{aligned} \mathrm{P}_o^{K-n} &= \mathrm{P}\left(z_{k,0}^{K-n} + \sum_{i=1}^{n} z_{k,i}^{K-n} < \epsilon_k, \forall k \in \{1, \cdots, K-n\}\right) \\ &\le \sum_{k=1}^{K-n} \mathrm{P}\left(z_{k,0}^{K-n} + \sum_{i=1}^{n} z_{k,i}^{K-n} < \epsilon_k\right). \end{aligned} \tag{4.11}$$

Considering that $\mathrm{P}(a+b<c) \leq \mathrm{P}(a<c)+\mathrm{P}(b<c)$ and channel gains are independent, we have

$$\mathrm{P}_o^{K-n} \leq \sum_{k=1}^{K-n} \prod_{i=0}^{n} \mathrm{P}\left(z_{k,i}^{K-n} < \epsilon_k\right). \tag{4.12}$$

Note that all the elements in (4.10) have the same structure given by

$$z_{k,i}^{K-n} = \frac{a_{k,i}^{K-n} x}{b_{k,i}^{K-n} x + \frac{1}{\rho}}. \tag{4.13}$$

Further, under the assumption that x follows exponential distribution, the cumulative distribution function (CDF) of $z_{k,i}^{K-n}$ is given by

$$\mathrm{P}_{z_{k,i}^{K-n}}(Z < z) = \begin{cases} 1, & \text{if } z \geq \frac{a_{k,i}^{K-n}}{b_{k,i}^{K-n}} \\ 1 - e^{-\frac{k-z}{\rho\left(a_{k,i}^{K-n} - b_{k,i}^{K-n} z\right)}}, & \text{otherwise.} \end{cases} \tag{4.14}$$

When SNR is high, $\frac{\epsilon_k}{\rho\left(a_{k,i}^{K-n} - b_{k,i}^{K-n} z\right)}$ approaches 0. At this time, the probability that $z_{k,i}^{K-n} < \epsilon_k$ holds can be approximated using the power series of exponential functions (Cover et al. 1991) as follows:

$$\mathrm{P}_{z_{k,i}^{K-n}}\left(Z < \epsilon_k\right) = 1 - e^{-\frac{\epsilon_k}{\rho\left(a_{k,i}^{K-n} - b_{k,i}^{K-n} e\right)}} \approx \frac{\epsilon_k}{\rho a_{k,i}^{K-n}}, \tag{4.15}$$

and it requires $\epsilon_k < \frac{a_{k,i}^{K-n}}{b_{k,i}^{K-n}}$.

As for the distribution functions of $z_{k,0}^{K-n}$, it can be derived as follows based on order statistics (David et al. 2003).

$$\mathrm{P}_{z_{k,0}^{K-n}}(Z < z) = \begin{cases} 1, & \text{if } z \geq \frac{a_{k,0}^{K-n}}{b_{k,0}^{k-n}} \\ \int_0^{\frac{\frac{z}{\rho}}{a_{k,0}^{K-n} - b_{k,0}^{K-n} z}} \frac{e^{-x}}{(K-n-1)!} x^{K-n-1} dx, & \text{otherwise.} \end{cases} \tag{4.16}$$

Using high SNR approximation and referring to Jeffrey et al. (2007), the probability of $\mathrm{P}\left(z_{k,0}^{K-n} < \epsilon_k\right)$ can be approximated with the help of the power series of exponential functions as follows:

$$\begin{aligned}\mathrm{P}\left(z_{k,0}^{K-n} < \epsilon_k\right) &= \int_0^{\frac{\epsilon_k}{\rho\left(a_{k,i}^{K-n} - b_{k,i}^{K-n}\epsilon_k\right)}} \frac{x^{K-n-1}e^{-x}}{(K-n-1)!}dx \\ &\approx \frac{\epsilon_k^{K-n}}{(K-n)!\left(a_{k,i}^{K-n}\right)^{K-n}\rho^{K-n}},\end{aligned} \tag{4.17}$$

which is conditioned on $\epsilon_k < \frac{a_{k,0}^{K-n}}{b_{k,0}^{K-n}}$. Similarly the probability for the event $z_{K-n,n}^{K-n} < \epsilon_k$ can be approximated as

$$\mathrm{P}\left(z_{K-n,n}^{K-n} < \epsilon_k\right) \approx \frac{\epsilon_k}{q_{K-n+1,K-n}^2\rho}, \tag{4.18}$$

since $z_{K-n,n}^{K-n}$ is actually a special case of (4.13).

Combining the above derivations, the diversity order that can be achieved by the cooperative NOMA scheme can be obtained.

Based on Proposition 4.1 and the assumption on the independence among the channels, we have the following lemma.

Lemma 4.1 *If $\epsilon_k < \frac{a_{k,i}^{K-n}}{b_{k,i}^{K-n}}$ for $1 \le k \le (K-n)$ and $0 \le i \le n$, user n is guaranteed to achieve a diversity order of K with the proposed cooperative NOMA. Generally speaking, through the cooperation among users, the cooperative NOMA is capable of providing all the users with a diversity order of K, which is irrelevant to the channel conditions. However, for traditional non-cooperative NOMA, a diversity order of just n can be achievable for user n.*

Further, to reduce system complexity, user pairing is essential, which identifies which users are grouped together to perform cooperative NOMA. Without loss of generality, selecting only two users is considered. It is assumed that users are ordered in the same way as before and users m and n $(m<n)$ are paired. With time division multiple access (TDMA), data rates for both users are given by $\bar{R}_i = \frac{1}{2}\log\left(1 + \rho\left|h_i\right|^2\right), i \in \{m, n\}$. Considering the complication of date rate expressions with cooperative NOMA, conventional NOMA is considered with data rates $R_m = \log\left(1 + \frac{\rho|h_m|^2 p_m^2}{\rho|h_m|^2 p_n^2+1}\right)$ and $R_n = \log\left(1 + \rho p_n^2\left|h_n\right|^2\right)$. Note that R_n is achievable since $\log\left(1 + \frac{|h_n|^2 p_m^2}{|h_n|^2 p_n^2+1}\right) \ge R_m$.

When SNR is high, the difference between the sum rate under TDMA and that under conventional NOMA is

$$\begin{aligned} & R_m + R_n - \bar{R}_m - \bar{R}_n \\ & \approx \log\left(1 + \frac{p_m^2}{p_n^2}\right) + \log \rho p_n^2 \left|h_n\right|^2 - \frac{\log \rho \left|h_m\right|^2}{2} - \frac{\log \rho \left|h_n\right|^2}{2} \\ & = \frac{\log \left|h_n\right|^2}{2} - \frac{\log \left|h_m\right|^2}{2}. \end{aligned} \tag{4.19}$$

It can be seen that the gap is influenced by the degree of differentiation between the two users' channels. Therefore, it is the best to cluster two users with the experienced channel fading considerably different into a single group. This observation holds for cooperative NOMA as well. In particular, for the m-th user, although R_m can be as large as $\log\left(1 + \frac{\rho|h_m|^2 p_m^2}{\rho|h_m|^2 p_n^2+1} + \rho\left|g_{n,m}\right|^2\right)$, the data rate for the m-th user is bounded as $R_m \leq \log\left(1 + \frac{\rho|h_n|^2 p_m^2}{\rho|h_n|^2 p_n^2+1}\right)$, since the n-th user needs to decode the m-th user's information. Because of $\log\left(1 + \frac{\rho|h_n|^2 p_m^2}{\rho|h_n|^2 p_n^2+1}\right) \approx \log\left(1 + \frac{p_m^2}{p_n^2}\right)$, the conclusion obtained for conventional NOMA also fits into cooperative NOMA. Further details about user pairing for non-cooperative NOMA can be found in Ding et al. (2016).

4.1.4 Performance Analysis

Figure 4.1 shows the comparison results of three schemes, namely orthogonal MA, non-cooperative NOMA, and cooperative NOMA, in terms of outage probabilities as functions of SNR. $K=2$ and $p_1^2 = \frac{4}{5}$. It can seen that the cooperative NOMA scheme outperforms the other two schemes, using which the maximum diversity gain is achievable by all the users.

In Fig. 4.2, outage capacities under three schemes are evaluated with the setting of $R_1 = R_2$. Under each targeted data rate, the corresponding value of the vertical axis is calculated as one minus the outage probability. Particularly, when the outage probability is 10% and the transmit SNR is 15 dB, the cooperative NOMA scheme significantly outperforms the other two schemes in terms of bits per channel use (BPCU).

In Fig. 4.3, outage probabilities with cooperative NOMA are presented, where local short-range communications are not exploited. Specifically, when local short-range communication is unavailable, cooperative MA schemes can be revised as follows. Consider a scenario with only two users whose channels are ordered as before. At this time, the first two time slots are allocated to the two users, respectively, and one additional time slot is consumed for user cooperation. Thus, three time slots are needed. During the first two time slots, the base station serves the two users simultaneously and User 2 helps User 1 during the third time slot. Assume that the relay uses a codebook independent of the one at the source. By applying (Cover et al. 1991, Theorem 15.7 .2), the achievable rate at User 1 is

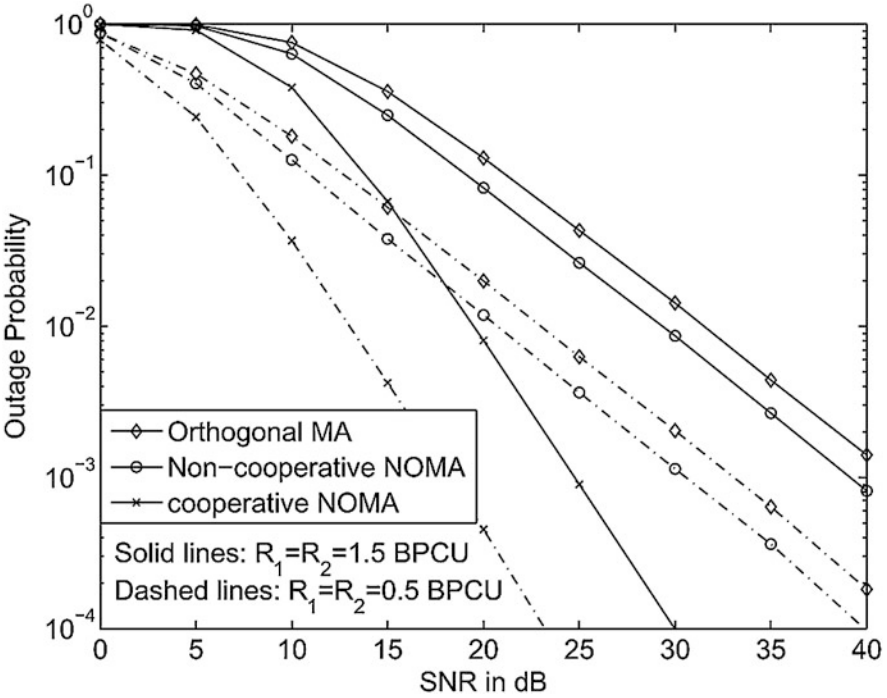


Fig. 4.1 Outage probability with cooperative NOMA

$\frac{1}{3}\left(2\log\left(1+SINR_{1,1}\right)+\log\left(1+\rho\left|g_{2,1}\right|^2\right)\right)$ and the achievable rate at User 2 is $\frac{2}{3}\log\left(1+SINR_{2,2}\right)$, if User 1's message is correctly decoded at User 2. Based on the results in Fig. 4.3, it can be found that cooperative NOMA is still better than the baselines.

Figure 4.4 shows how user pairing affects the performance of cooperative and non-cooperative NOMA. Assume one of the two paired users is the user with the best channel condition. When pairing this user with the user that has the worst channel, sum rate gain achieves 1.5 BPCU at 10 dB, while the gain decreases to 0.2 BPCU if the other user is the one with the second best channel.

4.2 Contract Based Cooperative Signal Processing

4.2.1 Background

Owing to the centralized cooperative signal processing in the base band unit (BBU) pool, inter-RRH interference can be completely avoided in F-RANs. However, it is still challenging to solve the inter-tier interference between RRHs and macro base

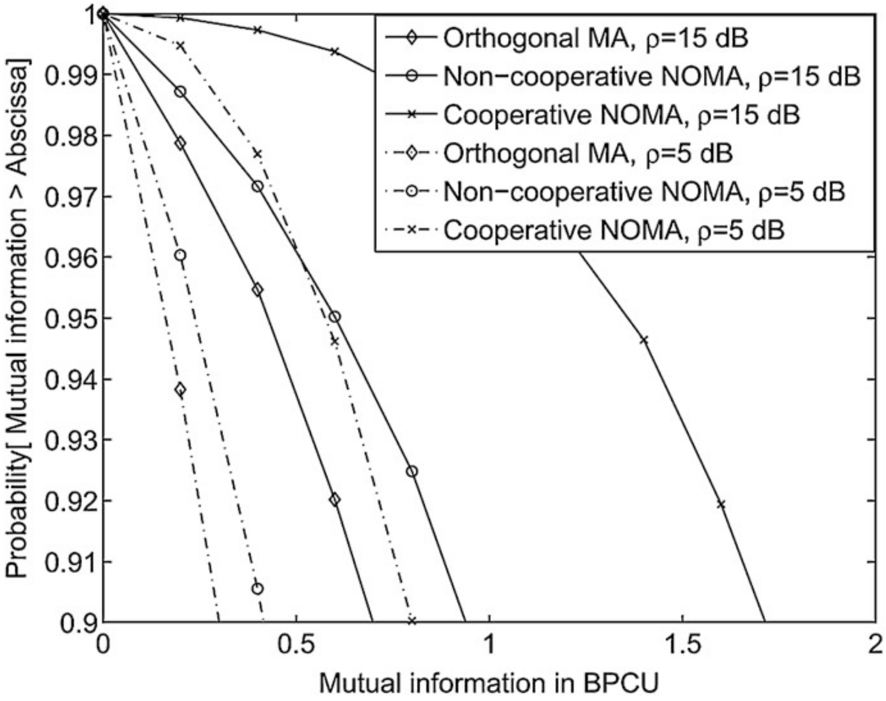


Fig. 4.2 Outage capacity achieved by cooperative NOMA

stations (MBSs). On the other hand, game theory has been considered as a promising model to handle the inter-tier interference problem. In Kang et al. (2012), authors jointly optimize macrocell and femtocell objectives by a Stackelberg game under the constraint of limiting inter-tier interference to the MBS. In Han et al. (2014), the authors develop the framework of hierarchical game to investigate the uplink power allocation to mitigate the inter-tier interference in two-tier femtocell networks. Inspired by the existing works, this chapter goes a further step to handle the inter-tier interference in F-RANs via cooperative signal processing based on advanced contract theory, which is suitable in situations with information asymmetry.

4.2.2 System Model and Cooperative Signal Processing Framework

In the considered downlink scenario, there are one BBU pool, one MBS and K RRHs, and the BBU pool connects to each RRH and the MBS through an ideal fronthaul link and a backhaul link, respectively. All the RRHs operate on the same radio resource with the MBS, causing inter-tier interference. RRHs jointly serve M

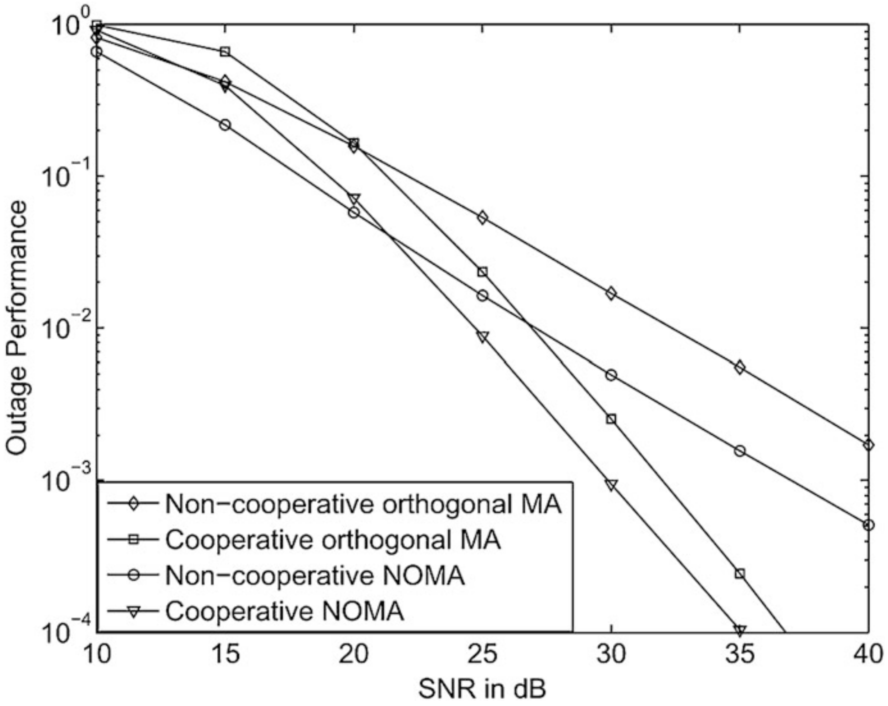


Fig. 4.3 Outage probability achieved by cooperative NOMA without using local short-range communications. R1 = 1.2 BPCU and R2 = 1.9 BPCU

RUEs with $M < K$ via zero forcing precoding while the MBS serves one MUE. In addition, the MBS, the MUE, each RRH, and each RUE all equip with a single antenna. Define the coefficient of the channel from RRH k to RUE m as g_{mk} whose matrix is denoted as $\mathbf{G}$ with g_{mk} being the (m, k)-th entry, and denote the coefficient of the channel from the MBS to the MUE as g_B. In addition, define the coefficient of the channel from RRH k to the MUE as f_{kM}, and f_{Bm} denotes the coefficient of the channel from the MBS to RUE m. The symbol delivered by RRHs to RUE m and from the MBS to the MUE are denoted as s_m and s_{M+1}, respectively. Let $\mathbf{A}$ of $K \times M$ denote the precoding matrix.

The received signal at all RUEs is given by

$$\mathbf{y}_R = \mathbf{GAs} + \underbrace{\mathbf{f}_B s_{M+1}}_{\text{interference}} + \mathbf{n}_C, \tag{4.20}$$

where $\mathbf{n}_C$ is an AWGN vector of $M \times 1$ with the variance of each entry being σ_n^2, $\mathbf{f}_B = [f_{B1}, f_{B2}, \ldots, f_{BM}]^T$, and $\mathbf{s} = [s_1, s_2, \ldots, s_M]^T$. In addition, we use $\mathbf{G} = \mathbf{D}^{\frac{1}{2}}\mathbf{H}$ to model the channel fading, where $\mathbf{H}$ of $M \times K$ is the matrix made up of fast fading coefficients with i.i.d. zero-mean complex Gaussian with unity variance and

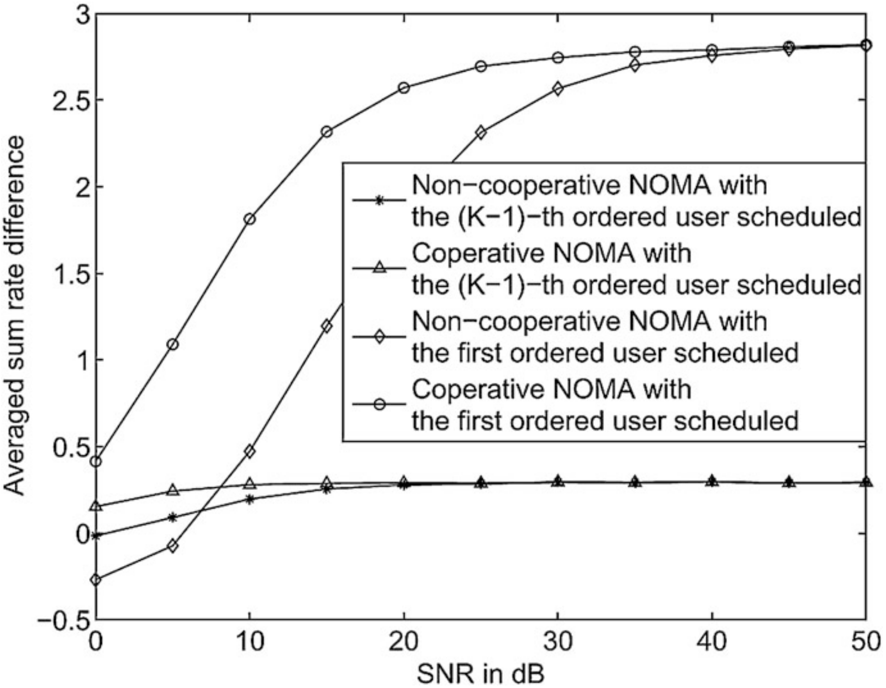


Fig. 4.4 The impact of user pairing on the sum rate. $K = 10$

$\mathbf{D}^{\frac{1}{2}}$ of $M \times M$ is a diagonal matrix with $[\mathbf{D}]_{mm} = \upsilon_m$. Moreover, the m-th element in $\mathbf{f}_B$ is with the variance υ_{Bm}.

For the MUE, its observed signal is

$$y_B = g_B s_{M+1} + \underbrace{\mathbf{f}_M \mathbf{A}\mathbf{s}}_{\text{interference}} + n_B, \tag{4.21}$$

where n_B is AWGN with variance σ_n^2 and $\mathbf{f}_M = [f_{1M}, f_{2M}, \ldots, f_{KM}]$ whose elements are with variance υ_{M+1}.

In order to mitigate the mutual interference between RRH transmission and MBS transmission, the following transmission schemes are introduced. In the first scheme, named as *RRH-alone with UEs-all*, all K RRHs serve M RUEs and the single MUE, while the MBS keeps idle. Since $K \geq M + 1$, cooperative signal processing is performed in the BBU pool. In the second scheme, named as *RRH-alone with RUEs-only*, only transmission from all the RRHs to all the RUEs occurs. In the last scheme, named as *RRH-MBS with UEs-separated*, RRHs and the MBS serve their own UEs simultaneously and their transmission power is determined by fairness power control. Based on these schemes, an interference coordination framework is proposed, which harmonizes all the above three schemes in the time domain. Specifically, each time transmission interval (TTI) of time length T_0 is partitioned into three phases.

- **Phase I** : This phase lasts from 0 to t_1 with the first scheme employed.
- **Phase II** : This phase lasts from t_1 to t_1+t_2 with the second scheme employed.
- **Phase III** : This phase lasts from $T_0-t_1-t_2$ to T_0 with the last scheme employed.

To facilitate the cooperation between the BBU pool and the MBS and fully reap the benefits of cooperative signal processing at the BBU pool, contract theory is utilized in the following to result in a win-to-win situation.

4.2.3 Optimal Contract Design Under Complete CSIs

Rate-Based Utility Definition

- **Phase I**: Let P_{C1} and P_{M1} denote the received symbol power of s_m and s_{M+1}, respectively. Then RUE sum rate and the MUE data rate are calculated as

$$R_{C1}=M\log\left(1+\frac{P_{C1}}{\sigma_n^2}\right), \tag{4.22}$$

$$R_{M1}=\log\left(1+\frac{P_{M1}}{\sigma_n^2}\right), \tag{4.23}$$

respectively. In addition, the total transmit power of RRHs is constrained by

$$\mathscr{E}\left\{\mathrm{tr}\left(\mathbf{F}\mathbf{F}^H\right)^{-1}\mathbf{\Lambda}\right\}\leq P_{\max}, \tag{4.24}$$

where $\mathbf{F}=[\mathbf{G}^T,\mathbf{f}_M^T]^T$, $\mathbf{\Lambda}$ of $(M+1)\times(M+1)$ is a diagonal matrix whose $(M+1)$-th diagonal element is P_{M1} with the remaining diagonal elements being P_{C1}, and $P_{\max}$ is a pre-defined allowable maximum transmit power. We re-written the transmit power constraint as

$$\mathrm{tr}\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathscr{E}\left\{(\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H)^{-1}\right\}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{\Lambda}\right)\leq P_{\max}, \tag{4.25}$$

where $\tilde{\mathbf{D}}$ is a $(M+1)\times(M+1)$ diagonal matrix with $\tilde{\mathbf{H}}=[\mathbf{H}^T,\mathbf{h}_f^T]^T$. Equation (4.25) can be given by

$$\sum_{m=1}^{M}\frac{P_{C1}}{(K-M-1)\upsilon_m}+\frac{P_{M1}}{(K-M-1)\upsilon_{M+1}}\leq P_{\max}. \tag{4.26}$$

- **Phase II**: Assume that all the RUEs possess the same received symbol power P_{C2} and then RUE sum rate is calculated by

$$R_{C2}=M\log\left(1+\frac{P_{C2}}{\sigma_n^2}\right). \tag{4.27}$$

Similarly, we can write down the following constraint for the long-term average transmit power.

$$\mathscr{E}\left\{P_{C2}\mathrm{tr}\left(\mathbf{G}\mathbf{G}^H\right)^{-1}\right\}=\sum_{m=1}^{M}\frac{P_{C2}}{\upsilon_m\left(K-M\right)}\leq P_{\max}. \tag{4.28}$$

To achieve maximal sum rate, $P_{C2}=\frac{(K-M)P_{\max}}{\varepsilon_1}$, where $\varepsilon_1=\sum_{m=1}^{M}\frac{1}{\upsilon_m}$.

- **Phase III**: Under fairness power control, the transmit power of the MBS, P_B, and the transmit power of the RRH, P_{C3}, are calculated as

$$\{P_{C3},\,P_B\}=\arg\max_{P_{C3},P_B}\min\left\{R_{C3},\,R_{M3}\right\}, \tag{4.29}$$

$$s.t.\quad 0\leq P_{C3}\leq P_{\max},\ 0\leq P_B\leq P_{\max}. \tag{4.30}$$

The sum rate of all RUEs, R_{C3}, and the data rate, R_{M3}, are as follows:

$$R_{C3}=\sum_{m=1}^{M}\log\left(1+\frac{P_{C3}}{|f_{Bm}|^2P_B+\sigma_n^2}\right), \tag{4.31}$$

$$R_{M3}=\log\left(1+\frac{|g_B|^2P_B}{P_{C3}\mathbf{f}_M\mathbf{G}^H\left(\mathbf{G}\mathbf{G}^H\right)^{-2}\mathbf{G}\mathbf{f}_M^H+\sigma_n^2}\right). \tag{4.32}$$

Based on the above derivation, we can define the rate-based utility for the BBU pool as follows:

$$U_C=t_1R_{C1}+\left(T_0-t_1-t_3\right)R_{C2}-\left(T_0-t_3\right)R_{C3}. \tag{4.33}$$

The utility of the MBS is calculated as

$$U_M=t_1R_{M1}+t_3R_{M3}. \tag{4.34}$$

Contract Design Under Perfect CSIs

To design a feasible contract that identifies the time length of each phase, the BBU pool should guarantee that the MBS's utility is at least as good as the utility achieved when there is no interference coordination. This constraint is called individual rational (IR), whose definition is given as follows.

(*Individual Rational*) When the MBS's utility is higher than its reserved utility u, which is

$$t_1 R_{M1} + t_3 R_{M3} \geq u = T_0 R_{M3}, \tag{4.35}$$

the corresponding contract is individual rational, whose item is $(t_1, t_3, P_{M1}, P_{C1})$.

The optimal contract $\left(t_1^*, t_3^*, P_{M1}^*, P_{C1}^*\right)$ can be obtained by solving

$$\left(t_1^*, t_3^*, P_{M1}^*, P_{C1}^*\right) = \arg \max_{t_1, t_3, P_{M1}, P_{C1}} \left\{ t_1 R_{C1} + (T_0 - t_1 - t_3)\, R_{C2} - (T_0 - t_3)\, R_{C3} \right\}, \tag{4.36}$$

$$s.t. \quad t_1 R_{M1} + t_3 R_{M3} \geq T_0 R_{M3}, \tag{4.37}$$

$$t_3 \geq 0, t_1 \geq 0, T_0 - t_1 - t_3 \geq 0, \tag{4.38}$$

$$\frac{P_{C1}}{(K-M-1)\,\varepsilon_1} + \frac{P_{M1}}{(K-M-1)\,\upsilon_{M+1}} \leq P_{\max}. \tag{4.39}$$

Using Karush–Kuhn–Tucker conditions, the optimal contract $\left(t_1^*, t_3^*, P_{M1}^*, P_{C1}^*\right)$ satisfies the following conditions:

$$t_1^* R_{M1}^* - \left(T_0 - t_3^*\right) R_{M3}^* = 0, \tag{4.40}$$

$$t_3^* = 0, \tag{4.41}$$

$$\frac{P_{C1}^*}{(K-M-1)\,\varepsilon_1} + \frac{P_{M1}^*}{(K-M-1)\,\upsilon_{M+1}} = P_{\max}. \tag{4.42}$$

Then the optimization problem can be transformed to

$$\max_{P_{M1}} \quad \frac{T_0 R_{M3}\left[M\log\left(1+\frac{(K-M-1)P_{\max}}{\varepsilon_1\sigma_n^2}-\frac{P_{M1}}{\varepsilon_1\sigma_n^2\upsilon_{M+1}}\right)-R_{C2}\right]}{\log\left(1+\frac{P_{M1}}{\sigma_n^2}\right)}+T_0(R_{C2}-R_{C3}) \tag{4.43}$$

$$s.t. \quad 0\le P_{M1}\le(K-M-1)\,\upsilon_{M+1}P_{\max}. \tag{4.44}$$

The objective function is only with respect to P_{M1}. Once P_{M1}^* is obtained, t_1^* and P_{C1}^* can be directly determined.

4.2.4 Contract Design Under Practical Channel Estimation

Channel Estimation

To acquire the downlink channel state information, RUE m and the MUE need to transmit training sequences $\boldsymbol{\psi}_m$ and $\boldsymbol{\psi}_{M+1}$ to RRHs and the MBS, respectively. All $\boldsymbol{\psi}_m$'s are with the $N(>M)$ symbol length. $\|\boldsymbol{\psi}_{M+1}\|^2=NP_b$ and $\|\boldsymbol{\psi}_m\|^2=NP_s$ with P_s and P_b being the training power. Moreover, $\boldsymbol{\Psi}=\left[\boldsymbol{\psi}_1,\boldsymbol{\psi}_2,\ldots,\boldsymbol{\psi}_M\right]$ satisfies $\boldsymbol{\Psi}^H\boldsymbol{\Psi}=NP_s\mathbf{I}_M$ and $\boldsymbol{\Psi}^H\boldsymbol{\psi}_{M+1}=\mathbf{0}_M$.

During the training stage in the uplink, the observations at the k-th RRH is given by

$$\mathbf{x}_k=\sum_{m=1}^{M} g_{mk}\boldsymbol{\psi}_m+f_{kM}\boldsymbol{\psi}_{M+1}+\mathbf{w}_k, \tag{4.45}$$

where $\mathbf{w}_k$ of $N\times 1$ is an AWGN vector whose covariance is $\sigma_n^2\mathbf{I}_N$.

For channel coefficient g_{mk}, its least-squares estimate (LSE) is calculated as

$$\hat{g}_{mk}=\frac{\boldsymbol{\psi}_m^H}{\|\boldsymbol{\psi}_m\|^2}\mathbf{x}_k, \tag{4.46}$$

and the LSE of channel coefficient between k-th RRH and MUE f_{kM} is

$$\hat{f}_{kM}=\frac{\boldsymbol{\psi}_{M+1}^H}{\|\boldsymbol{\psi}_{M+1}\|^2}\mathbf{x}_k. \tag{4.47}$$

Similarly, the observation at the MBS is given by

$$\mathbf{x}_B=\boldsymbol{\psi}_{M+1}g_B+\sum_{m=1}^{M}\boldsymbol{\psi}_m f_{Bm}+\mathbf{w}_B, \tag{4.48}$$

where $\mathbf{w}_B$ of $N \times 1$ is an AWGN vector whose covariance is $\sigma_n^2 \mathbf{I}_N$. The estimation of g_B and f_{Bm} are expressed as

$$\hat{g}_B = \frac{\boldsymbol{\psi}_{M+1}^H}{\|\boldsymbol{\psi}_{M+1}\|^2} \mathbf{x}_B, \tag{4.49}$$

$$\hat{f}_{Bm} = \frac{\boldsymbol{\psi}_m^H}{\|\boldsymbol{\psi}_m\|^2} \mathbf{x}_B, \tag{4.50}$$

respectively.

The MSEs of $\hat{g}_{mk}$ and $\hat{f}_{Bm}$, denoted by δ_{mk} and δ_{Bm}, are equal and can be given by

$$\delta_{mk} = \delta_{Bm} = \frac{\mathscr{E}\left\{\|\boldsymbol{\psi}_m^H \mathbf{w}_k\|^2\right\}}{\|\boldsymbol{\psi}_m\|^2} = \frac{\mathscr{E}\left\{\|\boldsymbol{\psi}_{M+1}^H \mathbf{w}_k\|^2\right\}}{\|\boldsymbol{\psi}_{M+1}\|^2} = \underbrace{\frac{\sigma_n^2}{N P_s}}_{\delta_1}. \tag{4.51}$$

Similarly, the MSEs of $\hat{g}_B$ and $\hat{f}_{kM}$, denoted by δ_B and δ_{kM}, are calculated as

$$\delta_B = \delta_{kM} = \underbrace{\frac{\sigma_n^2}{N P_b}}_{\delta_2}. \tag{4.52}$$

Contract Design Under Imperfect CSIs

Under incomplete CSIs, define $\hat{R}_{C1}$, $\hat{R}_{M1}$, $\hat{R}_{C2}$, $\hat{R}_C$, and $\hat{R}_M$, which are compatible with R_{C1}, R_{M1}, R_{C2}, R_C, and R_M when CSIs are complete. For all RUEs and the MUE, their data rates are derived in the following.

Phase I The RUE sum rate and the MUE data rate are calculated as

$$\hat{R}_{C1} = M \log\left(1 + \frac{P_{C1}}{\delta_1 P_{C1} \sum_{m=1}^{M} \left[\hat{\mathbf{F}}\hat{\mathbf{F}}^H\right]_{m,m}^{-1} + \delta_2 P_{M1} \left[\hat{\mathbf{F}}\hat{\mathbf{F}}^H\right]_{M+1,M+1}^{-1} + \sigma_n^2}\right), \tag{4.53}$$

$$\hat{R}_{M1} = \log\left(1 + \frac{P_{M1}}{\delta_1 P_{C1} \sum_{m=1}^{M}\left[\hat{\mathbf{F}}\hat{\mathbf{F}}^H\right]_{m,m}^{-1} + \delta_2 P_M \left[\hat{\mathbf{F}}\hat{\mathbf{F}}^H\right]_{M+1,M+1}^{-1} + \sigma_n^2}\right), \tag{4.54}$$

respectively, where the total transmit power is constrained by

$$\sum_{m=1}^{M} \frac{P_{C1}}{(K-M-1)(\upsilon_m+\delta_1)} + \frac{P_{M1}}{(K-M-1)(\upsilon_{M+1}+\delta_2)} \le P_{\max}. \tag{4.55}$$

Phase II The sum rate of all RUEs is given by

$$\hat{R}_{C2} = M\log\left(1 + \frac{P_{C2}}{\delta_1 P_{C2}\mathrm{tr}\left\{\left[\hat{\mathbf{G}}\hat{\mathbf{G}}^H\right]^{-1}\right\} + \sigma_n^2}\right), \tag{4.56}$$

where the total transmit power is constrained by

$$\sum_{m=1}^{M} \frac{P_{C2}}{(\upsilon_m+\delta_1)(K-M)} \le P_{\max}. \tag{4.57}$$

To achieve the maximal sum rate of RUEs, P_{C2} must be set to $P_{C2} = \frac{(K-M)P_{\max}}{\varepsilon_2}$ with $\varepsilon_2 = \sum_{m=1}^{M} \frac{1}{\upsilon_m+\delta_1}$.

Phase III The RUE sum rate and the MUE data rate are calculated as

$$\hat{R}_{C3} = \sum_{m=1}^{M} \log\left(1 + \frac{P_{C3}}{\delta_1 P_{C3}\mathrm{tr}\left\{\left[\hat{\mathbf{G}}\hat{\mathbf{G}}^H\right]^{-1}\right\} + |f_{Bm}|^2 P_B + \sigma_n^2}\right), \tag{4.58}$$

$$\hat{R}_{M3} = \log\left(1 + \frac{|g_B|^2 P_B}{P_{C3}\mathbf{f}\hat{\mathbf{G}}^H\left(\hat{\mathbf{G}}\hat{\mathbf{G}}^H\right)^{-2}\hat{\mathbf{G}}\mathbf{f}^H + \sigma_n^2}\right), \tag{4.59}$$

respectively. P_{C3} and P_B under the fairness power control are determined by solving the following optimization problem:

$$\begin{aligned}\{P_{C3}, P_B\} &= \arg \max_{P_{C3}, P_B} \min \left\{ \hat{R}_{C3}, \hat{R}_{M3} \right\}, \\ s.t. \quad & 0 \le P_{C3} \le P_{\max}, \; 0 \le P_B \le P_{\max}. \end{aligned} \tag{4.60}$$

With practical channel estimation, although the CSI of links from RRHs to RUEs and the MUE can be estimated by the BBU pool, the CSI $|g_B|^2$ of the MBS-MUE link cannot be exactly known, leading to information asymmetry. Assume that $|g_B|^2$ takes discrete values whose set is denoted as $\Xi = \{\xi_1, \xi_2, \ldots, \xi_L\}$ with $\xi_1 < \xi_2 < \ldots < \xi_L$, and an additional constraint, called incentive compatible (IC), is involved to design the optimal contract.

Incentive Compatible As contract is incentive compatible if the MBS with $|g_B|^2 = \xi_l$ prefers to choose the contract item $(t_1^l, t_3^l, P_{M1}, P_{C1})$, i.e.,

$$t_1^l \hat{R}_{M1} + t_3^l \hat{R}_{M3}^l \ge t_1^{l'} \hat{R}_{M1} + t_3^{l'} \hat{R}_{M3}^l, \forall l, l' \in \{1, 2, \ldots, L\}, \tag{4.61}$$

in which the MBS is allocated with t_1^l and t_3^l in Phase I and III, respectively, when $|g_B|^2 = \xi_l$. Meanwhile, $\hat{R}_{M3}^l$ follows (4.58) when $|g_B|^2 = \xi_l$.

Meanwhile, the IR constraint at this time is defined as follows.

Individual Rational For $|g_B|^2 = \xi_l$, a contract is individual rational only when the MBS's utility is higher than its reserved utility u_l by choosing the contract item $(t_1^l, t_3^l, P_{M1}, P_{C1})$, i.e.,

$$t_1^l \hat{R}_{M1} + t_3^l \hat{R}_{M3}^l \ge u_l = T_0 \hat{R}_{M3}^l, \forall l \in \{1, 2, \ldots, L\}, \tag{4.62}$$

in which the MBS is allocated with t_1^l and t_3^l in Phase I and III, respectively, when $|g_B|^2 = \xi_l$. Meanwhile, $\hat{R}_{M3}^l$ follows (4.58) when $|g_B|^2 = \xi_l$.

Assume that the pdf of $z = |f_{Bm}|^2$, the set $\Xi = \{\xi_1, \xi_2, \ldots, \xi_L\}$, and the variable q_l denoting the possibility of $|g_B|^2 = \xi_l$ are all known by the BBU pool. Obviously, $q_l \in [0, 1]$ and $\sum\limits_{l \in \{1,2,\ldots,L\}} q_l = 1$. Once the MBS accepts the contract when $|g_B|^2 = \xi_l$, the BBU pool's utility is given by

$$U_C = t_1^l \hat{R}_{C1} + \left(T_0 - t_1^l - t_3^l\right) \hat{R}_{C2} - \left(T_0 - t_3^l\right) \int_z p(z) \hat{R}_{C3} dz. \tag{4.63}$$

Similarly, the rate-based utility of the MBS is written as

$$U_M = t_1^l \hat{R}_{M1} + t_3^l \hat{R}_{M3}^l. \tag{4.64}$$

By addressing the problem below, optimal contract design can be derived when CSIs are incomplete.

$$\max_{\left\{\left(t_1^l, t_3^l, P_{M1}, P_{C1}\right), \forall l \in \{1,2,\ldots,L\}\right\}} \sum_{l \in \{1,2,\ldots,L\}} q_l \left[t_1^l \hat{R}_{C1} + \left(T_0 - t_1^l - t_3^l\right)\hat{R}_{C2} - \left(T_0 - t_3^l\right)\hat{R}_{C3}\right] \tag{4.65}$$

$$s.t. t_1^l \hat{R}_{M1} + t_3^l \hat{R}_{M3}^l \geq T_0 \hat{R}_{M3}^l, \forall l \in \{1, 2, \ldots, L\}, \tag{4.66}$$

$$t_1^l \hat{R}_{M1} + t_3^l \hat{R}_{M3}^l \geq t_1^{l'} \hat{R}_{M1} + t_3^{l'} \hat{R}_{M3}^l, \forall l, l' \in \{1, 2, \ldots, L\}, \tag{4.67}$$

$$t_3^l \geq 0, t_1^l \geq 0, T_0 - t_1^l - t_3^l \geq 0, \forall l \in \{1, 2, \ldots, L\}, \tag{4.68}$$

$$P_{M1} \geq 0,\ P_{C1} \geq 0, \tag{4.69}$$

$$\frac{P_{C1}}{(K-M-1)\,\varepsilon_2} + \frac{P_{M1}}{(K-M-1)\,(\upsilon_{M+1}+\delta_2)} \leq P_{\max}. \tag{4.70}$$

After some transformations, the problem can be solved by one-dimensional search.

4.2.5 *Numerical Results*

In the simulation, independent complex Gaussian random variables are used to model fast fading which are with zero means and unit variances. For the large-scale fading υ_m, $m = 1, \ldots, (M+1)$, we assume $\upsilon_m = z_m/(r_m/r_0)^v$, in which r_0 is 100 m, r_m is the distance from UE m to RRHs, z_m follows log-normal distribution whose standard deviation is 8 dB, and the path loss exponent satisfies $v = 3.8$. The

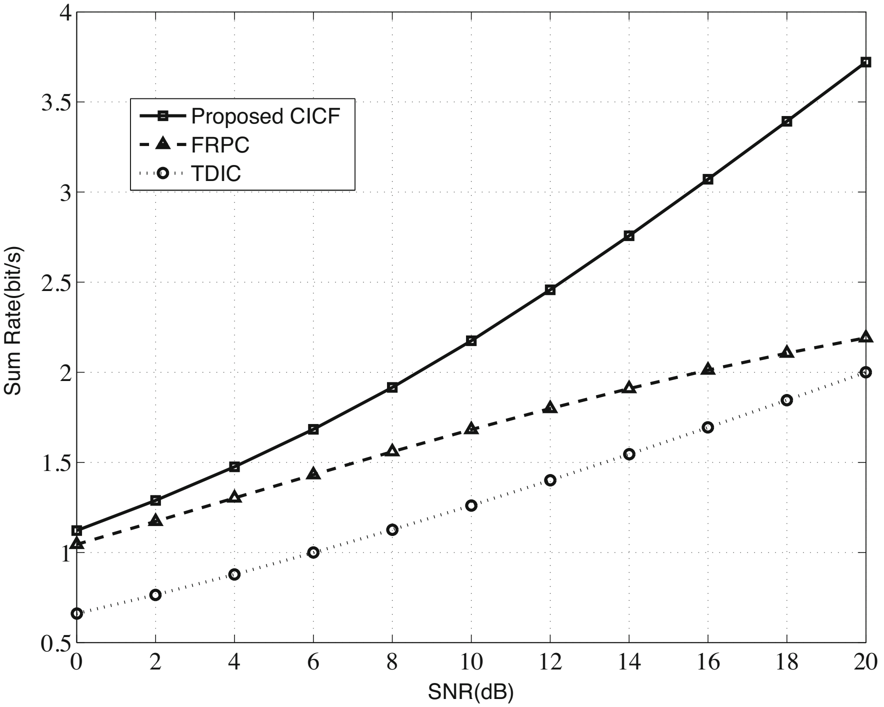


Fig. 4.5 Sum rates versus SNRs for all RUEs with different schemes

other parameters are set as $P_{\max} = 1$, $T_0 = 1$, $M = 4$, and $N = 10$. Moreover, two baselines *Frequency reuse with power control* (FRPC) and *Time domain interference cancelation* (TDIC) are considered. Specifically, in FRPC, the RRHs and MBS reuse the same frequency resource whose transmission power is determined by fairness based power control, while RRHs and the MBS transmit separately with the whole TTI equally divided, i.e., $t_R = t_M = \frac{T_0}{2}$.

Performance Evaluations Under Perfect CSIs

Figure 4.5 demonstrates the sum rate of all the RUEs under different SNRs and transmission schemes. It can be seen that contract theory based interference coordination (CICF) outperforms the other two schemes, verifying its effectiveness.

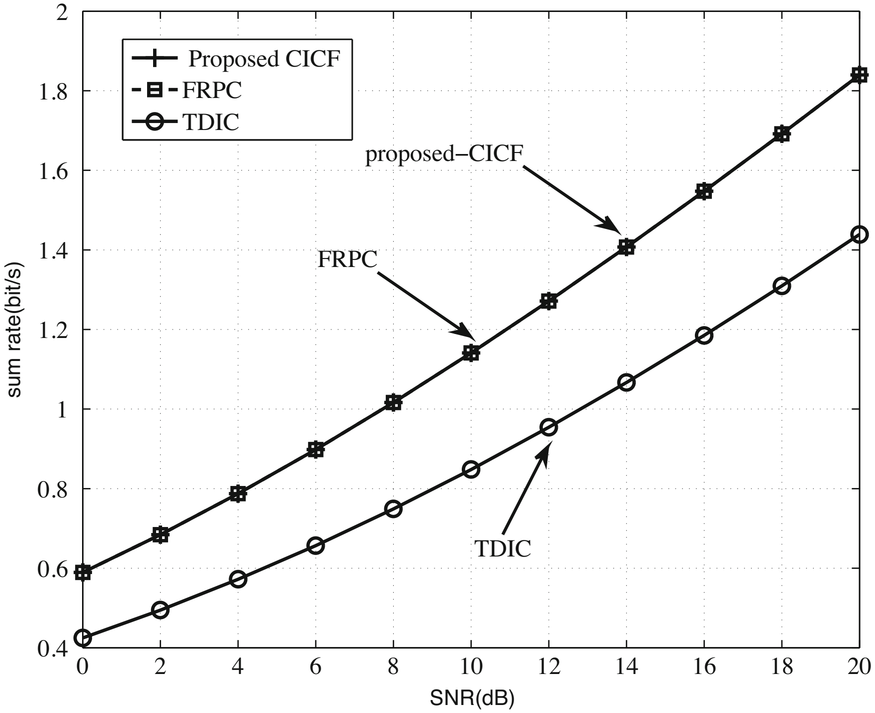


Fig. 4.6 Data rates versus SNRs for MUEs with different schemes

In Fig. 4.6, we can see that the proposed CICF leads to MUE data rate that is the same as the FRPC. This is because the IR constraint takes equality under complete and perfect CSIs when the contract design is optimal.

Performance Evaluations Under Partial CSIs

In Fig. 4.7, the proposed CICF can achieve a significant performance gain for all RUEs since both the time duration of three phases and the allocated power are optimized. In Fig. 4.8, the data rates of the MUE under different SNRs with partial CSIs are presented. From the figure, it is shown that the MBS can achieve

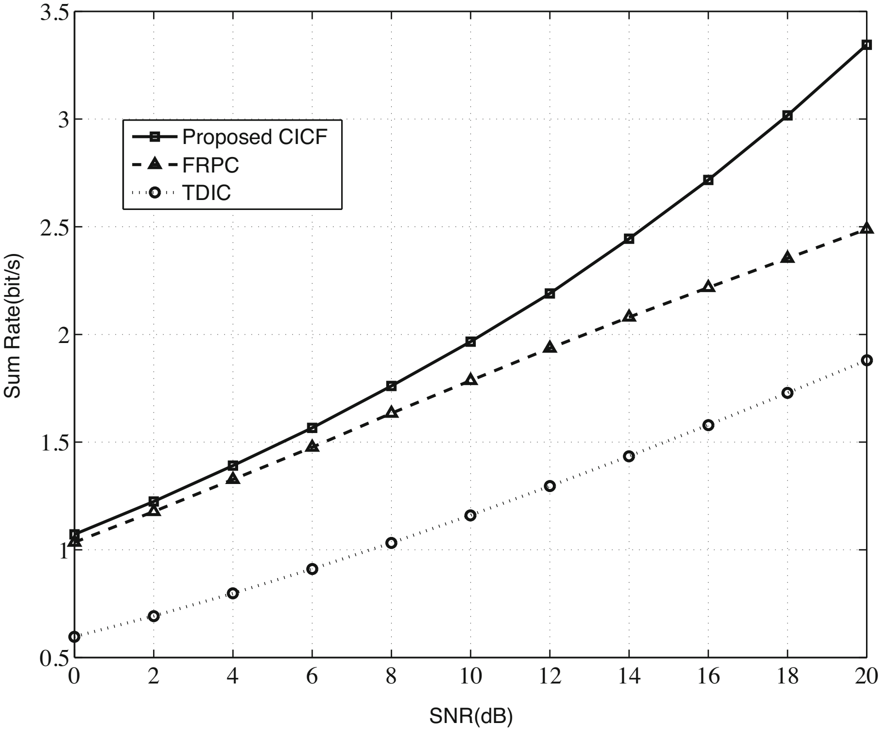


Fig. 4.7 Sum rates versus SNRs for RUEs with different schemes

a significant performance gain in the proposed CICF compared with the baselines. This is due to the fact that the optimal contract under partial CSIs can guarantee the MBS with $|g_B|^2 = \xi_l$ $(l < L)$ receives a larger utility than that obtained by not agreeing with the contract.

4.3 Summary

In this chapter, the cooperative signal processing technique in F-RANs has been studied. Specifically, cooperative non-orthogonal multiple access (NOMA) can achieve higher user data rate than traditional orthogonal multiple access and NOMA

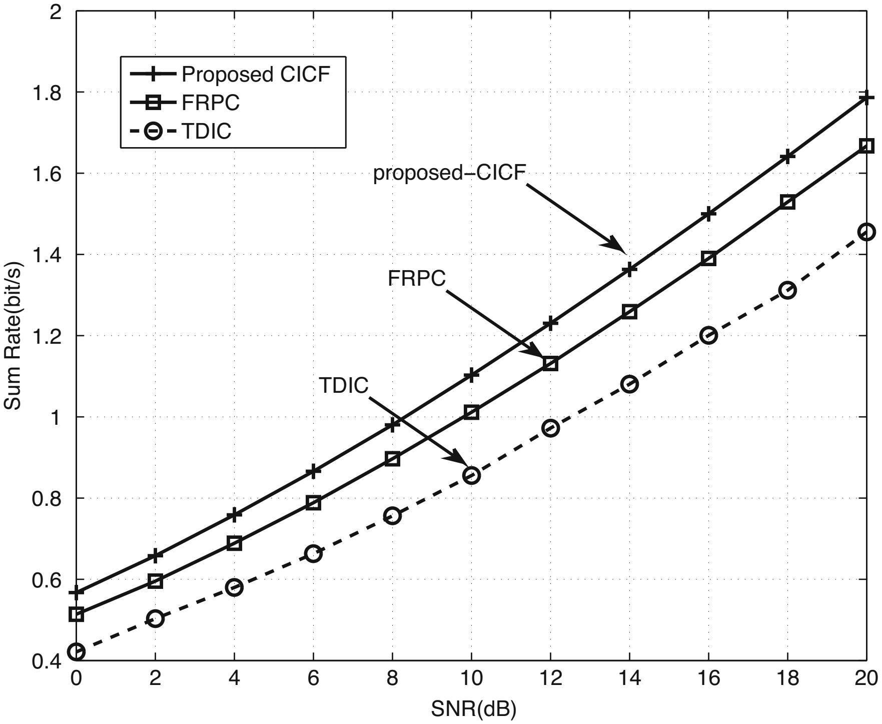


Fig. 4.8 MUE data rate under different schemes with varying SNRs

schemes, while contract theory based cooperation between RRHs and the macrocell can well overcome the information asymmetry, which outperforms traditional interference mitigation schemes.

References

Choi J (2014) Non-orthogonal multiple access in downlink coordinated two-point systems. IEEE Commun Lett 18(2):313–316

Cover TM et al (1991) Elements of information theory. Wiley-Interscience, New York

David HA et al (2003) Order statistics, 3rd edn. Wiley, Hoboken

Ding Z et al (2014) On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users. IEEE Signal Process Lett 21(12):1501–1505

Ding Z et al (2016) Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions. IEEE Trans Veh Technol 65(8):6010–6023

Han Q et al (2014) Hierarchical-game-based uplink power control in femtocell networks. IEEE Trans Veh Technol 63(6):2819–2835

Jeffrey A et al (2007) Table of integrals, series, and products. Elsevier Science, Amsterdam. https://books.google.com/books?id=aBgFYxKHUjsC

Kang X et al (2012) Price-based resource allocation for spectrum-sharing femtocell networks: a Stackelberg game approach. IEEE J Sel Areas Commun 30(3):538–549

Saito Y et al (2013) System-level performance evaluation of downlink non-orthogonal multiple access (NOMA). In: 2013 IEEE 24th annual international symposium on personal, indoor, and mobile radio communications (PIMRC). ACM, New York, pp 611–615

Chapter 5
Flexible Network Management in Fog Radio Access Networks

In Sect. 5.1, the access slicing paradigm for F-RANs will be presented, which is composed of a hierarchical management architecture and several key techniques. In Sect. 5.2, we mainly focus on resource management optimization for access slicing by involving Lyapunov theory.

5.1 The Access Slicing Paradigm

In this section, to facilitate flexible network management of F-RANs, a paradigm, called access slicing, is introduced, which is based on a hierarchical architecture. Then, radio and cache resource management and social awareness are identified as two key techniques for the paradigm, followed by the discussion of open issues.

5.1.1 Background

The fifth generation (5G) communication systems are envisioned to support various use cases (Chen et al. 2014). However, legacy network architectures are mainly broadband-service-oriented and cannot well apply in machine-type communications (mMTC) and ultra-reliable MTC (uMTC) scenarios (sec 2015). Recently, network slicing, as an emerging concept, has attracted a lot of attentions, and flexible and cost-efficient networking can be achieved for diverse services.

For existing network slicing paradigms, the creation of network slices is mainly driven by the demands of business use cases and does not take characteristics of radio access networks (RANs) into account. However, network slicing is an end-to-end solution and it needs the full support of RANs to better guarantee the customized requirements. As an attractive RAN architecture for the future wireless networks,

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0_5

fog radio access networks (F-RANs) are benefited from edge computing, cloud computing, and heterogenous networking and thus can deliver competitive network performance, in terms of spectral efficiency, energy efficiency, low latency, and high reliability. Hence, combining the concept of network slicing and F-RANs is very promising.

To cope with the drawbacks of previous network slicing solutions and unleash the F-RAN potentials, a new network slicing paradigm, termed by access slicing, is proposed. It is compatible with the existing core network (CN) based network slicing solution (NSC 2016), and it has a significant difference from both the CN-based network slicing and the RAN-based network slicing (Sallent et al. 2017). Specifically, our proposal makes full utilization of F-RANs and enjoys the superiorities of both CN-based slicing and RAN-based slicing. Moreover, our proposal also has the capability of information awareness, which helps improve various quality of service (QoS) and quality of experience (QoE) requirements.

Access slicing in this chapter is based on a hierarchical network management architecture, where a new management entity, called orchestrator for access slicing, is introduced to inherit the legacy function of CN-based network slicing. The orchestrator is responsible for network function orchestration and enabling the co-existence of multi-slices. In addition, to further enhance end-to-end performance of access slice instances, cache and radio resource allocation is then identified as one of the key techniques of the access slicing paradigm. Another key technique is the social awareness, which lays the foundation of wise decisions on the creation and adjustment of access slice instances.

5.1.2 A Hierarchical Management Architecture for Access Slicing

Key Components in the Access Slicing

Access slicing in F-RANs is illustrated in Fig. 5.1, which is divided into two layers, namely the centralized orchestration layer and the slice instance layer. In the slice instance layer, various access slice instances are included, such as mMTC, uMTC, and eMBB.

eMBB Instance In this instance, high power nodes (HPNs) are deployed mainly for control signaling delivery and providing basic data services. In addition, the performance of the HPN can be further enhanced by equipping massive MIMO, resulting in higher diversity and multiplexing gains. Via backhaul connection, the HPN can coordinate with the BBU pool. For example, centralized large-scale CoMP can significantly reduce the mutual interference between F-APs/RRHs and HPNs.

uMTC Instance For uMTC services, caching resource at F-APs is exploited and the served UEs can enjoy a considerable decrease in latency. Moreover, PHY layer, MAC layer, and network layer protocol functions can be revised to further lower the processing latency in the air interface.

mMTC Instance The mMTC instance features massive connections and mMTC devices can adopt a clustering mechanism for network access. Figure 5.1 demonstrates device clustering based on a mesh or tree-like topology, where cluster-wise packet traffic is sent to an F-AP or HPN by the cluster head. In addition, traditional air interface protocols can be re-designed to degrade the cost incurred by massive transmission. Moreover, the technique of social awareness can be used to automatically detect the active UEs, which facilitates the distributed CRRM for both intra-clusters and inter-clusters.

The Hierarchical Architecture

In the access slicing paradigm, to achieve more adaptive and convenient management of slice instances, a hierarchical network management architecture is adopted as shown in Fig. 5.1. Specifically, the centralized orchestrator can collect multi-dimensional network information that has impacts on service QoS and user QoE.

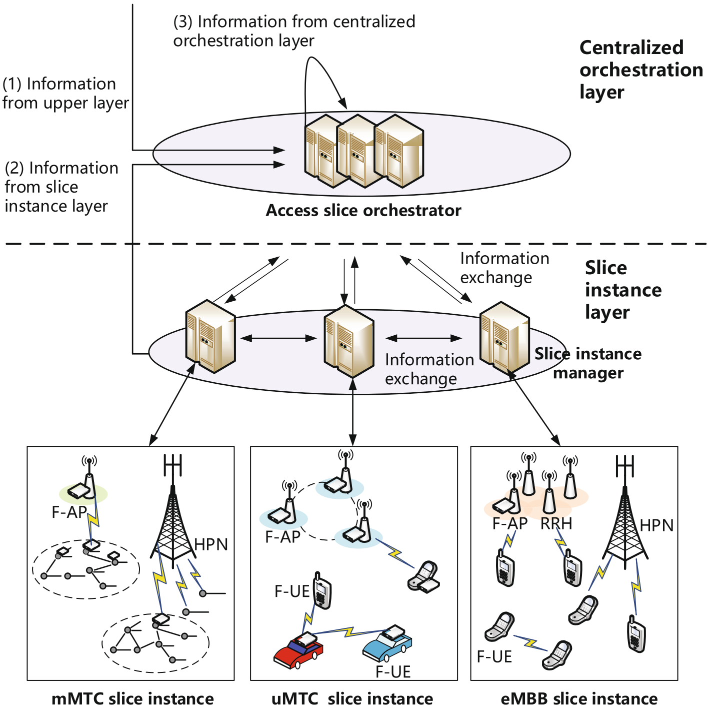

(3) Information from centralized orchestration layer
Centralized orchestration layer
(1) Information from upper layer
(2) Information from slice instance layer
Access slice orchestrator
Information exchange
Slice instance layer
Information exchange
Slice instance manager
F-AP
HPN
F-AP
F-UE
F-UE
F-AP
RRH
HPN
F-UE
mMTC slice instance
uMTC slice instance
eMBB slice instance

Fig. 5.1 An illustration of access slicing in F-RANs

Such information include the requested service and subscription information from the upper layer, RAN information and UE information from the slice instance layer, and the slice instance configuration information from the centralized orchestration layer. The RAN information further includes radio resource division among F-APs, resource utilization, interference level, and so on, while UE information further includes UE mobility, communication capabilities, battery life, and so on. Based on the multi-dimensional information, slice instance performance requirements can be met and meanwhile the whole F-RAN performance will be improved.

5.1.3 Key Techniques for Access Slicing in F-RANs

Resource Management

For access slice instances, resource management is the key to achieve slice isolation and meanwhile meet dynamic user demands as well as raise resource usage efficiency. In Zhu et al. (2016), a hierarchical combinatorial auction based resource allocation scheme in a sliced wireless network is proposed. To evaluate the performance, average social welfare, average resource utilization, and average user satisfaction are taken into account. As shown in Figs. 5.2 and 5.3, higher average social utility, average user satisfaction, etc. are achieved by the proposal.

For the access slicing in F-RANs, due to the coupling of radio and cache resource, the joint resource management is a challenging task. Particularly, the involvement of cache resource incurs many differences in terms of resource management. For example, traditional user association schemes mainly concern the channel gain between users and access points. However, in F-RANs, considering that cached contents contribute to lower latency, user association in F-RANs should not only take the channel gain into account but also whether the requested content is cached

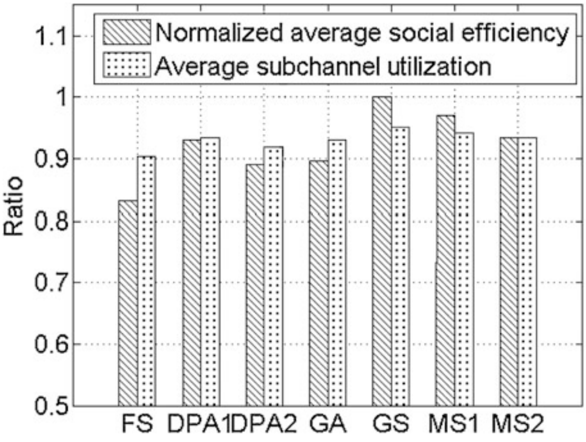


Fig. 5.2 Performance comparisons among different algorithms applied in the proposed hierarchical auction model (Zhu et al. 2016): normalized average social efficiency and average subchannel utilization under different hierarchical auction schemes

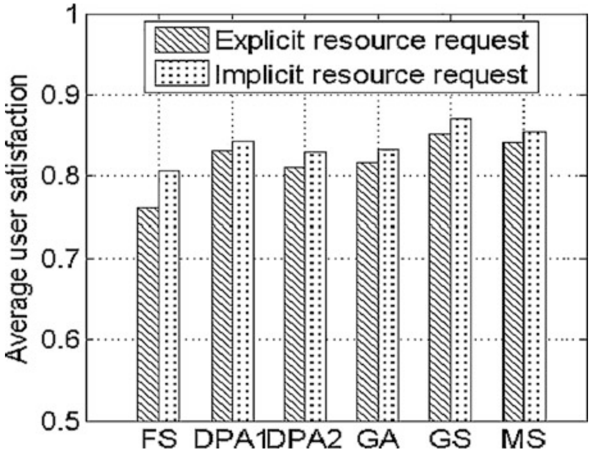


Fig. 5.3 Performance comparisons among different algorithms applied in the proposed hierarchical auction model (Zhu et al. 2016): average user satisfaction under different hierarchical auction schemes

at F-APs. Moreover, because of the potential availability of the requested content at F-APs, F-UEs, and the cloud, communication mode selection needs to be properly addressed together with resource optimization problems. As for the optimization objectives, in addition to energy efficiency, spectral efficiency, and latency, some new performance metrics have been proposed in literatures. In Dao et al. (2017), the serviceability is proposed as key foundational criteria fitting into the IoT paradigm in the 5G era. The serviceability is defined as the ability of a network to serve UEs under desired requirements (e.g., throughput, delay, and packet loss). Then, given the distribution of cached contents, a serviceability maximization problem with desired data rate constraints is formulated, which is solved by an adaptive resource balancing scheme.

Social-Aware Slicing Techniques in F-RANs

To realize access slicing in a convenient way, social awareness plays a key role. By endowing F-RANs with social-aware ability, UEs' social behaviors and interactions can be fully exploited to enhance the design and optimization of F-RANs. In Hu et al. (2015), the authors survey the cross-disciplinary research area applying social network analysis to telecommunication networking. In particular, two specific eMBB scenarios are studied to show the benefits brought by social awareness, namely the extending coverage scenario and the offloading traffic scenario. In Fig. 5.4, owing to the information of common interest (IoCI) carried by the roaming UEs, the delivery ratio of the IoCI in the extending coverage scenario has a significant improvement by 55%, which is delivered over the opportunistic links between the roaming UEs and other UEs.

As for the traffic offloading scenario, a large number of UEs are in a small area, which causes the burden of base stations and lower the spectrum efficiency. To overcome this issue, a large-scale opportunistic network is utilized, which

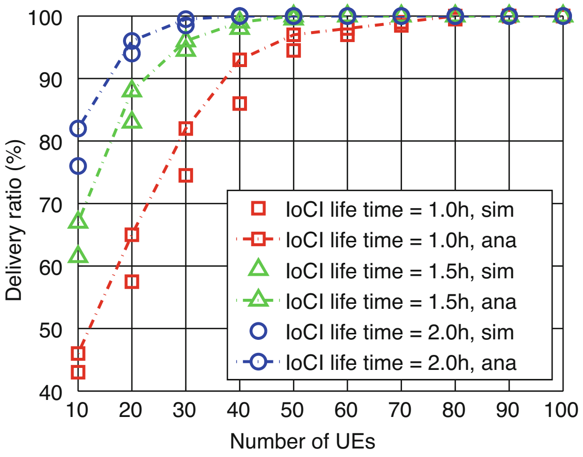


Fig. 5.4 Performance results for the extending coverage and offloading traffic scenarios in which opportunistic communication are applied (Hu et al. 2015): the average successful delivery ratio before the expiry of IoCI

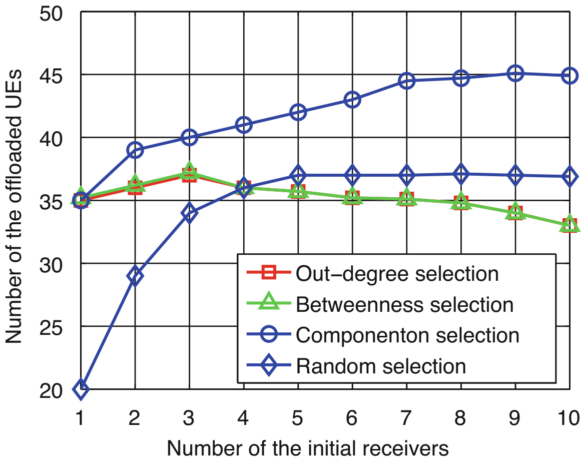


Fig. 5.5 Performance results for the extending coverage and offloading traffic scenarios in which opportunistic communication are applied (Hu et al. 2015): the average number of UEs receiving the IoCI under different selection schemes of the initial receiver set

helps offload user traffic from cellular infrastructure. Specifically, the opportunistic communication between UEs can be established and the heavy traffic is offloaded. To verify the benefits of opportunistic communication in the offloading traffic scenario, the contact profile of 78 UEs is first studied based on a realistic UE mobility trace set. Simulation results in Fig. 5.5 demonstrate the advantages of opportunistic communication in terms of traffic offloading, which leads to a considerable reduction of cellular traffic by 58% (i.e., 45 UEs).

5.1.4 Challenges and Open Issues

Currently, several challenges and open issues should be addressed to further accelerate the development of F-RAN access slicing. Firstly, there is lack of standardization on access slicing. Actually, although standardization on network slicing is on the road and attracts active participation of multiple vendors and organizations, such as the Next Generation Mobile Networks (NGMN) Alliance, 5G Public Private Partnership (5GPPP), and 3GPP. However, related studies and discussions emphasize the use cases supported by network slicing, the construction procedure of slice instances as well as the architecture of CN-based slicing. Considering the advantages of access slicing over CN-based slicing and RAN-based slicing, it is essential to pay more attention to access slicing. Second, there is still lack of open test beds related to network slicing. Up to now, lots of efforts have been made on the prototype verification of CN-based network slicing. For example, in 2016, Deutsche Telekom and Huawei jointly demonstrated the world-wide first end-to-end 5G system at Mobile World Congress to support diverse 5G use cases. The demonstration shows that network slice instances can be constructed in an optimized and autonomous way within sub-minute time. Encouraged by the recent advances in CN-based network slicing test beds, it can be anticipated in the near future that open test beds for access slicing come out, which allows researchers all over the world to verify their ideas in a convenient and cost-efficient way.

5.2 Resource Management in Sliced F-RANs

5.2.1 Background

RAN slicing has been recognized as an essential way to enhance the performance for an end-to-end network (Chen et al. 2019). However, several challenges should be overcome first for its further development. First, the performance requirements of emerging applications become more stringent. Second, as indicated in TS (TS3 2019), a RAN node needs to serve more than one slice instances. Considering the differentiation in node capability, the associated nodes should be optimized.

On the other hand, the F-RAN has been considered as a revolutionary architecture to tackle performance requirements in 5G (Peng et al. 2016). To leverage the superiority brought by F-RAN slicing, this chapter adopts a RAN slicing paradigm with a hierarchical architecture, where network resource should be properly allocated and communication mode needs to be properly selected. Particularly, UEs can operate in C-RAN mode to acquire higher data rate that is benefited from centralized signal processing at the cloud, while local data processing at F-APs and F-UEs is preferred to alleviate transmission burdens on fronthaul and save system power. Since mode selection and resource allocation are usually coupled, jointly optimizing both of them is critical to improve the performance of slice instances. In this aspect, intelligent decision-making mechanisms, such as deep reinforcement learning, can be employed.

5.2.2 System Model

Figure 5.6 shows the considered scenario, where there are L_1 single-antenna RRHs and M_0 F-APs with $L_0(L_0 < L_1)$ antennas. In addition, there are K_0 single-antenna traditional UEs and K_1 single-antenna F-UEs, whose sets are denoted as $\mathscr{K}_0$ and $\mathscr{K}_1$, respectively. The traditional UEs desire low power consumption and feature random traffic arrivals, while F-UEs are with a buffer that is sufficiently large. Each F-UE is served by a slice instance and demands high-data rate. Moreover, another slice instance is created for traditional UEs, in which C-RAN mode and F-AP mode are available, and the instance aims to lower the power consumption of traditional UEs while achieving the stability of transmission delay. Moreover, the performance of both slice instances can be further enhanced by F-UEs via D2D communication. Specifically, the traffic of an F-UE can be relayed by another F-UE, while F-UEs can allow more traditional UEs to be connected simultaneously by data aggregation. The F-RAN operates in a slotted fashion, and a decision slot is indexed by $t \in \{0, 1, 2, \ldots\}$.

The F-RAN system has N subchannels in total, and the bandwidth of each subchannel is W_0. For subchannel allocation between slice instances, two strategies are considered, namely the orthogonal and multiplexed subchannel strategies. In the first strategy, subchannel n can be only assigned to at most one traditional UE i or F-UE j to achieve hard slice isolation. In the second strategy, a subchannel can

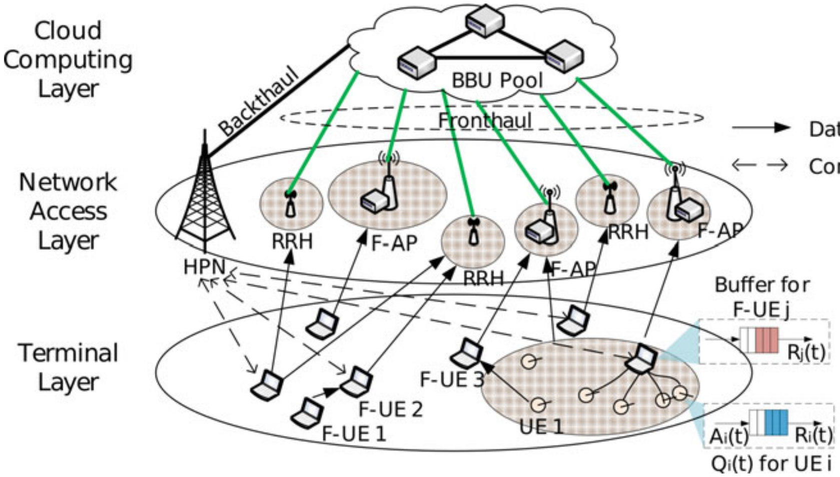


Fig. 5.6 The system model of the F-RAN slicing architecture (Xiang et al. 2020)

be shared by multiple traditional UEs and F-UEs. At this time, slice isolation can be guaranteed via proper mode selection and resource allocation. Although slice isolation in current works is realized by mainly utilizing the first strategy, it is still meaningful to study a multiplexed subchannel allocation strategy to improve spectrum utilization.

The Communication Model

For F-UE j, the associated F-AP or RRHs should be selected properly to meet its minimum rate R_{th}. Let $s_{j,m,n}^{TX}(t)$ denote the communication mode selection of F-UE j at slot t. When F-AP m ($m \in \{1, 2, \ldots, M_0\}$) is selected and subchannel n is allocated, $s_{j,m,n}^{TX}(t)$ equals to 1, and it equals to 0 otherwise. When $s_{j,0,n}^{TX}(t) = 1$, it means F-UE j connects to all the RRHs and occupies subchannel n. Once $s_{j,m,n}^{TX}(t) = 1$, the data rate of F-UE j in uplink at slot t is given as follows based on the principle of MMSE detection.

$$\begin{aligned} R_{j,m,n}(t) &= W_0 s_{j,m,n}^{TX}(t) \log\left(1 + \frac{P_{j,n}(t)\|\mathbf{v}_{j,m,n}^{H}(t)\mathbf{h}_{j,m,n}(t)\|^2}{Int_{j,m,n} + \sigma^2\|\mathbf{v}_{j,m,n}(t)\|^2}\right), \\ Int_{j,m,n} &= \sum_{k \neq j, k \in \mathscr{K}_0 \cup \mathscr{K}_1} P_{k,n}(t)\|\mathbf{v}_{j,m,n}^{H}(t)\mathbf{h}_{k,m,n}(t)\|^2, \end{aligned} \tag{5.1}$$

where σ^2 represents the noise power, $\mathbf{v}_{j,m,n}(t)$ denotes the MMSE detection vector, $P_{j,n}(t)$ denotes the transmission power of F-UE j over subchannel n, and $\mathbf{h}_{k,m,n}(t)$ represents the channel vector from UE k to F-AP m over subchannel n. Similar to the data rate calculation of F-UE j, the rate $R_i(t)$ of traditional UE i can be obtained. In addition to ensuring the rate threshold R_i^{min}, we aim at achieving the stability of the queue backlog at traditional UE i with traffic arriving randomly. The dynamics of queue backlog is given by

$$Q_i(t+1) = \max\{Q_i(t) - R_i(t), 0\} + A_i(t), \tag{5.2}$$

where $Q_i(t)$ is the queue backlog for traditional UE i in slot t and $A_i(t)$ denotes the data volume in bits of traditional UE i for uplink transmission at slot t, whose mean value is λ_i. Next, the definition of queue stability is presented below to bound the average queue backlog.

Queue Stability Neely (2010) The queue backlog $Q_i(t)$ which is a discrete time process would be mean-rate stable if

$$C0: \lim_{t\to\infty} \frac{\mathbb{E}\{|Q_i(t)|\}}{t} = 0, i \in \mathscr{K}_0. \tag{5.3}$$

Since each F-UE is assumed to be with a large buffer, it can assist to transmit the data from other F-UEs as well as traditional UEs. In Fig. 5.6, the neighbor of F-UE 1, i.e., F-UE 2, helps relay its data, while F-UE 3 helps relay the data of traditional UE 1 towards the F-AP to overcome the limitation of tradition UE 1's transmit power. Overall, for F-UE j at slot t, it not only needs to upload its own data of R_{th} bits but also needs to help transmit the data of other UEs generated at last slot. The data volume needs to be uploaded by F-UE j at slot t is $\sum\limits_{k=1}^{K_0+K_1} \mathbb{1}\{\sum\limits_{n=1}^{N} s_{k,j,n}^{TX}(t) \geq 1\} R_k(t-1) + R_{th}$, where $\mathbb{1}\{\sum\limits_{n=1}^{N} s_{k,j,n}^{TX}(t) \geq 1\}$ is an indicator function that equals to 1 when $\sum\limits_{n=1}^{N} s_{k,j,n}^{TX}(t) \geq 1$ holds and equals to 0 otherwise.

The Computing Model

In this chapter, the computing model is given by following that in Liao et al. (2017), which is related to computing resource consumption of baseband processing and MMSE detector generation. The former further includes constant computing resource C_{cons} incurred by IFFT and the computing resource $\mu_1 R_k(t)$ required by demodulation and decoding. The latter depends on the number of antennas and is given by $\mu_0 L_1^3$ when $s_{k,0,n}^{TX}(t) = 1$. Overall, computing resource consumption for UE k is modeled as

$$\begin{aligned} C_k(t) =& \mu_0 \sum_{n=1}^{N} \left(\sum_{m=1}^{M_0} s_{k,m,n}^{TX}(t) L_0^3 + s_{k,0,n}^{TX}(t) L_1^3 \right) + \\ & \mu_1 R_k(t) + C_{cons}, k \in \mathcal{K}_0 \cup \mathcal{K}_1, \end{aligned} \tag{5.4}$$

where μ_0 and μ_1 are the slopes.

Then, the following constraint on computing resource consumption needs to be considered.

$$\begin{aligned} C1: D_m^{CPU} \geq & \sum_{k=1}^{K_0+K_1} \mathbb{1} \left\{ \sum_{n=1}^{N} s_{k,m,n}^{TX}(t) \geq 1 \right\} C_k(t), \\ & m \in \{0, 1, 2, \ldots, M_0\}, \end{aligned} \tag{5.5}$$

where D_m^{CPU} is the computing resource available at F-AP m. Based on the computing model (5.4), more computing resource will be consumed by a UE when it operates in C-RAN mode compared to the case of operating in F-AP mode due to the involvement of more antennas ($L_0 < L_1$). Moreover, there is no computing resource consumption for the UEs choosing D2D mode.

5.2.3 Problem Formulation and Lyapunov Optimization

Problem Formulation

In uplink, the system power consumption is sum of power consumption led by fronthaul transmission and wireless transmission, and it is expressed as

$$\begin{aligned} P(t) = & \sum_{i=1}^{K_0} \sum_{n=1}^{N} \frac{1}{\eta_0} P_{i,n}(t) + \sum_{j=1}^{K_1} \sum_{n=1}^{N} \frac{1}{\eta_1} P_{j,n}(t) \\ & + \sum_{k=1}^{K_0+K_1} \sum_{n=1}^{N} s_{k,0,n}^{TX}(t) P^{fronthaul}, \end{aligned} \tag{5.6}$$

where η_0 and η_1 are the power amplifier efficiency of a traditional UE and an F-UE, respectively, and $P^{fronthaul}$ is the power consumption induced by fronthaul transmission, which is considered as a constant.

Moreover, the rate constraints of traditional UE i and F-UE j are given by

$$\begin{aligned} & C2 : R_i(t) \geqslant R_i^{min}, i \in \mathscr{K}_0, \\ & C3 : R_j(t) \geqslant \sum_{k=1}^{K_0+K_1} \mathbb{1}\{\sum_{n=1}^{N} s_{k,j,n}^{TX}(t) \geq 1\} R_k(t-1) \\ & \qquad + R_{th}, j \in \mathscr{K}_1. \end{aligned} \tag{5.7}$$

Based on above formulations, the aim in this chapter is to optimize system power by efficient mode selection and resource allocation that are denoted by $\{s_{k,m,n}^{TX}(t), P_{k,n}(t)\}$. The formal optimization problem is as follows.

$$\min_{\{s_{k,m,n}^{TX}(t), P_{k,n}(t)\}} \bar{P} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{P(t)\} \tag{5.8}$$

subjects to

$$
\begin{aligned}
& C0, C1, C2, C3, \\
& C4 : P_{k,n}(t) \leq \nVdash \{ \sum_{m=0}^{M_0+K_1} s_{k,m,n}^{TX}(t) = 1 \} P_{k,n}^{max}, \forall k, n, \\
& C5 : s_{k,m,n}^{TX}(t) \in \{0, 1\}, \forall k, m, n, \\
& C6 : \sum_{m=0}^{M_0+K_1} s_{k,m,n}^{TX}(t) \in \{0, 1\}, \forall k, n, \\
& C7 : \sum_{m=0}^{M_0+K_1} \sum_{n=1}^{N} s_{k,m,n}^{TX}(t) \in \{0, 1\}, \forall k,
\end{aligned}
$$

where C0 aims to keep the stability of the traditional UE's queue backlog, C1 is used to limit computing resource consumption, C2 and C3 provide the minimum data rate guarantee to traditional UEs and F-UEs, respectively, and C4 states that the transmission power of UE k equals 0 on the subchannel not assigned to it and is limited by the maximum transmission power $P_{k,n}^{max}$ otherwise. C5 is used to constrain the selection of communication mode, C6 states that a UE can select only one communication mode on a subchannel, and each UE can only be assigned with one subchannel as implied by C7.

It can be seen that problem (5.8) is a mixed integer programming and traditional optimization algorithms can lead to high computation complexity and meanwhile the problem relies on future network information that is time-varying and difficult to predict precisely. Therefore, the optimization of $\{s_{k,m,n}^{TX}(t), P_{k,n}(t)\}$ is very challenging.

General Lyapunov Optimization

With Lyapunov optimization (Neely 2010), constraints C0 can be re-formulated as a queue mean-rate stable problem that can be further resolved by using only the instantaneous information at each time slot, including channel state information and queue backlogs. Define the set of queue backlog as $\mathbf{Q}(t) = \{Q_i(t)\}$. To leverage Lyapunov optimization, we define a Lyapunov function as follows, which measures the degree of queue congestion.

$$
L\left(\mathbf{Q}(t)\right) \triangleq \frac{1}{2} \sum_{i=1}^{K_0} Q_i^2(t). \tag{5.9}
$$

Then we present the definition of Lyapunov drift aiming to alleviate the congestion of the queue backlog while maintaining stable queues.

$$\Delta\left(\mathbf{Q}(t)\right) \overset{\Delta}{=} \mathbb{E}\left\{L\left(\mathbf{Q}(t+1)\right) - L\left(\mathbf{Q}(t)\right)|\mathbf{Q}(t)\right\}. \tag{5.10}$$

To comprehensively take the queue backlog and system power consumption into account, define the drift-plus-penalty as $\Delta(\mathbf{Q}(t)){+}V\mathbb{E}\left\{P(t)|\mathbf{Q}(t)\right\}$ and V is involved to balance the system power and queue delay. Assume the expectation of $P(t)$ satisfies $P_{min} \le \mathbb{E}\left\{P(t)\right\} \le P_{max}$ with P_{min} and P_{max} being two finite constants.

Denote the optimal value of (5.8) as P^*. Assume there exist positive constants B, ε and V make the following inequality hold for the drift-plus-penalty function for all slots t and all possible $\mathbf{Q}(t)$.

$$\Delta\left(\mathbf{Q}\left(t\right)\right) + V\mathbb{E}\left\{P(t)|\mathbf{Q}(t)\right\} \le B + VP^* - \epsilon\sum_{i=1}^{K_0} Q_i(t). \tag{5.11}$$

Then C0 holds and the average system power satisfies

$$\overline{P} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\{P(t)\right\} \le P^* + \frac{B}{V}. \tag{5.12}$$

Take the average queue length as a measurement of the queue delay that meets

$$\overline{Q} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{K_0}\mathbb{E}\left\{Q_i(t)\right\} \le \frac{B+V\left(P^* - P_{min}\right)}{\epsilon}. \tag{5.13}$$

Note that the adjustment of parameter V will result in an average system power that is near to the optimum P^*. Also, the average system power and the average queue delay have a $[\mathcal{O}(1/V), \mathcal{O}(V)]$ tradeoff, which means the decrease of power consumption leads to the increase in queuing delay. Hence, parameter V for delay tolerable UEs can be set as a larger value.

Next, the upper bound of drift-plus-penalty is derived based on the queue dynamics and the Lyapunov drift. Specifically, under any control policy, we have

$$\begin{aligned}&\Delta\left(\mathbf{Q}\left(t\right)\right) + V\mathbb{E}\left\{P(t)|\mathbf{Q}(t)\right\}\\ &\quad\le B + V\mathbb{E}\left\{P(t)|\mathbf{Q}(t)\right\} - \sum_{i=1}^{K_0} Q_i(t)\mathbb{E}\left\{R_i(t) - A_i(t)|\mathbf{Q}(t)\right\},\end{aligned} \tag{5.14}$$

in which $B > \frac{1}{2}\sum_{i=1}^{K_0}\mathbb{E}\left\{R_i^2(t) + A_i^2(t)|\mathbf{Q}(t)\right\}$ is a positive and finite constant.

Leveraging the concept of opportunistic expectation minimization, we can minimize $P(t)$ based on the observation of $\mathbf{Q}(t)$ at each slot to realize the minimization of $\mathbb{E}\left\{P(t)|\mathbf{Q}(t)\right\}$. Since $Q_i(t)$, $A_i(t)$, and B are all independent of the policy at slot

t, we can solve the following problem at slot t to minimize the upper bound of the drift-plus-penalty.

$$\begin{aligned} &\min_{\{s^{TX}_{k,m,n}, P_{k,n}\}} \quad VP - \sum_{i=1}^{K_0} Q_i R_i \\ &s.t. \quad C1 \sim C7. \end{aligned} \tag{5.15}$$

As demonstrated in (5.15), the above objective is non-convex with regard to both optimization variables $s^{TX}_{k,m,n}$ and $P_{k,n}$.

5.2.4 Solution for Orthogonal and Multiplexed Subchannel Strategies

Centralized RL-Based Solution Under Orthogonal Subchannel Allocation

To select communication modes for UEs effectively, a Q-learning based approach is proposed. To decrease the dimensions of the Q-table, system state is defined as $\mathbf{s} = \{k_0,\ s_k | k = 1,\ 2,\ \ldots,\ K_0 + K_1\}$. Note that only UE k_0 would reselect a mode according to the action, and $s^{TX}_{k,m,n} = 1$ when $s_k = n + mN$, which represents subchannel n has been assigned to UE k accessing F-AP m. After selecting action a, k_0 and s_{k_0} in $\mathbf{s}$ are updated, leading to the state transition.

The update rule of the Q-value is given by

$$Q_{k,m,n} \leftarrow (1-\alpha) Q_{k,m,n} + \alpha W_{k,m,n}, \tag{5.16}$$

in which $W_{k,m,n}$ represents the feedback reward by adopting action a and $\alpha \in (0,\ 1)$ denotes the learning rate. The reward is defined as

$$W_{k,m,n} = \begin{cases} 1 - \frac{V_0 P_{k,n} + s^{TX}_{k,0,n} V P^{fronthaul} - Q_k R_{k,m,n}}{V_0 P^{max}_{k,n} + s^{TX}_{k,0,n} V P^{fronthaul} - Q_k R^{min}_k}, & k \in \mathscr{K}_0, \\ 1 - \frac{V_1 P_{k,n} + s^{TX}_{k,0,n} V P^{fronthaul}}{V_1 P^{max}_{k,n} + s^{TX}_{k,0,n} V P^{fronthaul}}, & k \in \mathscr{K}_1, \end{cases} \tag{5.17}$$

where $V_0 = \frac{V}{\eta_0}$ and $V_1 = \frac{V}{\eta_1}$.

The probability that UE k associates with F-AP m and occupies subchannel n is given by

$$Pr_{k,m,n} = \frac{e^{\frac{Q_{k,m,n}}{\tau}}}{\sum_{m'=0}^{M_0+K_1} \sum_{n'=1}^{N} e^{\frac{Q_{k,m',n'}}{\tau}}}, k \in \mathscr{K}_0 \cup \mathscr{K}_1, \tag{5.18}$$

where $\tau = \tau_0 / \log(1 + t_{epi})$ is the temperature parameter.

After $\{s^{TX}_{k,m,n}\}$ are identified via Q-learning, problem (5.15) is simplified into the following problem.

$$\begin{aligned}\min_{\{P_{k,n}\}} \quad & \sum_{i=1}^{K_0}\sum_{n=1}^{N} V_0 P_{i,n} + \sum_{j=1}^{K_1}\sum_{n=1}^{N} V_1 P_{j,n} - \sum_{i=1}^{K_0} Q_i R_i \\ s.t. \quad & C1 \sim C4.\end{aligned} \tag{5.19}$$

Suppose $\{P^*_{k,n}\}$ is the extreme point of the targeted convex function. Once $\{P^*_{k,n}\}$ lies in the feasible region described by C1 $\sim$ C4, it is the optimum to problem (5.19). Otherwise, the optimal solution can be found by an iterative algorithm shown in Algorithm 1.

Algorithm 1 An iterative approach to solving problem (5.19)

```
1: For the targeted optimization function in (5.19), derive its partial derivative;
2: Get the extreme point $\{P^*_{k,n}\}$ of the targeted convex function.
3: Initialize $\{P_{k,n}\} = \{P^*_{k,n}\}$ and a fixed step $\triangle P$ is defined;
4: repeat
5:   With $P_{k,n}$ fixed, calculate the partial derivative $f'(P_{k,n})$;
6:   Take $k^* = arg\ min_k f'(P_{k,n})$;
7:   Let $P_{k^*,n} = P_{k^*,n} - \triangle P$;
8: until $\{P_{k,n}\}$ is in the feasible region.
```

Distributed RL-Based Solution Under Multiplexed Subchannel Allocation

When subchannels are allowed to be multiplexed among UEs, we develop a RL-based method that makes UEs perform distributed communication mode selection. Compared to centralized approaches, each UE at this time only cares about its own selection probabilities. By just taking the states of neighbor nodes into account, each UE can store the Q-table in a more affordable way because of the reduction of Q-table size. After UE k has chosen RRHs, or an F-AP or an F-UE together with a subchannel n, $Q_{k,m,n}$ is updated by following (5.16), in which $m = 0$ means selecting RRHs, $m = \{1, 2, \ldots, M_0\}$ means selecting an F-AP and $m = \{M_0 + 1, M_0 + 2, \ldots, M_0 + K_1\}$ means selecting an F-UE. As for the definition of $W_{k,m,n}$, the same definition holds as that in (5.17) if constraints C1, C2, and C3 are all satisfied. Otherwise, $W_{k,m,n}$ is set 0.

After distributed Q-learning outputs $\{s^{TX}_{k,m,n}\}$, problem (5.15) can now be transformed into the following problem.

$$\begin{aligned}
\min_{\{P_{k,n(k)}\}} \quad & \sum_{i=1}^{K_0} V_0 P_{i,n(i)} + \sum_{j=1}^{K_1} V_1 P_{j,n(i)} - \sum_{i=1}^{K_0} Q_i R_i \\
s.t. \quad & C1, C4 \\
& D2: \sqrt{\sum_{k'=1}^{K_0+K_1} P_{k',n(k')} \|\mathbf{v}_{k,m(k),n(k)}^H \mathbf{h}_{k',m(k),n(k)}\|^2 + \sigma^2 \|\mathbf{v}_{k,m(k),n(k)}\|^2} \\
& \leq \sqrt{1+\frac{1}{\gamma_k^{QoS}}} \mathbf{Re}\{\mathbf{v}_{k,m(k),n(k)}^H \mathbf{h}_{k,m(k),n(k)}\} P_{k,n(k)}^{\frac{1}{2}}, k \in \mathscr{K}_0 \cup \mathscr{K}_1,
\end{aligned} \tag{5.28}$$

in which γ_k^{QoS} represents the SINR that corresponds to data rate R_i^{min} and sum rate threshold in the right side of C3. D2 is a second-order cone constraint got from C2 and C3.

Next, the following proposition is useful to solve the above problem.

Equivalent Problem The following problem possesses the same optimum as problem (5.28).

$$\begin{aligned}
\min_{\{w_k, u_k, P_{k,n(k)}^{\frac{1}{2}}\}} \quad & \sum_{i=1}^{K_0} Q_i \{w_i e_i - \log w_i\} \\
& + \sum_{i=1}^{K_0} V_0 P_{i,n(i)} + \sum_{j=1}^{K_1} V_1 P_{j,n(j)}, \\
s.t. \quad & C1, C4, D2,
\end{aligned} \tag{5.29}$$

in which $u_k \in \mathbb{C}$ denotes the receiver variable, w_k is the weight for UE k, and e_k is the corresponding mean-squared-error (MSE) that is given by

$$\begin{aligned}
e_k \triangleq & \|u_k \sum_{k'} \mathbf{v}_{k,m(k),n(k)}^H \mathbf{h}_{k',m(k),n(k)} P_{k',n(k')}^{\frac{1}{2}}\|^2 \\
& - 2\mathbf{Re}\{u_k \mathbf{v}_{k,m(k),n(k)}^H \mathbf{h}_{k,m(k),n(k)}\} P_{k,n(k)}^{\frac{1}{2}} \\
& + \sigma^2 \|u_k \mathbf{v}_{k,m(k),n(k)}\|^2 + 1.
\end{aligned} \tag{5.30}$$

Note that the block coordinate descent method can be utilized to obtain a stationary point of problem (5.29).

Table 5.1 Simulation parameters

Fronthaul power $P^{fronthaul}$	0.35 W
Noise power spectral density	− 164 dBm/Hz
Subchannel bandwidth W_0	180 kHz
Power amplifier efficiencies η_0, η_1	0.05, 0.05
Pathloss model	$127 + 25\log_{10}$ (d)(km)

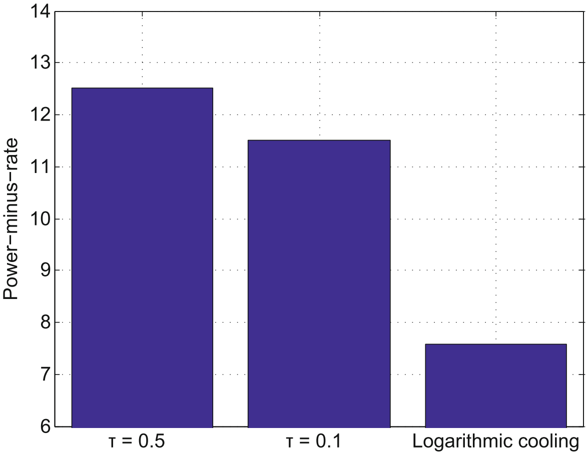


Fig. 5.7 The impacts of the temperature parameter τ in the orthogonal subchannel strategy

5.2.5 Simulation Results

In simulation, the number of RRHs and the number of F-APs are $L_1 = 10$ and $M_0 = 3$, respectively. All the RRHs and F-APs are located in a region of 1000×1000 m, and we equip each F-AP with $L_0 = 6$ antennas. The minimum required bit rates of each F-UE and each traditional UE are set to 0.6 and 0.06 Mbits/slot, respectively. For all the traditional UEs, the mean data arrival rate is assumed to be the same. Other parameter settings are presented below (Table 5.1).

The Impacts of Different Parameters

From Fig. 5.7, It can be observed that a smaller value of τ achieves a better performance, and the value of the power-minus rate function becomes larger with total computing resource decreasing as demonstrated in Fig. 5.8. Although raising V can reduce system power consumption, such improvement is less significant when V is sufficiently large according to the result in Fig. 5.9, and meanwhile a larger V will lead to worse queue delay as illustrated in Fig. 5.10.

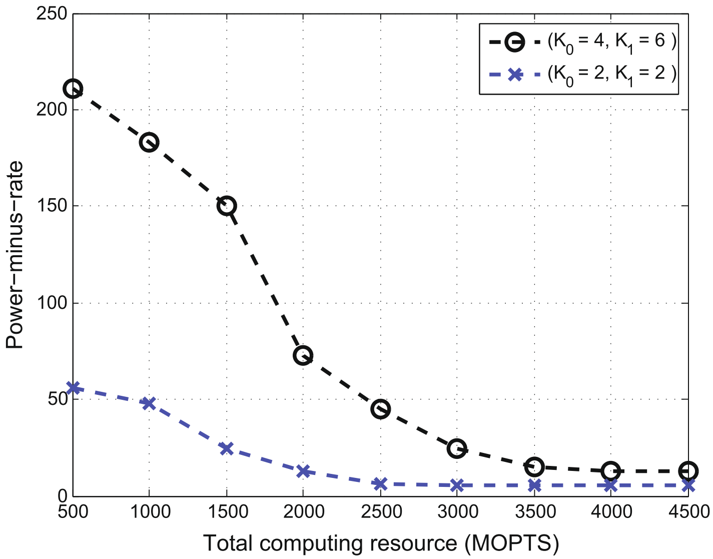


Fig. 5.8 Power-minus-rate vs. total computing resource at slot *t*

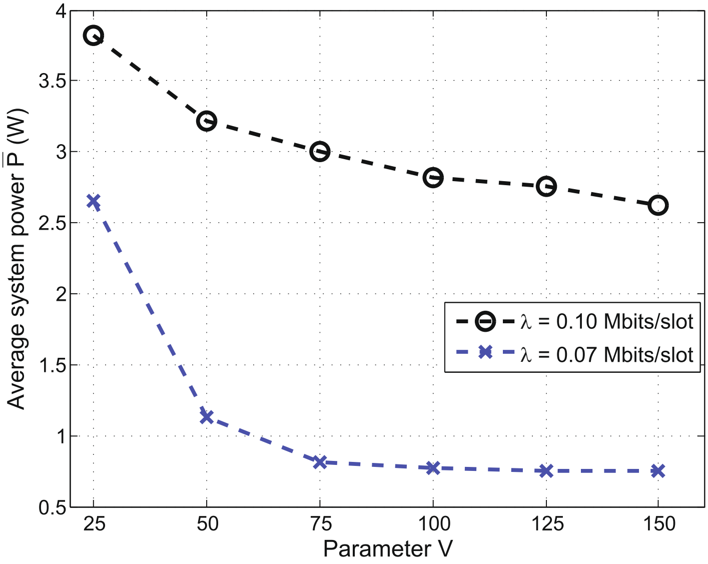


Fig. 5.9 Average system power $\overline{P}$ vs. parameter *V*

Performance Comparison with Benchmarks

In Fig. 5.11, we compare the performance of the proposed Q-learning approach with different baseline schemes. All the RRHs scheme means all the UEs access

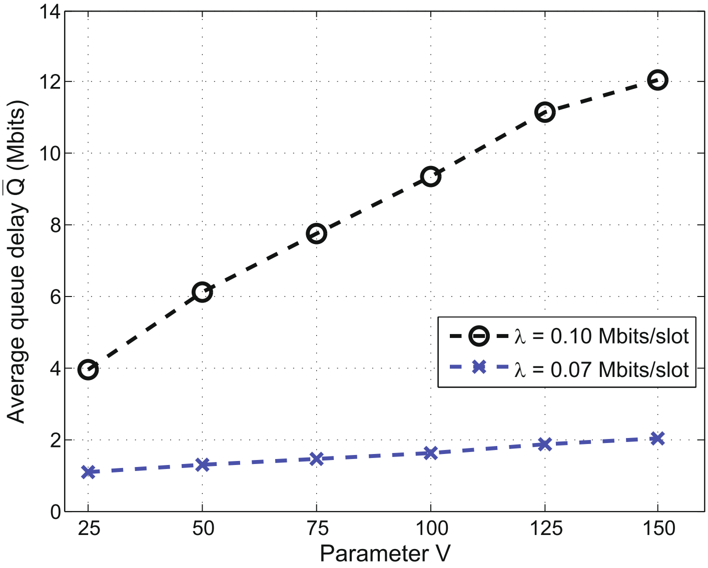


Fig. 5.10 Average queue delay $\overline{Q}$ vs. parameter V

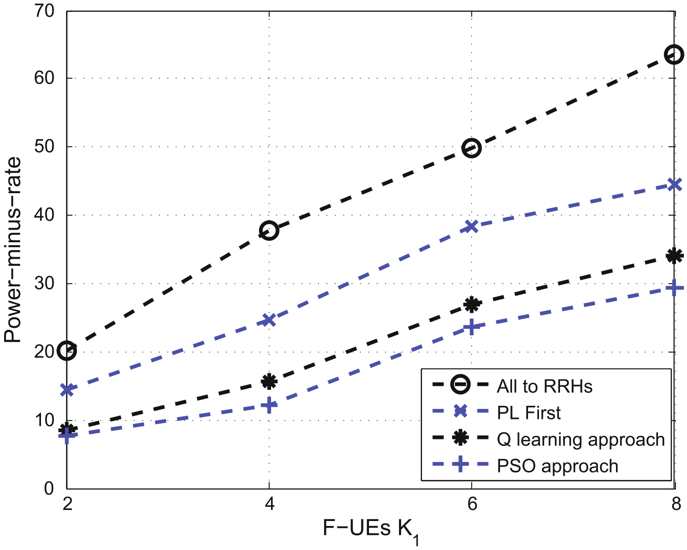


Fig. 5.11 Performance evaluation at slot t under multiplexed subchannel allocation with F-UE number varying

the RRHs, while PL First scheme means UEs access F-APs or RRHs that bring the best propagation loss. It can be seen that our proposal always reaches better performance than these two schemes. Moreover, the proposed approach results in competitive performance relative to the PSO approach but with much lower computation complexity. Specifically, the PSO approach takes around 35 min while our proposal only takes 2 min, when the number of F-UEs is 6.

5.3 Summary

In this chapter, access slicing based flexible network management for fog radio access networks has been introduced. The paradigm is realized based on a hierarchical management architecture consisting of a centralized orchestration layer and a slice instance layer. Meanwhile, radio and cache resource management and social awareness have been presented as two key techniques for access slicing. For radio resource management, we have proposed a Lyapunov optimization based scheme for a scenario with two slices, whose superior performance has been verified by extensive simulation.

References

Chen S et al (2014) The requirements, challenges and technologies for 5G of terrestrial mobile telecommunication. IEEE Commun Mag 52(5):36–43

Chen X et al (2019) Optimized computation offloading performance in virtual edge computing systems via deep reinforcement learning. IEEE Internet Things 6(3):4005–4018

Dao N et al (2017) Adaptive resource balancing for serviceability maximization in fog radio access networks. IEEE Access 5:14548–14559

Hu J et al (2015) Bridging the social and wireless networking divide: information dissemination in integrated cellular and opportunistic networks. IEEE Access 3:1809–1848

IMT-vision-framework and overall objectives of the future development of IMT for 2020 and beyond (2015). Recommendation ITU-R M

Liao Y et al (2017) How much computing capability is enough to run a cloud radio access network? IEEE Commun Lett 21(1):104–107

Neely M (2010) Stochastic network optimization with application to communication and queuing systems. Morgan and Claypool, San Rafael

Nr and ng-ran overall description (2019). 3GPP, TS 38300

Peng M et al (2016) Fog computing based radio access networks: issues and challenges. IEEE Netw 30(4):46–53

Sallent O et al (2017) On radio access network slicing from a radio resource management perspective. IEEE Wireless Commun 24(5):166–174

Study on architecture for next generation system (2016). 3GPP TR 23799

Xiang H et al (2020) Mode selection and resource allocation in sliced fog radio access networks: a reinforcement learning approach. IEEE Trans Veh Technol 69:4271–4284. https://doi.org/10.1109/TVT.2020.2972999

Zhu K et al (2016) Virtualization of 5G cellular networks as a hierarchical combinatorial auction. IEEE Trans Mobile Comput 15(10):2640–2654

Chapter 6
Dynamic Resource Allocation in Fog Radio Access Networks

In Sect. 6.1, cost-aware energy efficient resource allocation will be studied based on traditional centralized optimization theory. In Sect. 6.2, game theory-based resource optimization will be adopted to mitigate intra-tier interference among fog access points. In Sect. 6.3, resource allocation for dynamic F-RANs will be addressed by the advanced deep reinforcement learning method.

6.1 Centralized Cost-Aware Energy Efficiency Optimization in F-RANs

6.1.1 Background

To overcome the existing challenges in the cloud radio access networks (C-RANs), the fog radio access networks (F-RANs) have been proposed as a solution to provide high spectral efficiency and energy efficiency (Peng et al. 2016). In addition to remote radio heads (RRHs) in C-RANs, F-RANs also involve fog access points (F-APs) that possess local resource management, signal processing, and caching capabilities, which can alleviate the burdens on fronthaul and the BBU pool.

In this chapter, a metric called economical energy efficiency (E^3) is introduced to balance the energy efficiency (EE) improvement and the corresponding cost that have a close relationship with edge caching to achieve an efficient, green, and economically acceptable network.

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0_6

6.1.2 System Model and Problem Formulation

This chapter considers an F-RAN with N_1 RRHs, N_2 F-APs, and K single-antenna UEs. Each AP is equipped with T antennas. The set of RRHs, F-APs, and active UEs are denoted by $\mathscr{N}_1 = \{1, 2, 3, \ldots, N_1\}$, $\mathscr{N}_2 = \{N_1 + 1, N_1 + 2, N_1 + 3, \ldots, N_1 + N_2\}$, and $\mathscr{K} = \{1, 2, 3, \ldots, K\}$, respectively. The set of all APs can be denoted as $\mathscr{N} = \mathscr{N}_1 \bigcup \mathscr{N}_2 = \{1, 2, 3, \ldots, N\}$, in which $N = N_1 + N_2$. The capacity-constrained fronthaul links connect the RRHs and the BBU pool. The set of all contents requested by UEs is denoted by $\mathscr{D} = \{D_1, D_2, D_3, \ldots, D_F\}$. Let $\pi_k \in \mathscr{D}$ denote the content requested by UE k. Let S_C ($S_C < F$) denote the size of the centralized cache and S_n ($S_n < F$) denote the cache size at F-AP n. $\mathscr{S}_C$ and $\mathscr{S}_n$ denote the set of the contents stored in the centralized cache and those cached by F-AP n, respectively. When a UE accesses RRHs, its requested content is downloaded from the cloud via fronthaul, and the cloud has to fetch the content from the internet if the content is not cached. When a UE accesses an F-AP, the content is delivered from the local cache. We define a binary matrix $\boldsymbol{M} \in \{0, 1\}^{K \times N}$ to represent the node selection.

In this chapter, we adopt a novel performance metric, called **E**conomical **E**nergy **E**fficiency (E^3), which is the ratio of effective system throughput to energy consumption weighted by cost coefficient (Yan et al. 2017). For F-RANs, it can be formulated as

$$E^3 = \frac{\sum\limits_{k \in \mathscr{K}} \alpha_k \, R_k}{\sum\limits_{n \in \mathscr{N}} (P_{Tn} + P_{0n} \, C_n)}, \tag{6.1}$$

in which R_k denotes the data rate of UE k, α_k is involved to prioritize different UEs, C_n represents the cost coefficient of AP n, and P_{0n} and P_{Tn} are the static power and the load-dependent power, respectively. Without loss of generality, it assumes that $\alpha_k > 0, \forall k \in \mathscr{K}$.

The Communication Model

Let $\mathbf{w}_k = \left[\mathrm{w}_{k,1}, \mathrm{w}_{k,2}, \ldots, \mathrm{w}_{k,N}\right] \in \mathbb{C}^{1 \times NT}$ denote the beamforming vector for UE k and $\mathbf{h}_k = \left[h_{k,1}, h_{k,2}, \ldots, h_{k,N}\right]^T \in \mathbb{C}^{NT \times 1}$ the channel state information vector for UE k. Then, the received downlink signal at UE k can be written as

$$y_k = \mathbf{w}_k \mathbf{h}_k x_k + \sum_{j \in \mathscr{K}, j \neq k} \mathbf{w}_j \mathbf{h}_k x_j + z_k, \forall k \in \mathscr{K}, \tag{6.2}$$

in which x_k represents UE k's signal that is independent and identically distributed according to $\mathscr{CN}(0, 1)$, and z_k is the noise following distribution $\mathscr{CN}(0, \sigma^2)$.

UE k's data rate is given by

$$R_k\left(\mathbf{w}\right) = B_0 \log\left(1 + \frac{\mathbf{w}_k \mathbf{h}_k (\mathbf{w}_k \mathbf{h}_k)^H}{\sigma^2 + \sum\limits_{j \in \mathscr{K}, j \neq k} \mathbf{w}_j \mathbf{h}_k \left(\mathbf{w}_j \mathbf{h}_k\right)^H}\right), \tag{6.3}$$

where B_0 is the available bandwidth of the system.

The Power Model

The circuit power of AP n can be denoted as a constant, P_{0n}, and the transmitting power can be written as

$$P_{Tn}\left(\mathbf{w}\right) = \sum_{k \in \mathscr{K}} \varphi_n \left\| \mathrm{w}_{k,n} \right\|_2^2, \forall n \in \mathscr{N}, \tag{6.4}$$

where φ_n is the efficiency of the power amplifier.

Cost Coefficient

The cost on fetching the content from the internet can be denoted as $\sum\limits_{k \in \mathscr{K}, \pi_k \in \overline{\mathscr{S}}_C} c_A(\pi_k)$, where $\overline{\mathscr{S}}_C$ is the set of the contents that is not stored in the centralized cache, and $c_A(\pi_k)$ represents the cost on acquiring content π_k from the internet (Wang et al. 2014). The cost on content caching is related to the size of the cache and the content refreshment. Assume that cost on centralized cloud is equally divided among RRHs, and hence the content acquiring part of the cost coefficient is given by

$$c_{Cn} = \frac{1}{N_1}\left(c_P\, S_C + c_{RC} + \sum_{k \in \mathscr{K}, \pi_k \in \overline{\mathscr{S}}_C} c_A(\pi_k)\right), \tag{6.5}$$

where c_P denotes the price induced by consuming unit cache resource, while c_{RC} is the content refreshing cost for the cloud.

$\forall n \in \mathscr{N}_2$, c_{Cn} can be given by

$$c_{Cn} = c_P\, S_n + c_{Rn}, \tag{6.6}$$

where c_{Rn} stands for the cost on content refreshment for F-AP n.

For RRH n, the cost coefficient on fronthaul transmission can be written as

$$c_{Fn}(\mathbf{M}) = c_{Tn} \sum_{k \in \mathscr{K}} m_{k,n},\ \forall n \in \mathscr{N}_1, \tag{6.7}$$

where c_{Tn} represents the per-UE cost of fronthaul transmission that relates to fronthaul condition. For F-APs, since contents are delivered directly from their local caches, $c_{Fn}(\mathbf{M}) = 0,\ \forall n \in \mathcal{N}_2$. Overall, the cost coefficient is written as

$$C_n(\mathbf{M}) = c_{O_n} + c_{Cn} + c_{Fn}(\mathbf{M}), \forall n \in \mathcal{N}, \tag{6.8}$$

in which constant c_{O_n} represents the other cost.

Problem Formulation

The E^3-optimizing resource allocation problem in F-RANs under various constraints is formulated as

$$\begin{aligned} &\max_{\{\mathbf{w},\mathbf{M}\}} \frac{\sum\limits_{k\in\mathcal{K}} \alpha_k\, R_k(\mathbf{w})}{\sum\limits_{n\in\mathcal{N}} \left(P_{Tn}(\mathbf{w}) + P_{0n}\, C_n(\mathbf{M})\right)} \\ &\text{s.t.} C1 : \sum_{k\in\mathcal{K}} R_k(\mathbf{w}) \nVdash \left\{\left\|\mathrm{w}_{k,n}\right\|_2^2\right\} \le R_C^n, \forall n \in \mathcal{N}_1, \\ &C2 : \sum_{n\in\mathcal{N}} \left\|\mathrm{w}_{k,n}\right\|_2^2 \le P_{\mathrm{MAX}}^n, \forall n \in \mathcal{N}, \\ &C3 : \nVdash \left\{\left\|\mathrm{w}_{k,n}\right\|_2^2\right\} \le m_{k,n}, \forall k \in \mathcal{K}, n \in \mathcal{N}, \\ &C4 : m_{k,n} \in \{0, 1\}, \forall k \in \mathcal{K}, n \in \mathcal{N}, \\ &C5 : \pi_k \in \mathcal{S}_n, \forall m_{k,n} > 0, n \in \mathcal{N}_2, \\ &C6 : m_{k,i} + \sum_{j\in\mathcal{N}_2} m_{k,j} \le 1, \forall k \in \mathcal{K}, i \in \mathcal{N}_1, \end{aligned} \tag{6.9}$$

where $\nVdash\{\cdot\}$ is the indicator function defined as

$$\nVdash\{x\} = \begin{cases} 0, & \text{if } x = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{6.10}$$

P_{MAX}^n presents the upper bound of AP n's transmit power and R_C^n represents RRH n's fronthaul capacity. $C3$ requires that the power allocated to a UE by an AP is 0 if the UE is not associated with this AP. $C5$ means each UE can only access the F-AP that has its requested content. $C4$ and $C6$ are constraints regarding transmission mode selection.

To handle the non-convexity of $C3$, we rewrite it as follows:

$$C3 : \left\|w_{k,n}\right\|_2^2 \beta_{k,n} \le m_{k,n}, \forall k \in \mathcal{K}, n \in \mathcal{N}, \tag{6.11}$$

in which $\beta_{k,n}$ is the constant weight factor corresponding to the beamforming vector of AP n for UE k and makes the following update iteratively:

$$\beta_{k,n} = \frac{1}{\|\mathrm{w}_{k,n}\|_2^2 + \tau}, \ \forall k \in \mathscr{K}, n \in \mathscr{N}, \tag{6.12}$$

in which $\tau > 0$ is a constant for regularization and $\left\|\mathrm{w}_{k,n}\right\|_2^2$ is calculated based on the result from the last iteration.

The transmitting rate $R_k(\boldsymbol{w})$ in both the objective function and the fronthaul constraints makes it difficult to solve problem (6.9). To address the problem, $C1$ is reformulated as

$$C1 : \sum_{k \in \mathscr{K}} \hat{R}_k \left\|w_{k,n}\right\|_2^2 \beta_{k,n} \leq R_C^n, \forall n \in \mathscr{N}. \tag{6.13}$$

We solve the E^3 optimization problem (6.9) iteratively with fixed $\beta_{k,n}$ and $\hat{R}_k$ obtained from the previous iteration.

6.1.3 Problem Solution and Algorithm Design

Optimization Problem Reformulation

As it can be observed, *Problem 1* can be classified as a nonlinear fractional program (Peng et al. 2015). Let $U_R(\mathbf{w}, \mathbf{M}) = \sum\limits_{k \in \mathscr{K}} \alpha_k R_k(\mathbf{w})$, and

$$U_{PC}(\mathbf{w}, \mathbf{M}) = \sum_{n \in N} (P_{Tn}(\mathbf{w}) + P_{0n} C_n(\mathbf{M})), \tag{6.14}$$

then we can denote E^3 as a nonnegative variable q, i.e.,

$$q = \frac{U_R(\mathbf{w}, \mathbf{M})}{U_{PC}(\mathbf{w}, \mathbf{M})}, \tag{6.15}$$

whose optimal value is denoted as $q^* = U_R(\boldsymbol{w}^*, \boldsymbol{M}^*)/U_{PC}(\boldsymbol{w}^*, \boldsymbol{M}^*)$.

Let $F(q) = \max\limits_{\{\mathbf{w}, \mathbf{M}\}} U_R(\mathbf{w}, \mathbf{M}) - qU_{PC}(\mathbf{w}, \mathbf{M})$. q^* can be achieved if and only if

$$\begin{aligned} F(q^*) &= \max_{\{\mathbf{w}, \mathbf{M}\}} U_R(\mathbf{w}, \mathbf{M}) - q^* U_{PC}(\mathbf{w}, \mathbf{M}) \\ &= U_R(\mathbf{w}^*, \mathbf{M}^*) - q^* U_{PC}(\mathbf{w}^*, \mathbf{M}^*) = 0, \end{aligned} \tag{6.16}$$

in which $\{\boldsymbol{w}, \boldsymbol{M}\}$ is any feasible solution of *Problem 1* meeting $C1$–$C6$. Given q^*, the same policies are shared by the *Problem 1* and the problem below.

Problem 2 (Transformed E^3 Optimization)

$$\begin{aligned} &\max_{\{\mathbf{w},\mathbf{M}\}} U_R(\mathbf{w},\mathbf{M}) - q^* U_{PC}(\mathbf{w},\mathbf{M}) \\ &\text{s.t.} \quad C1{-}C6. \end{aligned} \tag{6.17}$$

Algorithm Framework

The upper bound of $R_k(\boldsymbol{w})$ and the lower bound of $C_n(\boldsymbol{M})$ can be written as:

$$R_k^{\max} = B_0 \log\left(1 + \frac{\sum\limits_{n\in\mathcal{N}} P_{\text{MAX}}^n}{\sigma^2}\right), \tag{6.18}$$

$$C_n^{\min} = c_{O_n} + c_{C_n}. \tag{6.19}$$

Accordingly, the value of q is limited by

$$0 \le q \le \frac{\sum\limits_{k\in\mathcal{K}} \alpha_k R_k^{\max}}{\sum\limits_{n\in\mathcal{N}} P_{0n} C_n^{\min}}. \tag{6.20}$$

We employ bi-section method (Peng et al. 2014) to search q^*, whose accuracy can be ensured after limited number of execution. The final proposed algorithm includes an outer loop to obtain the optimal E^3 using bi-section method and an inner loop to derive the optimal policy $\{\boldsymbol{w}, \boldsymbol{M}\}$ given q. The optimization problem to be solved by the inner loop is as follows.

Problem 3 (Policy Optimization in the Inner Loop)

$$\begin{aligned} &\max_{\{\mathbf{w},\mathbf{M}\}} U_R(\mathbf{w},\mathbf{M}) - q U_{PC}(\mathbf{w},\mathbf{M}) \\ &\text{s.t.} \quad C1{-}C6, \end{aligned} \tag{6.21}$$

in which the outer loop updates q iteratively. Because $q > 0$, we have the following equivalent transformation under a given q for *Problem 3*:

$$\begin{aligned} \max_{\{\mathbf{w},\mathbf{M}\}} &\sum_{k\in\mathcal{K}} \alpha_k R_k(\mathbf{w}) - q \sum_{k\in\mathcal{K},n\in\mathcal{N}} \varphi_n \left\|\mathrm{w}_{k,n}\right\|_2^2 \\ &- q \sum_{k\in\mathcal{K},n\in\mathcal{N}_1} P_{0n}\, c_{Tn}\, m_{k,n} \\ \text{s.t. } & C1{-}C6. \end{aligned} \tag{6.22}$$

Algorithm for the Inner Loop

As it can be observed, it is still difficult to solve the problem since the optimization variables ***w*** and ***M*** are highly coupled. In order to develop low-complexity algorithms, we decouple these variables as follows:

1. *The Initialization of* ***M***: Suppose that all UEs associate with RRHs initially, i.e.,

$$m_{k,n} = \begin{cases} 1, & \text{if } n \in \mathcal{N}_1, \\ 0, & \text{otherwise}, \end{cases} \tag{6.23}$$

2. *Power Allocation Optimization Given* ***M***: Given ***M***, we can transform problem (6.22) into

$$\begin{aligned} &\max_{\{\mathbf{w}\}} \sum_{k \in \mathcal{K}} \alpha_k \, R_k(\mathbf{w}) - q \sum_{k \in \mathcal{K}, n \in \mathcal{N}} \varphi_n \left\| \mathrm{w}_{k,n} \right\|_2^2 \\ &\text{s.t. } C1 - C3. \end{aligned} \tag{6.24}$$

As pointed out in Christensen et al. (2008), the E^3 maximization problem (6.24) has the same optimal solution as the following weighted minimum mean square error (WMMSE) problem:

$$\begin{aligned} &\min_{\{\mathbf{w}, \boldsymbol{\rho}, \mathbf{u}\}} \sum_{k \in \mathcal{K}} \alpha_k \cdot (\rho_k \cdot e_k - \log \rho_k) + q \cdot \sum_{k \in \mathcal{K}} \mathbf{w}_k \mathbf{J}_k \mathbf{w}_k^H \\ &\text{s.t. } C1 - C3, \end{aligned} \tag{6.25}$$

where $\boldsymbol{J}_k = \mathrm{diag}\{\varphi_1, \varphi_2, \ldots, \varphi_N\}$, ρ_k represents the MSE weight factor of user k. Given receiver $u_k \in \mathbb{C}$, the MSE e_k is calculated as

$$\begin{aligned} e_k &= \mathbb{E}\left[\left\| u_k^H \cdot y_k - x_k \right\|_2^2 \right] \\ &= u_k^H \left(\sum_{j \in \mathcal{K}} \mathbf{w}_j \, \mathbf{h}_k \, \mathbf{h}_k^H \, \mathbf{w}_j^H + \sigma^2 \right) u_k \; - 2Re\{u_k^H \mathbf{w}_k \mathbf{h}_k\} + 1. \end{aligned} \tag{6.26}$$

Under fixed ***w*** and ρ_k, the optimal u_k is given by

$$u_k = \frac{\mathbf{w}_k \, \mathbf{h}_k}{\sum\limits_{j \in \mathcal{K}} \mathbf{w}_j \, \mathbf{h}_k \, \mathbf{h}_k^H \, \mathbf{w}_j^H + \sigma^2}. \tag{6.27}$$

Under fixed ***w*** and u_k, the optimal ρ_k is given by

$$\rho_k = e_k^{-1}. \tag{6.28}$$

Problem (6.25) can be solved efficiently through the block coordinate descent method by iterating among ρ_k, u_k, and $\boldsymbol{w}_k$, and then we can get the optimal $\boldsymbol{w}$ by addressing the problem below, keeping u_k and ρ_k fixed.

$$\begin{aligned} &\min_{\{\mathbf{w}\}} \sum_{k\in\mathcal{K}} \mathbf{w}_k \mathbf{X}_k \mathbf{w}_k^H - 2\alpha_k \rho_k Re\{u_k \mathbf{w}_k \mathbf{h}_k\} \\ &\text{s.t. } C1{-}C3, \end{aligned} \tag{6.29}$$

where $\boldsymbol{X}_k$ is given by

$$\mathbf{X}_k = \sum_{j\in\mathcal{K}} \alpha_j\ \rho_j\ u_k^H\ \mathbf{h}_k\ \mathbf{h}_k^H\ u_k + q\ \mathbf{J}_k. \tag{6.30}$$

Since $q > 0$, $\boldsymbol{J}_k = \text{diag}\{\varphi_1, \varphi_2, \ldots, \varphi_N\}$, $\varphi_n > 0$, $\forall k, n$, we can observe that $\boldsymbol{X}_k$ is a positive-definite matrix. Therefore, we can conclude that problem (6.29) is QCQP and can be resolved with CVX.

3. *Modification of $\boldsymbol{M}$ Under Given $\boldsymbol{w}$*: Let $\mathcal{N}_C(k) = \{n \mid \pi_k \in \mathcal{S}_n,\ n \in \mathcal{N}_2\}$ denote the set of F-APs with the ability to serve UE k, $\mathcal{K}_R = \{k \mid \sum_{n\in\mathcal{N}_2} m_{k,n} = 0\}$ the set of UEs currently served by RRHs, and $\mathcal{K}_C = \{k \mid \mathcal{N}_C(k) \neq \emptyset,\ k \in \mathcal{K}_R\}$ the set of UEs served by F-APs. We use $E^3(\boldsymbol{w}, \boldsymbol{M})$ to denote the E^3 metric given $\boldsymbol{w}$ and $\boldsymbol{M}$. Given a pair composed of $k \in \mathcal{K}_C$ and $n \in \mathcal{N}_C(k)$, the improvement of E^3 owing to the offloading, denoted by $\Delta E^3_{(k,n)}$, can be written as

$$\Delta E^3_{(k,n)} = E^3(\mathbf{w}', \mathbf{M}') - E^3(\mathbf{w}, \mathbf{M}). \tag{6.31}$$

Particularly, when

$$m'_{i,j} = \begin{cases} m_{i,j}, & \text{if } i \neq k, \\ 1, & \text{if } i = k,\ j = n, \\ 0, & \text{otherwise}, \end{cases} \tag{6.32}$$

we can get w′ by solving the problem below:

$$\begin{aligned} &\max_{\{\mathrm{w}'\}}\ E^3(\mathrm{w}', \mathbf{M}') \\ &\text{s.t. } C1{-}C3, \\ &\qquad C7 : w'_{i,j} = w_{i,j},\ \forall i \neq k,\ j \in \mathcal{N}. \end{aligned} \tag{6.33}$$

Similar to problem (6.24), we can solve problem (6.33) by WMMSE method, but the complexity is reduced significantly.

When $\Delta E^3_{(k,n)}$ is calculated for each $k \in \mathcal{K}_C$ and $n \in \mathcal{N}_C(k)$, we select the pair that satisfies

$$(k^*, n^*) = \max_{\{k \in \mathscr{K}_C, n \in \mathscr{N}_C(k)\}} \Delta E^3_{(k,n)}. \tag{6.34}$$

If $\Delta E^3_{(k^*,n^*)} > 0$, $\boldsymbol{M}$ will be updated according to

$$m_{i,j} = \begin{cases} 1, & \text{if } i = k^*,\ j = n^*, \\ 0, & \text{if } i = k^*,\ j \neq n^*. \end{cases} \tag{6.35}$$

6.1.4 Simulation Results

In this part, an F-RAN system with $N_1 = 3$ single-antenna RRHs and $N_2 = 2$ single-antenna F-APs is considered. There are 2 UEs located in the concerned area of 0.2 km × 0.2 km. The available bandwidth is 10 MHz and the noise power spectrum density is $\sigma^2 = -174$ dBm/Hz. The channel coefficient depends on both small scale fading and distance related fading. The former is modeled as an identically distributed Rayleigh random variable while the latter is calculated by $38.5 + 40.0 * \log_{10}(d)$ in which d is the distance in meters between the transmitter and receiver. Moreover, it is assumed that $P^n_{\text{MAX}} = 25\,\text{dBm}$, $P_{0n} = 27\,\text{dBm}$, $\varphi_n = 3$, $\forall n \in \mathscr{N}$, $R^n_C = R_C$, $c_{Tn} = c_T$, $\forall n \in \mathscr{N}_1$, $S_n = S_E$, $c_{Rn} = c_{RE}$, $\forall n \in \mathscr{N}_2$, $c_P = 0.03$, $c_{O_n} = 0.7, \forall n \in \mathscr{N}$, and $c_A(\pi_k) = 1$, $\forall \pi_k \in \overline{\mathscr{S}}_C$. The probability requesting each of $F = 100$ contents is assumed to follow a Zipf distribution and the skewness parameter α is set to 1.

Impact of Fronthaul and Transmitting Method Selection

We compare three algorithms, including totally RRH strategy, F-AP involved strategy, and the proposed algorithm (ATMS). According to Fig. 6.1, in the lower R_C regime, the E^3 performance resulted from all the three algorithms improves as R_C increases. Nevertheless, the rate of increment for E^3 gradually slows down since increasing R_C cannot result in considerable throughput improvement, as shown in Fig. 6.2. In addition, we can see that our proposal always outperforms the base algorithms, verifying the effectiveness of the proposal.

Moreover, as shown in Fig. 6.2, in the higher R_C regime, the total transmitting rate of ATMS algorithm is surpassed by totally RRH algorithm. This is because the throughput gain cannot catch up with the growth of c_T and more UEs will access F-APs for a lower cost, which makes the data rate decrease due to no cooperation gain.

Impact of Content Caching

As shown in Fig. 6.3, the introduction of centralized content caching can improve the E^3 performance due to the reduction of the cost induced by acquiring the content. Moreover, with the cache cost increasing, the resulted benefit cannot be significantly improved, and hence the E^3 performance tends to saturate and then decrease as S_C keeps growing. The impact of content refreshment is shown in Figs. 6.4 and 6.5. As it can be observed, a higher η_C/η_E contributes to more reduction of cost incurred by content acquisition. However, the benefit should be balanced when c_{RC}/c_{RE} increases during E^3 optimization. In addition, compared to centralized cache, content refreshment strategies have more impact on edge cache due to the differentiation in cache size and the way of use.

6.2 Cooperative Game Based Interference Management

6.2.1 Background

As an indispensable part of F-RANs, cloud computing based C-RAN technology simplifies the RRH's functionalities, making it possible to deploy RRHs in a large scale to boost F-RAN's capacity. However, due to limited spectrum, the dense deployment of RRHs can lead to severe interference, which is a potential bottleneck

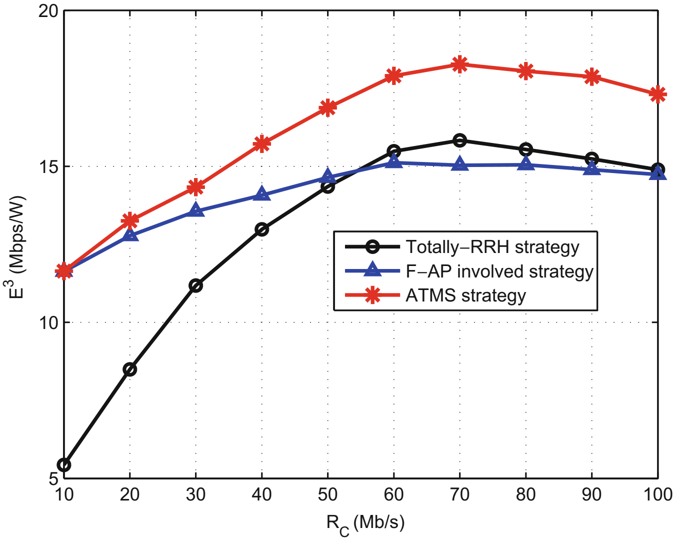


Fig. 6.1 E^3 versus fronthaul capacity constraint (R_C)

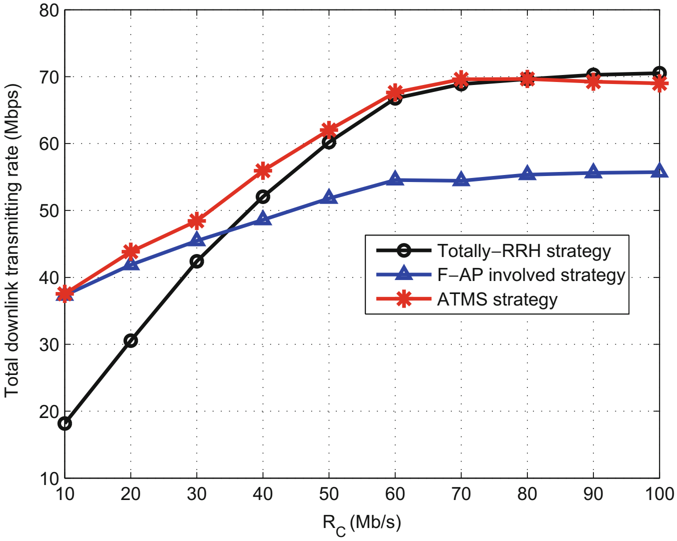


Fig. 6.2 The total transmitting rate versus fronthaul capacity constraint (R_C)

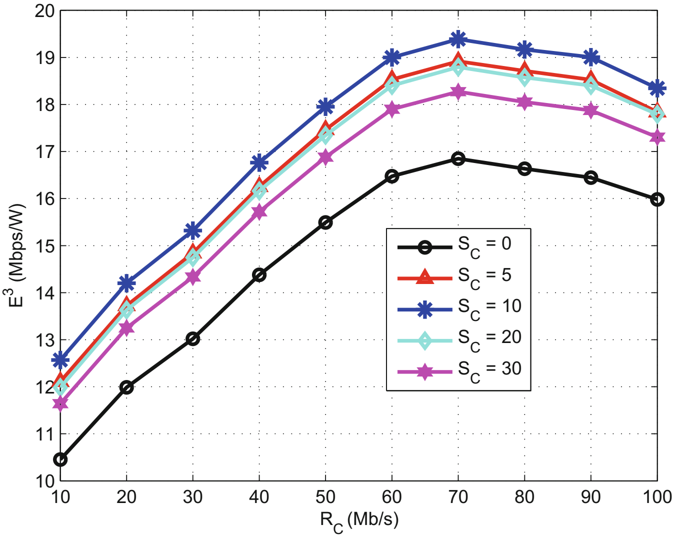


Fig. 6.3 The impact of centralized cache size (S_C) on E^3

to F-RAN's performance. In such a case, studying interference mitigation among RRHs is essential.

On the other hand, with the idea of cooperation becoming a new networking paradigm, coalition game has been recognized as a useful tool to model cooperation in communication networks (Han et al. 2012). In Zhou et al. (2014), the authors

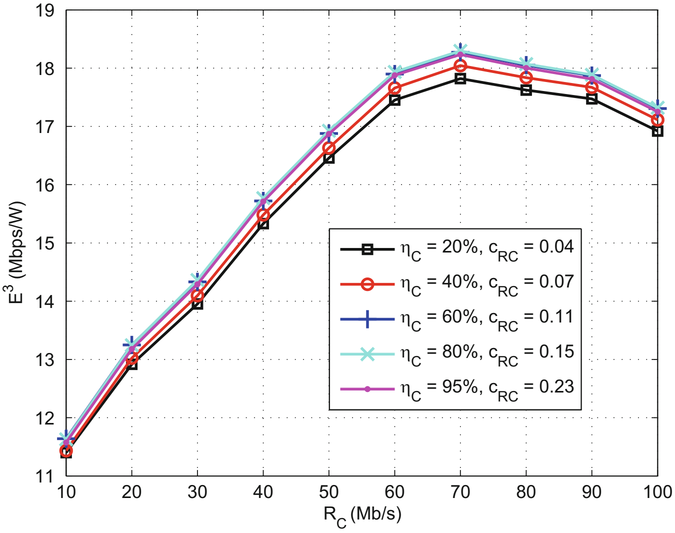


Fig. 6.4 The impact of content refreshment strategies for centralized cache

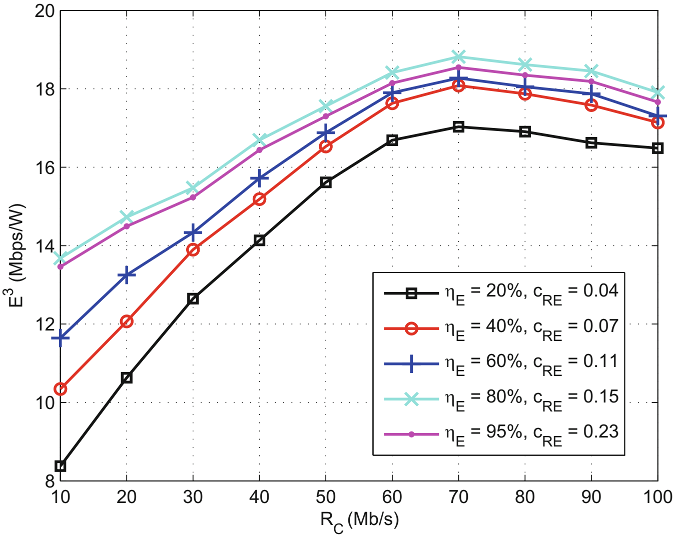


Fig. 6.5 The impact of content refreshment strategies for edge cache

propose cooperative approaches to mutual information optimization in an MIMO transmission scenario, in which coalition formation game and coalition graph formation game are adopted. In Pantisano et al. (2011), a coalitional game based interference alignment scheme is developed for intra-tier interference suppression in a femtocell network, and the authors in Ma et al. (2013) also study the coalition game based interference management for multiple access points that deliver a multimedia service. In this chapter, motivated by the previous works, a coalition formation game based RRH clustering scheme is proposed to coordinate the transmission of users within the same cluster in the time domain.

6.2.2 System Model

Consider a downlink transmission scenario with densely deployed RRHs and UEs. The cloud BBU pool connects with each RRH through an ideal fronthaul link. RRHs are all with a single antenna and transmit over the same frequency band. Initially, each RRH only serves one UE who share the same label with the associated RRH. In addition, suppose that the BBU pool has complete global channel state information (CSI).

Denote the RRH set by $\mathscr{N} = \{1, \ldots, N\}$ and denote the channel gain between RRH j and UE i by $g_{i,j}$. The received signal to interference plus noise ratio (SINR) of UE i is given by

$$\mathrm{SINR}_i = \frac{p_i A_i^T g_{i,i} A_i^R}{\sigma^2 + I_{i,\mathscr{N}} + I_0}, \tag{6.36}$$

where A_i^T and A_i^R be the transmitting antenna gain of RRH i and the receiving antenna gain of UE i, respectively, p_i is the transmission power of RRH i, σ^2 is the variance of additive Gaussian white noise, $I_{i,\mathscr{N}} = \sum\limits_{j \neq i}^{j \in \mathscr{N}} p_j A_j^T g_{i,j} A_i^R$, and I_0 is the interference from other RRHs.

The achievable Shannon rates of UE i are expressed as

$$R_i = B_i \cdot \log_2 \left(1 + \frac{p_i A_i^T g_{i,i} A_i^R}{\sigma^2 + I_{i,\mathscr{N}} + I_0} \right), \tag{6.37}$$

where B_i is the bandwidth allocated to UE i. The utility of RRH i is defined as a function of UE i's spectrum efficiency, which is given by

$$u_i(\mathscr{N}) = \frac{R_i}{B_i} = \log_2 \left(1 + \frac{p_i A_i^T g_{i,i} A_i^R}{\sigma^2 + I_{i,N} + I_0} \right). \tag{6.38}$$

From the above equation, it can be seen that co-channel interference has a negative impact on the utility of an RRH. In order to reduce such interference, RRHs can perform some inter-RRH cooperation by forming multiple coalitions.

6.2.3 Coalitional Game Formulation

In the formulated coalitional game, the set of RRHs $\mathcal{N}$ is partitioned into multiple disjoint subsets, each of which is called a coalition denoted by $S_k \subseteq \mathcal{N}$. Then, we have $\cup S_i = \mathcal{N}$ and $S_i \cap S_j = \varnothing$, $\forall i, j \in \{1, \ldots, K\}$ and $i \neq j$ with K being the number of coalitions. In each coalition, time division multiple access (TDMA) based RRH transmission is employed and $\forall i \in S_k$, its transmission time is a fraction of the normalized time slot α_i, which satisfies $\sum\limits_{i \in S_k} \alpha_i = 1, 0 < \alpha_i \leq 1$. Consequently, the interference is avoided within the coalition. However, the interference from the simultaneous transmission of other coalitions still exists.

Next, we formally model the cooperation behavior of RRHs as a coalitional game in simplified-partition form with transferable utility (TU), which is defined as follows.

Definition 6.1 A coalitional game in simplified-partition form with transferable utility is described by a pair $G = (\mathcal{N}, v)$, in which $\mathcal{N}$ represents the player set and v maps the coalition members to their total utility that is a real number denoted by $v(S, \mathcal{N})$. This reveals that the value of a coalition depends only on the members of coalition S given $\mathcal{N}$ and is independent of how $\mathcal{N} \backslash S$ is partitioned.

Assume non-coherent joint transmission is utilized within each coalition. Consequently, for every UE of RRHs in S_k, the interference caused by all RRHs in $\mathcal{N} \backslash S_k$ is independent of the specific way in which those RRHs are partitioned. The received signal strength and the suffered interference of UE i when RRH i joins a coalition can be written as

$$\begin{aligned} P_i^{\text{received}} &= p_i A_i^T g_{i,i} A_i^R + \sum_{\substack{l \neq i}}^{l \in S_k} p_l A_l^T g_{i,l} A_i^R, \\ I_{i,\mathcal{N} \backslash S_k} &= \sum_{j \in \mathcal{N} \backslash S_k} p_j A_j^T g_{i,j} A_i^R. \end{aligned} \tag{6.39}$$

The utility of RRH $i \in S_k$ is given by

$$u_i\left(S_k, \mathcal{N}\right) = \alpha_i \log_2 \left(1 + \frac{P_i^{\text{received}}}{\sigma^2 + I_{i,\mathcal{N}} S_k + I_0}\right). \tag{6.40}$$

Further, define the value of each coalition as the sum of the utility of all RRHs in the coalition, i.e.,

$$u\left(S_k, \mathcal{N}\right) = \sum_{i \in S_k} u_i\left(S_k, \mathcal{N}\right). \tag{6.41}$$

6.2.4 Coalition Formation Algorithm

Defection Order

Definition 6.2 For a given coalition composing of l players $S_k = \{k_1, k_2, \ldots, k_l\}$ and a newly formed coalition $S_k^{\text{new}} = S_k \cup \{i\}$, the defection order $S_k^{\text{new}} \supset_D S_k$ holds if the following conditions are satisfied:

$$S_k^{\text{new}} \supset_D S_k \Leftrightarrow \begin{cases} u\left(S_k^{\text{new}}, \mathscr{N}\right) > \sum\limits_{j \in S_k^{\text{new}}} u_j(\mathscr{N}) \\ \frac{u(S_k^{\text{new}}, \mathscr{N})}{|S_k^{\text{new}}|} > \frac{u(S_k, \mathscr{N})}{|S_k|} \\ u_i\left(S_k^{\text{new}}, \mathscr{N}\right) > u_i\left(S^{\text{old}}, \mathscr{N}\right), \end{cases} \tag{6.42}$$

where S^{old} is the coalition in which RRH i stays before joining coalition S_k, $|S_k| = l$, $\left|S_k^{\text{new}}\right| = l + 1$, $u_i\left(S_k^{\text{new}}, \mathscr{N}\right)$ is the utility of RRH i after joining coalition S_k^{new}, $u_i\left(S^{\text{old}}, \mathscr{N}\right)$ is RRH i's utility staying in the original coalition S^{old}, $\frac{u(S_k^{\text{new}}, \mathscr{N})}{|S_k^{\text{new}}|}$ and $\frac{u(S_k, \mathscr{N})}{|S_k|}$ are the average utility of members in S_k^{new} and S_k, respectively.

The Proposed Algorithm

Algorithm 1 Distributed algorithm for RRH coalition formation

```
Initial Stage: Each RRH transmits non-cooperatively, which means coalition partitioned is $\mathscr{H} = \{1, \ldots i, \ldots, N\}$.
Coalition Formation Stage: The cloud repeats:
for RRH $i = 1 : N$ do
  1) Discover the existing coalitions $\mathscr{H}$ and collect global information.
  2) Check every coalition $S_k$. If the defection order is met and the newly formed coalition is not included in the history $H(i)$, then record $u_i\left(S_k^{\text{new}}, \mathscr{N}\right)$ and $\frac{u(S_k^{\text{new}}, \mathscr{N})}{|S_k^{\text{new}}|}$.
  3) Let the RRH join the coalition which provides the maximum $u_i\left(S_k^{\text{new}}, \mathscr{N}\right)$ and $\frac{u(S_k^{\text{new}}. \mathscr{N})}{|S_k^{\text{new}}|}$. Add the newly formed coalition into $H(i)$, and $\mathscr{H}$ is updated.
end forUntil a stable partition is reached.
Transmission Stage: TDMA based non-coherent joint transmission is conducted in each coalition.
```

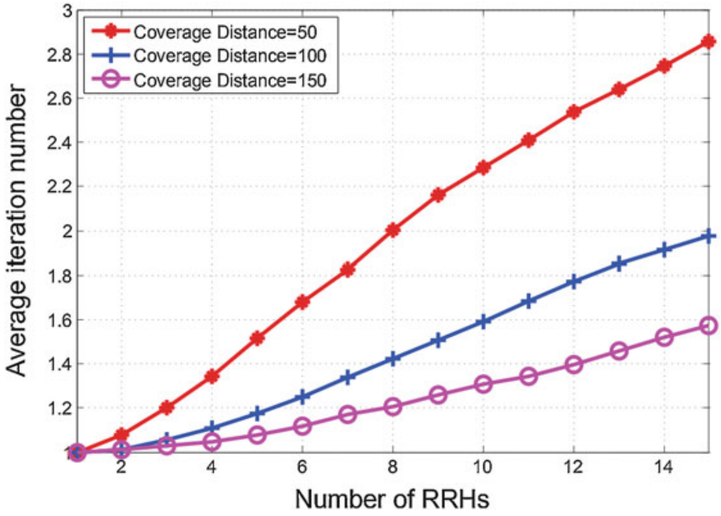


Fig. 6.6 Iterations vs. RRH number

6.2.5 Simulation Results and Analysis

In the simulation, a circular region is considered, where N RRHs are randomly deployed. The coverage radius of each RRH is set to 20 m. Moreover, suppose that $\alpha_i = \frac{1}{|S_k|}$ for each coalition S_k, which means all the RRHs within the same coalition are assigned with the same transmission time. The transmission power of each RRH is set to 30 dBm and all the antenna gains are normalized to 1. The noise power is set to $-$ 30 dBm and the pathloss model is given below.

$$\mathrm{PL(dB)} = 18.7 \times \log_{10}(\mathrm{d}) + 46.8 + 20 \times \log_{10}(2.7/5). \tag{6.43}$$

Algorithm Convergence

Figure 6.6 illustrates how the number of RRHs and coverage distance influence the complexity of the proposed coalition algorithm. Particularly, it can be observed that the iteration number increases when the coverage distance decreases. This is because smaller coverage distance generally means larger possibility that cooperation can occur and hence more coalition options are available. In addition, it can be seen that our proposed algorithm converges quickly even under a dense deployment of RRHs.

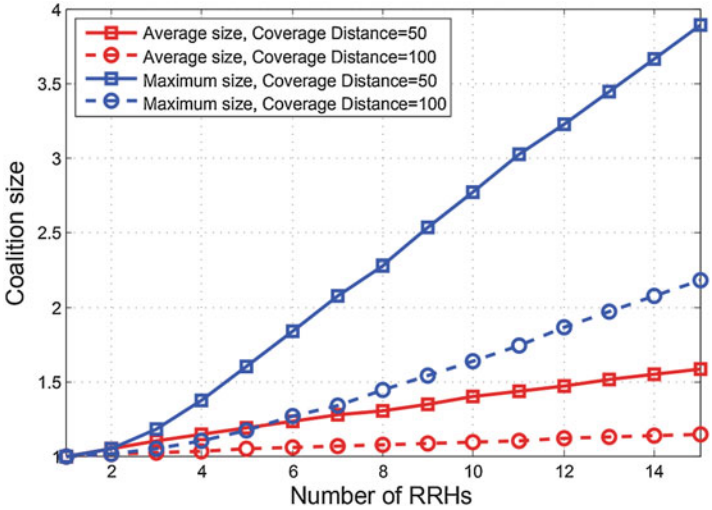


Fig. 6.7 Coalition size vs. number of RRHs

Coalition Size

In Fig. 6.7, the average and maximum coalition size under various number of RRHs are illustrated. For a fixed coverage distance, the increment of N makes both the average and maximum coalition size larger. In addition, when the number of RRHs is fixed, the coalition sizes, both in average and maximum, increase as the coverage distance decreases. This is because low RRH density causes little inter-RRH interference and hence each RRH behaves non-cooperatively. As the density of RRHs increases, the interference increases and RRHs tend to cooperate with nearby RRHs to efficiently suppress the severe interference, which leads to the increase of coalition size.

6.3 Deep Reinforcement Learning-Based Resource Management

6.3.1 Background

In F-RANs, communication mode selection plays a key role in boosting system performance, whose optimization is usually NP-hard (Yu 2014). Meanwhile, the dynamics of edge caching complicate the network environment and communication modes of UEs need to be frequently updated, making algorithms with high complexity less applicable. To facilitate the computer learn control policies directly from high-dimensional raw data, the authors in Mnih (2015) propose to combine deep learning with reinforcement learning, which is called deep reinforcement learning (DRL), and the Q-function is approximated by using a deep neural network called deep Q network (DQN). The DRL algorithm features a replay memory and a target DQN, which helps achieve stable training. In this chapter, we propose a DRL based scheme for the network controller to optimize user communication mode and network resource in a dynamic F-RAN.

6.3.2 System Model

The studied downlink F-RAN system is composed of multiple RRHs, multiple UEs, and one cloud, and each UE has a paired D2D transmitter. In the cloud, there are processors that provide computing resource for signal processing and we use D_n to denote processor n's computing capability (Liao 2017). Each processor can operate in two states, namely on state and off state, and we use $s_n^{\text{processor}} = 1$ and $s_n^{\text{processor}} = 0$ to represent them. Define the set of processors, RRHs, and UEs as $\mathscr{N} = 1, 2, \ldots, N$, $\mathscr{K} = 1, 2, \ldots, K$, and $\mathscr{M} = 1, 2, \ldots, M$, respectively. For each UE, it can operate in two different communication modes, namely D2D mode or C-RAN mode, and we use $s_m^{\text{mode}} = 1$ and $s_m^{\text{mode}} = 0$ to represent them. In C-RAN mode, the UE will be served by RRHs.

Communication Model

The received symbol of UE m in C-RAN mode is given by

$$\begin{aligned} y_m^C = & \sum_{k \in \mathscr{K}} \mathbf{h}_{m,k}^H \mathbf{v}_{m,k} x_m \\ & + \sum_{m' \in \mathscr{M}, m' \neq m, s_{m'}^{\text{mode}}=0} \sum_{k \in \mathscr{K}} \mathbf{h}_{m,k}^H \mathbf{v}_{m',k} x_{m'} + z_m, \end{aligned} \tag{6.44}$$

in which x_m represents the desired signal for UE m, $\mathbf{h}_{m,k}$ denotes the channel from RRH k, and UE m, $\mathbf{v}_{m,k}$ denotes the precoding vector of RRH k for UE m, and z_m is the noise at UE m, which is modeled by $\mathscr{CN}\left(0, \sigma^2\right)$. For UE m, its data rate is calculated as

$$R_m^C = \log\left(1 + \frac{\left|\sum\limits_{k \in \mathscr{K}} \mathbf{h}_{m,k}^H \mathbf{v}_{m,k}\right|^2}{\sum\limits_{m' \in \mathscr{M}, m' \neq m, s_{m'}^{\text{mode}}=0} \left|\sum\limits_{k \in \mathscr{K}} \mathbf{h}_{m,k}^H \mathbf{v}_{m',k}\right|^2 + \sigma^2}\right). \tag{6.45}$$

The received symbol of UE m in D2D mode is given by

$$y_m^D = \sqrt{p_m} h_m x_m + \sum_{m' \in \mathscr{M}, m' \neq m, s_{m'}^{\text{mode}}=1} \sqrt{p_{m'}} h_{m,m'} x_{m'} + z_m, \tag{6.46}$$

in which p_m represents the D2D transmission power for UE m, h_m denotes the D2D channel coefficient of UE m, $h_{m,m'}$ denotes the interfering D2D channel coefficient of UE m'. For UE m, its data rate in D2D mode is as follows:

$$R_{\mathrm{m}}^{(D)} = \log\left(1 + \frac{p_m |h_m|^2}{\sum_{m' \in \mathscr{M}, m' \neq m, s_{m'}^{\text{mode}}=1} p_{m'} |h_{m,m'}|^2 + \sigma^2}\right). \tag{6.47}$$

Computing Model

The modeling of computing resource consumption in the cloud follows that in Liao (2017). The computing resource consumption corresponding to the coding and modulation of UE m's signal is given by

$$D_{m,1} = \beta R_m, \tag{6.48}$$

where R_m denotes UE m's data rate. Moreover, to calculate the transmit signal for UE m, the cloud needs to consume computing resource given by

$$D_{m,2} = \alpha \| \mathbf{v}_m \|_0. \tag{6.49}$$

Then, overall computing resource consumed by the system is given by

$$\begin{aligned} D_{\text{system}} &= \sum_{m, s_m^{\text{mode}}=0} \left(D_{m,1} + D_{m,2}\right) \\ &= \beta \sum_{m, s_m^{\text{mode}}=0} R_m + \alpha \sum_{m, s_m^{\text{mode}}=0} \| \mathbf{v}_m \|_0. \end{aligned} \tag{6.50}$$

Caching Model

When the requested content is cached in the D2D transmitter, the value of the cache state at a D2D transmitter is defined as **True**, and the cache state is **False** otherwise. To characterize cache state dynamics, we use Markov process similar to He (2017) and the transition matrix is as below.

$$Pr_{\text{cache}} = \begin{bmatrix} Pr_{\text{True,True}} & Pr_{\text{True,False}} \\ Pr_{\text{False,True}} & Pr_{\text{False,False}} \end{bmatrix}, \tag{6.51}$$

where $Pr_{\text{True,False}}$ represents probability that the cache state at a D2D transmitter transits from **True** to **False**.

Energy Consumption Model

Following Tang (2017), the energy consumed by processor n in Watts is given by

$$E_n^{\text{processor}} = s_n^{\text{processor}} \mu D_n^3. \tag{6.52}$$

The wireless transmission power related to UE m is as follows:

$$E_m^{\text{wireless}} = \left(1 - s_m^{\text{mode}}\right) \frac{1}{\eta_1} \|\mathbf{v}_m\|_2^2 + s_m^{\text{mode}} \frac{1}{\eta_2} P_m, \tag{6.53}$$

where η_1 and η_2 denote the RRH's and UE's power amplifier efficiency, respectively (Tang 2017).

Moreover, the fronthaul energy consumption incurred by transmitting UE m's data is calculated as

$$E_m^{\text{fronthaul}} = \left(1 - s_m^{\text{mode}}\right) P_m^{\text{front}}, \tag{6.54}$$

where P_m^{front} is the energy consumption led by transmitting UE m's signal to its accessed RRHs over fronthaul.

Based on the above derivation, we can write down overall system energy consumption as follows:

$$E_{\text{system}} = \sum_n E_n^{\text{processor}} + \sum_m E_m^{\text{fronthaul}} + \sum_m E_m^{\text{wireless}}. \tag{6.55}$$

6.3.3 *Problem Formulation and Decoupling*

Problem Formulation

In the following, the system energy minimization problem is transformed into an MDP defined below:

- **State space**: The set of tuples $\mathscr{S} = \left\{\left\{\mathbf{s}^{\text{processor}}, \mathbf{s}^{\text{mode}}, \mathbf{s}^{\text{cache}}\right\}\right\}$ constitutes the state space $\mathscr{S}$. $\mathbf{s}_{\text{processor}}$ is a vector representing the current on–off states of all the processors, where the n-th element is $s_n^{\text{processor}}$. $\mathbf{s}_{\text{mode}}$ is a vector representing the current communication modes of all the UEs, where the m-th element is s_m^{mode}. $\mathbf{s}^{\text{cache}}$ is a vector containing the cache states of all D2D transmitters.

- **Action space**: The set of tuples $\mathscr{A} = \left\{\left\{a_{\text{processor}}, a_{\text{mode}}\right\}\right\}$ constitutes the action space $\mathscr{A}$. $a_{\text{processor}}$ represents turning on or turning off a certain processor. a_{mode} represents changing the communication mode of a certain UE.
- **Reward**: Here, we take the negative of system energy consumption defined in (6.55) as the reward U, and it is composed of the energy consumption led by working processors, transmission over fronthaul, and transmission over wireless channels.

After the controller takes an action using DRL, precoding is then optimized for RRH transmission by solving the following problem:

$$\begin{aligned}
&\min_{\{\mathbf{v}_m\}} \sum_{m, s_m^{\text{mode}}=0} \|\mathbf{v}_m\|_2^2 \\
&(a1)\; R_m \geq R_{m,\min}, \forall m, \\
&(a2) \sum_{m, s_m^{\text{mode}}=0} \left\|\mathbf{v}_{m,k}\right\|_2^2 \leq p_{\max}, \forall k, \\
&(a3)\; \beta \sum_{m, s_m^{\text{mode}}=0} R_m + \alpha \sum_{m, s_m^{\text{mode}}=0} \|\mathbf{v}_m\|_0 \leq \sum_n s_n^{\text{processor}} D_n,
\end{aligned} \tag{6.56}$$

where $\mathbf{v}_m$ is the network wide precoding vector for UE m.

6.3.4 DRL Based Mode Selection and Resource Management

Precoding Design with the Computing Resource Constraint

By rotating the phase of precoding (Tang 2017), constraint ($a1$) can be reformulated as a second-order cone constraint. In addition, as per Liao (2017) and Dai (2014), we can use reweighted l-1 norm to approximate the l-0 norm in constraint ($a3$). Inspired by the proposal in Dai (2014), problem (6.56) and the number of Eq. (21) should be (6.5.7) can be solved iteratively:

$$\begin{aligned}
&\min_{\{\mathbf{v}_m\}} \sum_{m, s_m^{\text{mode}}=0} \|\mathbf{v}_m\|_2^2 \\
&(e1) \sqrt{\sum_{m', s_{m'}^{\text{mode}}=0} \left|\mathbf{h}_m^H \mathbf{v}_{m'}\right|^2 + \sigma^2} \leq \sqrt{1 + \tfrac{1}{\gamma_m}} Re\left\{\mathbf{h}_m^H \mathbf{v}_m\right\}, \forall m, \\
&(e2)\; \text{Im}\left\{\mathbf{h}_m^H \mathbf{v}_m\right\} = 0, \forall m, \\
&(e3) \sum_{m, s_m^{\text{mode}}=0} \left\|\mathbf{v}_{m,k}\right\|_2^2 \leq p_{\max}, \forall k, \\
&(e4)\; \beta \sum_{m, s_m^{\text{mode}}=0} \tilde{R}_m + \\
&\quad \alpha \sum_{m, s_m^{\text{mode}}=0} \sum_k \sum_l \theta_{m,k,l} \left|v_{m,k,l}\right| \leq \sum_n s_n^{\text{processor}} D_n, \\
&(e5)\; v_{m,k,l} = 0, \textit{ if it is set to 0 in the last iteration},
\end{aligned} \tag{6.57}$$

in which $\mathbf{h}_m$ represents the channel vector between RRHs and UE m, $v_{m,k,l}$ denotes the precoding corresponding to the l-th antenna of RRH k for UE m, $\tilde{R}_m$ denotes UE m's data rate that is got from the previous iteration, $\sum\limits_{m,s_m^{\mathrm{mode}}=0} \sum\limits_k \sum\limits_l \theta_{m,k,l} \left|v_{m,k,l}\right|$ is the norm approximation of the term $\sum\limits_{m,s_m^{\mathrm{mode}}=0} \|\mathbf{v}_m\|_0$ in constraint $(a3)$. The algorithm process is summarized in Algorithm 2.

Algorithm 2 Precoding optimization with computing resource constraint

```
Stage 1:
    Precoding $v_{m,k,l}$, $\forall m$, $\forall k$, $\forall l$ is initialized by the controller that then computes the weight $\theta_{m,k,l}$ and data rate $\tilde{R}_m$.
Stage 2:
    Compute the optimal precoding by solving problem (2.57) using CVX;
    Update weight $\theta_{m,k,l}$ based on the precoding result; If $\mid v_{m,k,l} \mid$ is less than a given small threshold, $v_{m,k,l}$ is set to 0.
Stage 3:
    Repeat Stage 2 until the RRH transmit power consumption converges.
```

DRL Algorithm Design

Until now, the only task remained is to derive an MDP policy. As a classical approach to MDP, Q-learning has two drawbacks. The first one is that Q-learning faces the difficulty to store Q-values in large-scale networks and the second one is that manually defined input state is required. Fortunately, DRL proposed in Mnih (2015) can overcome these problems. Specifically, Q-values in DRL are represented by the weights of DQN and the controller with DRL is capable of directly learning from raw network data.

Here, it is assumed that the network controller uses DRL to minimize the discounted and accumulative system power consumption for T decision steps, which is given by $\mathbb{E}\left[\sum\limits_{t=0}^{T-1} \gamma^t U_t\right]$. The training procedure of the DRL model is summarized in Algorithm 3.

Algorithm 3 DRL based communication mode control and resource management

```
Stage 1:
    Randomly initialize a DQN with parameters $\mathbf{w}$ and construct the target DQN with parameters $\hat{\mathbf{w}} = \mathbf{w}$. Then, the controller randomly selects actions for a period of time to accumulate enough interaction samples for the replay memory, each of which is composed of the state transition, the action and reward.
```

```
2: Stage 2:
  For epoch $e = 0, 1, \ldots, E - 1$:
  Initialize the initial state $s_0$.
    For decision step $t = 0, 1, \ldots, T - 1$:
      A number $x$ between 0 and 1 is randomly generated. If $x \leq \varepsilon$: An action is randomly selected by the network controller.
      Else:
        The controller selects the action $a_t$ as $a_t = \arg \max_{a \in \mathscr{A}} Q(s_t, a, w)$.
      If end.
      The system state $s_t$ transits to $s_{t+1}$ according to $a_t$ and the cache state transition matrix. Optimize the precoding for UEs in C-RAN mode using Algorithm 3.
      If any UE reports a QoS violation to the HPN via the control channel, the HPN delivers the message to the controller via backhaul and the controller then executes the protecting operation.
      If end.
      The controller stores reward $U_t$ together with $s_t$, $s_{t+1}$, and $a_t$ into the replay memory.
      If the remainder when $t + 1$ is divided by $T'$ is 0:
        Randomly fetch a mini-batch of interaction samples from the replay memory, and make a single gradient update on the time-difference error with respect to $\mathbf{w}$.
      If end.
      Periodically set the value of parameters $\mathbf{w}$ to $\hat{}\ w$.
    For end.
  For end.
```

6.3.5 Simulation Results and Analysis

Consider a simulation scenario shown in Fig. 6.8. The distance between each pair of RRHs is 800 m. Each RRH is equipped with two antennas, and each UE is with one antenna. The channel coefficient of each communication link is generated by following the model composed of distance-dependent fading $distance^{-2}$, shadow fading of 8 dB, and fast fading that is modeled as $\mathscr{CN}(0, 1)$. We limit the transmit power of each RRH up to 1.5 W, while the transmit power of each D2D link is 100 mW. Moreover, we set the minimum required SINR of each UE as 5 dB.

There are 6 processors in total. Specifically, the power consumptions corresponding with these processors are 21.6 W, 6.4 W, 5 W, 8 W, 12.5 W, and 12.5 W, and their corresponding computing capabilities are 6 MOPTS, 4 MOPTS, 1 MOPTS, 2 MOPTS, 5 MOPTS, and 5 MOPTS. In addition, assume that $Pr_{\text{True,True}}$ and $Pr_{\text{False,True}}$ both equal to ρ_m for UE m.

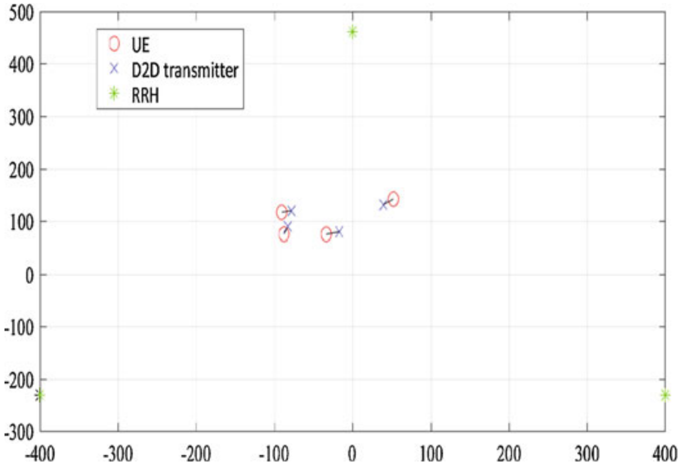


Fig. 6.8 The simulation scenario

Table 6.1 Simulation parameters

Parameter	Value	Parameter	Value
The learning rate of Adam optimizer	0.0001	RRH power efficiency	$\frac{1}{40}$
The capacity of replay memory	5000	UE power efficiency	$\frac{1}{20}$
The number of steps to update target DQN	480	Discounted factor	0.99
The number of steps to update DQN	3	Noise	10^{-13} W
The number of steps for ε linearly annealing from 1 to 0.01	3000	Fronthaul transmission power for each UE	5 W
Batch size for each DQN update	32	The initial steps to populate replay memory by random action selection	1000

As for the implemented DQN, it is based on a dense NN and contains one input layer of 14 neurons, two hidden layers of 24 neurons, and one output layer of 96 neurons. In each hidden layer, we use Relu as the activation function. Finally, all other parameters are summarized in Table 6.1.

The impacts of different batch size selection on system power consumption are evaluated in Fig. 6.9, where $\rho_m = 0.9, \forall m$, and it can be seen that the DRL algorithm achieves the best performance when the batch size is 32.

To accelerate the training process of the DRL model, the concept of transfer learning is involved. Specifically, a DRL model is first trained under $\rho_m = 0.9$ whose weights are then used for the weight initialization for the DRL model when ρ_m changes. Based on the results shown in Figs. 6.10 and 6.11, it can be seen that transfer learning-based DRL can achieve competitive performance compared to DRL learning from scratch while reducing training time significantly.

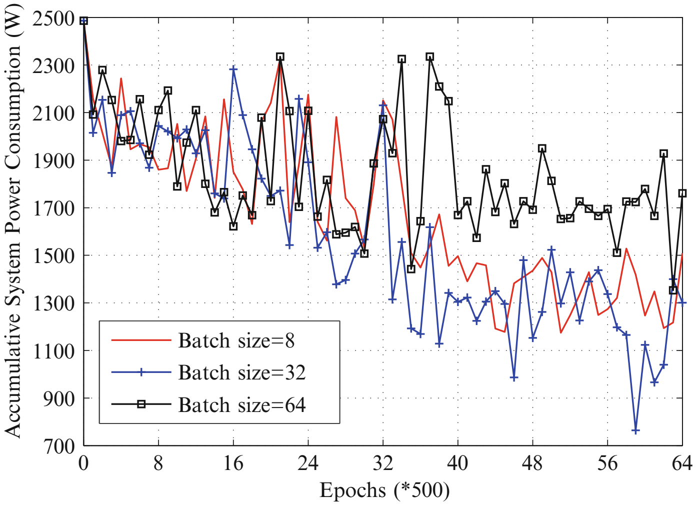


Fig. 6.9 The evolution of discounted accumulative system power consumption under different batch sizes

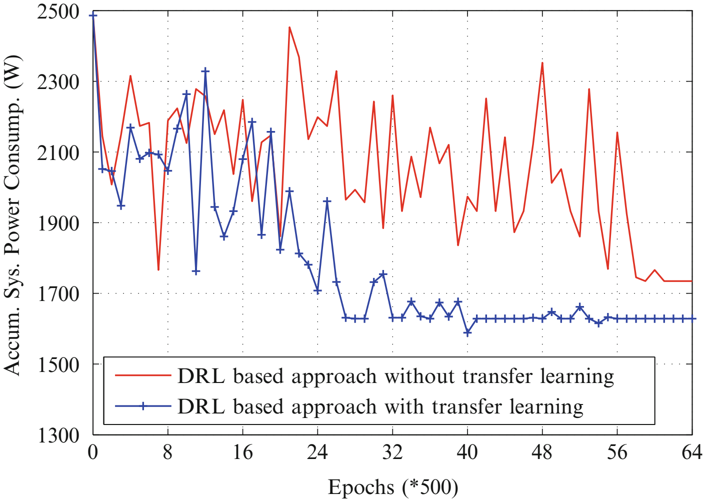


Fig. 6.10 The evolution of discounted accumulative system power consumption by transfer learning with $\rho_m = 0.6$, for all m

6.4 Summary

In this chapter, we have elaborated three different approaches to resource allocation in fog radio access networks, namely traditional centralized optimization approach, game theory-based approach, and deep reinforcement learning (DRL)

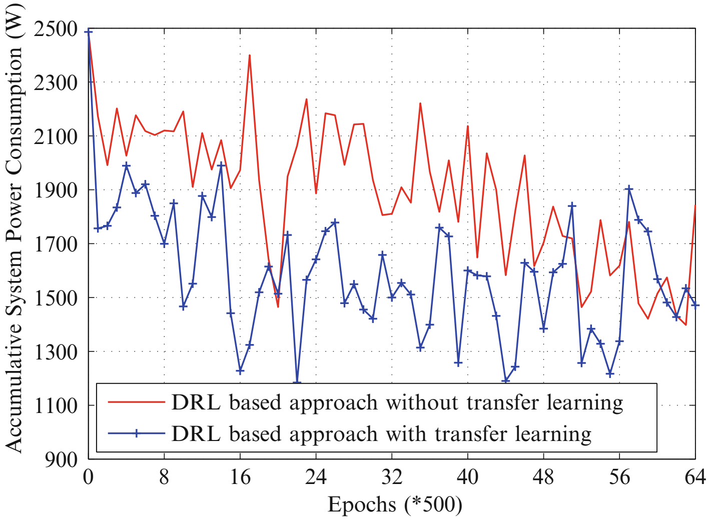


Fig. 6.11 The evolution of discounted accumulative system power consumption by transfer learning under $\rho_m = 0.75$, for all m

based approach. At the centralized cloud, traditional optimization approach has been applied to achieve high energy efficiency by optimizing network-wise precoding. In the fog computing tier, coalitional game has been adopted to mitigate intra-tier interference, hence contributing to the improvement of system throughput. For the DRL based approach, its superior performance has been demonstrated compared to various baselines in a dynamic F-RAN.

References

Christensen SS et al (2008) Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design. IEEE Trans Wirel Commun 7(12):4792–4799

Dai B (2014) Sparse beamforming and user-centric clustering for downlink cloud radio access network. IEEE Access 2:1326–1339

Han Z et al (2012) Game theory in wireless and communication networks: theory, models, and applications. Cambridge University Press, Cambridge. https://books.google.com/books?id=oSnz5ngb9YkC

He Y (2017) Deep reinforcement learning-based optimization for cache-enabled opportunistic interference alignment wireless networks. IEEE Trans Veh Technol 66(11):10433–10445

Liao Y (2017) How much computing capability is enough to run a cloud radio access network? IEEE Commun Lett 21(1):104–107

Ma B et al (2013) Interference management for multimedia femtocell networks with coalition formation game. In: 2013 IEEE international conference on communications (ICC). ACM, New York, pp 6112–6117

Mnih V (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533

Pantisano F et al (2011) Cooperative interference alignment in femtocell networks. In: 2011 IEEE global telecommunications conference - GLOBECOM 2011. ACM, New York, pp 1–6

Peng M et al (2014) Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies. IEEE Wirel Commun 21(6):126–135

Peng M et al (2015) Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. IEEE Trans Veh Technol 64(11):5275–5287

Peng M et al (2016) Fog-computing-based radio access networks: issues and challenges. IEEE Netw 30(4):46–53

Tang J (2017) System cost minimization in cloud ran with limited fronthaul capacity. IEEE Trans Wirel Commun 16(5):3371–3384

Wang X et al (2014) Cache in the air: exploiting content caching and delivery techniques for 5G systems. IEEE Commun Mag 52(2):131–139

Yan Z et al (2017) Economical energy efficiency: an advanced performance metric for 5G systems. IEEE Wirel Commun 24(1):32–37

Yu G (2014) Joint mode selection and resource allocation for device-to-device communications. IEEE Trans Commun 62(11):3814–3824

Zhou T et al (2014) Network formation games in cooperative MIMO interference systems. IEEE Trans Wirel Commun 13(2):1140–1152

Chapter 7
Content Caching in Fog Radio Access Networks

The rest of this chapter is organized as follows: A hierarchical cooperative content caching framework in F-RANs is introduced in Sect. 7.1. In Sect. 7.2, the content pushing and delivering schemes in F-RANs are studied, and the corresponding analytical performance is provided in Sect. 7.3. Section 7.4 studies joint optimization of cache management and resource allocation. The computation complexity analysis and simulation results are given in Sect. 7.5, followed by the conclusion in Sect. 7.6.

7.1 Hierarchical Cooperative Content Caching Framework in F-RANs

As a mainstream category of wireless service, multimedia delivery is with strict QoS requirements, and causes heavy loadings in F-RANs, since the data volume is large, especially for the high definition videos. To reduce the transmission latency and balance the loadings of fronthaul links, the content caches are deployed in F-RANs, which can be employed to keep the popular content objects locally.

Due to the hierarchical cloud and fog computing-based architecture, content caches can be equipped with both the cloud computing centers and edge equipment. Therefore, the caches can be with the cloud computing center, the F-APs, and the users. These three different categories of content caches have different tradeoff between the storage volume and the content deliver latency: The caches with the cloud computing center are with the large caching volume, but the content objects are delivered to the users through two hops, and the transport latency is long. When the content is with the F-AP cache, the latency caused by transport through fronthaul can be avoided, but its caching volume is smaller than the cloud cache. When the content is with the user cache, it can be served immediately when the user wants it, but the storage volume of user cache is quite limited.

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0_7

When the content objects are kept by the cloud and F-AP caches, they can be accessed via wireless channels, and the corresponding performance gains have been studied by deriving effective capacity in Sect. 3.2, Chap. 3. In this chapter, we mainly focus on the utility of user device caches. Moreover, we will discuss how to balance the tradeoff between the costs of content pushing and the utility of content caching, to ensure that the performance gains of content caching can be achieved in a cost-efficient way.

7.2 Content Pushing and Delivering Schemes in F-RANs

To ensure that the user requests can be responded locally, it requires that the contents objects should be pushed into the content caches proactively. Therefore, the content pushing and delivering procedures should be considered jointly, which are studied in this part.

7.2.1 System Model

Consider a content delivering scenario in F-RANs, where a single F-AP serves multiple users. We assume that the users locate in a disc region $D(B_{\mathrm{T}}, r)$, where B_{T} denotes its center, and r is the radius. The locations of users are modeled as a Matérn cluster process Ψ_{T} (Stoyan et al. 1995), and the number of users follows Poisson distribution, i.e., $L_T \sim$Pois(λ_{T}), λ_{T} denotes the density of users. In this part, all the content objects are assumed be with the same data volume, which can be denoted as K. Due to the status of user request, all the content objects are classified as follows:

- **Active content objects:** The active content objects are the content objects that are requested currently. For simplicity, it is assumed that the number of active content objects is finite, which can be denoted as a set $\Omega_a = \{x_1, \ldots, x_{M_a}\}$, and M_a is the number of active content objects of the current frame.
- **Proactive content objects:** Unlike active content objects, proactive content objects are not required currently. However, the popularity of content objects changes dynamically, and thus they may be requested in the future. The number of proactive content objects is also finite, which can be denoted as M_p. All these proactive content objects are denoted as a set Ω_a, $\Omega_p = \{y_1, \ldots, y_{M_p}\}$.

Similar to Chap. 3, a hierarchical content caching architecture is considered in this part. An F-AP cache Λ_{T} is equipped with B_{T}, while a user cache Π_k is equipped with U_k. Moreover, both the active and proactive content objects are kept by the F-AP cache Λ_{T}. In the user cache Π_k, only the active content objects are cached. Therefore, the size of Π_j is assumed to be $M_a K$, and all the active content objects can be kept by Π_j. To serve multiple users simultaneously, a multicast-based content

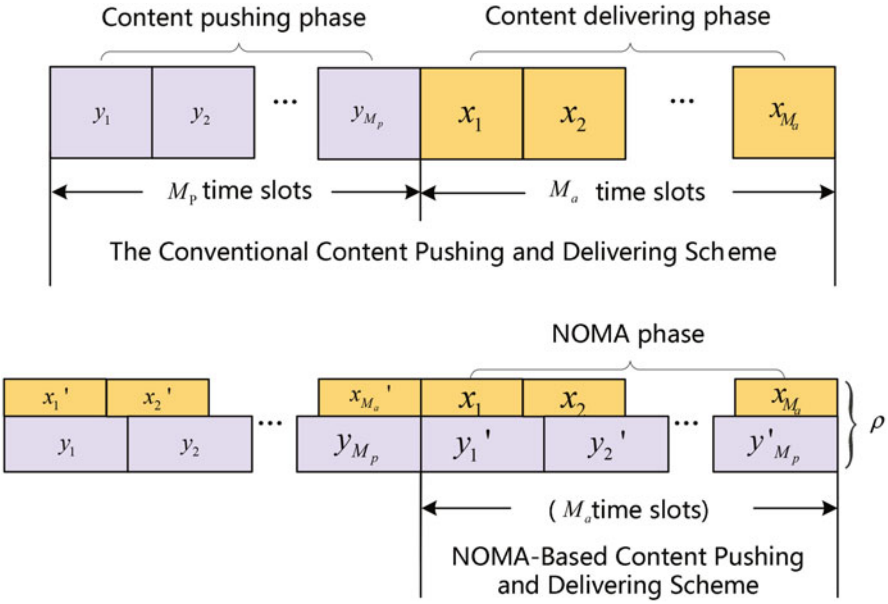


Fig. 7.1 The frame structures of OMA-MC and NOMA-MC schemes

delivering strategy is implemented, and thus the spectrum efficiency of F-RANs can be improved.

7.2.2 The Conventional Content Pushing and Delivering Scheme

In the conventional content pushing and delivering scheme, the delivering of each content object occupies an orthogonal radio resource block, which is illustrated in Fig. 7.1. Therefore, it consists of two phases, which are content pushing phase and delivering phase, respectively. During the content pushing phase, all M_p proactive content objects are pushed to the user caches via orthogonal time slots. During the content delivering phase, all the active content objects of Ω_a are multicasted to the users, and M_a orthogonal time slots are used.

Since the popularity of content objects is correlated in time domain, only a part of active content objects will be turned into proactive status in the next frame. In this part, we assume that M_p active content objects will become proactive content objects in the next frame, while all M_p proactive content objects will be activated. The number of active and proactive content objects follows the constraint $M_p \leqslant M_a$.

7.2.3 NOMA-Based Content Pushing and Delivering Scheme

Unlike the conventional scheme, the procedures of content pushing and delivering can be accomplished simultaneously by employing NOMA transmission techniques. The key idea of NOMA is to transmit two messages simultaneously, which can be detected successfully via SIC, as long as sophisticated power allocation strategy is employed. Therefore, during the m-th time slot, the active and proactive content objects can be mixed as a NOMA message, which can be expressed as follows:

$$s_m = \sqrt{P_a} x_m + \sqrt{P_p} y_{i_m},\ x_m \in \Omega_a, y_{i_m} \in \Omega_p, \tag{7.1}$$

where P_a and P_p are the transmit power of active and proactive content objects, respectively. Then, the observation of U_k can be expressed as

$$\begin{aligned} t_k &= r_k^{-\beta/2} g_k s_m + \sigma_k \\ &= r_k^{-\beta/2} g_k \big(\sqrt{P_a} x_m + \sqrt{P_p} y_{i_m}\big) + \sigma_k, \end{aligned} \tag{7.2}$$

where r_k is the distance between U_k and B_{T}, β denotes the pathloss exponent, g_k is the channel fading coefficient with respect to the wireless channel between U_k and B_{T}, and σ_k denotes the additive Gaussian noise with fixed power ω^2. In this paper, flat Rayleigh fading channel model is employed, i.e., $g_k \sim \mathscr{CN}(0, 1)$.

Since the number of proactive content objects is smaller than that of active content objects, the data rate of y_{i_m} should be lower than that of x_m. To guarantee the performance of NOMA transmission, y_{i_m} should be detected at first, and its corresponding received SINR is written as

$$\nu_p = \frac{r_k^{-\beta} |g_k|^2 \gamma_p}{r_k^{-\beta} |g_k|^2 \gamma_a + 1}, \tag{7.3}$$

where $\gamma_a = P_a/\omega^2$ is SNR of x_m, and $\gamma_p = P_p/\omega^2$ is defined similarly for y_{i_m}. To remove the interference caused by y_{i_m}, it should be eliminated via SIC before detecting x_m, and the transmit power should follow the constraint $P_p \geqslant P_a$. If y_{i_m} can be detected successfully, the received SNR of x_m is expressed as

$$\nu_a = r_k^{-\beta} |g_k|^2 \gamma_a. \tag{7.4}$$

Compared with the conventional scheme, the extra time slots occupied by content pushing can be avoided in NOMA-based content pushing and delivering schemes, since both the proactive content objects are pushed with the active content objects. Therefore, the spectrum efficiency of F-RANs can be improved significantly, especially when the number of proactive content objects is large.

7.3 The Performance of Content Pushing and Delivering in F-RANs

In this part, we mainly focus on the outage performance of NOMA-based content pushing and delivering scheme. First, a tractable expression of outage probability can be provided. Then, some asymptotic analysis results are derived to provide some insights.

7.3.1 The Outage Probability of NOMA-Based Scheme

In each frame of NOMA transmissions, it can be assumed that each user requests only one active content object. Without loss of generality, we focus on the multicasting of active content object x_m, which was also pushed in the previous frame. Therefore, the outage probability of x_m can be expressed as

$$\vartheta_{x_m}^{\mathrm{o}} = \vartheta_{x_m,p}^{\mathrm{o}} \vartheta_{x_m,a}^{\mathrm{o}}, \tag{7.5}$$

where $\vartheta_{x_m,p}^{\mathrm{o}}$ and $\vartheta_{x_m,a}^{\mathrm{o}}$ are the outage probability of pushing and delivering x_m in the previous and current frames, respectively.

Then, $\vartheta_{x_m}^{\mathrm{o}}$ can be further derived as follows based on (7.3)–(7.5):

$$\begin{aligned}\vartheta_{x_m}^{\mathrm{o,s}} &= \vartheta_{x_m,p}^{\mathrm{o,s}} \vartheta_{x_m,a}^{\mathrm{o,s}} \\ &= \Pr\left\{\nu_{p,1} \leqslant \xi_p\right\}\left(1 - \Pr\left\{\nu_{p,2} > \xi_p, \nu_a > \xi_a\right\}\right),\end{aligned} \tag{7.6}$$

where $\nu_{p,1}$ is the received SINR of pushing x_m in the previous frame, $\nu_{p,2}$ is defined similarly for y_n, ν_a is the received SNR of multicasting x_m in the current NOMA phase, ξ_p and ξ_a are the SINR and SNR thresholds with respect to $\nu_{p,1}$, $\nu_{p,2}$, and ν_a, respectively. Then, a tractable expression of $\vartheta_{x_m}^{\mathrm{o,s}}$ can be obtained, which is provided in the following theorem.

Theorem 7.1 *The outage probability of NOMA-based content pushing and delivering scheme can be written as*

$$\vartheta_{x_m}^{\mathrm{o,s}} = F\left(\frac{\xi_p}{\varphi_p}\right) \max\left\{F\left(\frac{\xi_p}{\varphi_p}\right), F\left(\frac{\xi_a}{P_a}\right)\right\}, \tag{7.7}$$

where $\varphi_p = P_p - \xi_p P_a$, *it follows the constraint* $\varphi_p > 0$, *and* $G(u)$ *is given as*

$$F\left(u\right) = 1 - \frac{2}{\beta d^2 \omega^{4/\beta} u^{2/\beta}} \nu\left(\frac{2}{\beta}, \omega^2 d^\beta u\right), \tag{7.8}$$

$\nu(a, x)$ *is the lower incomplete gamma function.*

Proof Based on (7.6), we have to derive $\vartheta_{x_m,p}^{\mathrm{o,s}}$ and $\vartheta_{x_m,a}^{\mathrm{o,s}}$ to obtain a closed-form expression of outage probability. By substituting (7.3) into (7.6), $\vartheta_{x_m,p}^{\mathrm{o,s}}$ is derived as follows:

$$\vartheta_{x_m,p}^{\mathrm{o,s}} = \Pr\left\{|g_k|^2 \leqslant \frac{\xi_p\omega^2}{\varphi_p} r_k^{\beta}\right\} = 1 - \underbrace{\mathbb{E}_{r_k}\left\{\exp\left(-\frac{\xi_p\omega^2}{\varphi_p} r_k^{\beta}\right)\right\}}_{\mathscr{I}_1}.$$

Please note that the locations of the users follow independently identical distribution, and the PDF of the distance between U_k and B_T is expressed as

$$g(r_k) = \tfrac{2r_k}{d^2}. \tag{7.9}$$

By substituting (7.9) into $\mathscr{I}_1$ given by (7.9), it can be expressed as

$$\begin{aligned}\mathscr{I}_1 &= \int_0^r \frac{2r_k}{d^2}\exp\left(-\frac{\xi_p\omega^2}{\varphi_p} r_k^{\beta}\right)\mathrm{d}r_k \\ &= \frac{2}{\beta d^2}\left(\frac{\varphi_p}{\xi_p\omega^2}\right)^{2/\beta}\nu\left(\frac{2}{\beta}, d^{\beta}\omega^2\frac{\xi_p}{\varphi_p}\right).\end{aligned} \tag{7.10}$$

In particular, $\vartheta_{x_m,p}^{\mathrm{o,s}} = 1$ when $\varphi_p \leqslant 0$.

Next, a closed-form expression of $\vartheta_{x_m,a}^{\mathrm{o,s}}$, which is given by (7.6), should be derived. It can be expressed as

$$\begin{aligned}\vartheta_{x_m,a}^{\mathrm{o,s}} &= 1 - \min\left\{\Pr\left\{|g_k|^2 > \frac{\xi_p\omega^2}{\varphi_p} r_k^{\beta}\right\},\ \Pr\left\{|g_k|^2 > \frac{\xi_a\omega^2}{P_a} r_k^{\beta}\right\}\right\} \\ &= \max\left\{1 - \Pr\left\{|g_k|^2 > \frac{\xi_p\omega^2}{\varphi_p} r_k^{\beta}\right\},\ 1 - \Pr\left\{|g_k|^2 > \frac{\xi_a\omega^2}{P_a} r_k^{\beta}\right\}\right\}.\end{aligned} \tag{7.11}$$

Then $\vartheta_{x_m,a}^{\mathrm{o,s}}$ can be derived similarly to $\vartheta_{x_m,p}^{\mathrm{o,s}}$, and the proof has been finished.

7.3.2 Further Discussion of Theorem 7.1

Recalling (7.7), to minimize the outage probability of NOMA-based scheme, the transmit power allocation of active and proactive content objects should be optimized, which can be formulated as following optimization problem:

$$\min\ \vartheta_{x_m}^{\mathrm{o,s}}(P_p, P_a) = F\left(\frac{\xi_p}{\varphi_p}\right)\max\left\{F\left(\frac{\xi_p}{\varphi_p}\right), F\left(\frac{\xi_a}{P_a}\right)\right\}, \tag{7.12a}$$

$$\text{s.t. } P_p - \xi_p P_a > 0, \tag{7.12b}$$

$$P_p + P_a \leqslant P,\ \ P_p, P_a \geqslant 0, \tag{7.12c}$$

where P is the maximum transmit power of F-AP, and the constraint given by (7.12b) is to guarantee $\varphi_p > 0$ in (7.7).

To obtain the solution of (7.12), we first verify the monotonicity of $G(u)$ given by (7.8). Its derivative can be expressed as

$$\begin{aligned}\nabla_u F(u) = \tfrac{4}{\beta^2 d^2 (\omega^2 u)^{2/\beta+1}} & \left[\nu\left(\frac{2}{\beta}, \omega^2 d^\beta u\right) \right. \\ & \left. + \frac{\beta}{2}(\omega^2 u)^{2/\beta} d^\beta \exp(-\omega^2 d^\beta u) \right] > 0. \end{aligned} \tag{7.13}$$

(7.13) indicates that $G(u)$ keeps increasing as u increases. Then, (7.12) can be transformed into the following two optimization problems.

1. *When* $P_p \leqslant \frac{\xi_p}{\xi_a}(\xi_a + 1)P_a$: In this case, the received SINR in (7.12a) follows the constraint $\xi_p/\varphi_p \geqslant \xi_a/P_a$. Due to the monotonicity, the following relationship between $G(\frac{\xi_p}{\varphi_p})$ and $G(\frac{\xi_a}{P_a})$ can be obtained:

$$F\left(\tfrac{\xi_p}{\varphi_p}\right) \geqslant F\left(\tfrac{\xi_a}{P_a}\right). \tag{7.14}$$

Based on (7.14), the optimization problem given by (7.12) can be transformed as

$$\min F\left(\frac{\xi_p}{\varphi_p}\right)^2, \tag{7.15a}$$

$$\text{s.t. } P_p \leqslant \frac{\xi_p}{\xi_a}(\xi_a + 1)P_a,\ (7.12b),\ (7.12c). \tag{7.15b}$$

The objective function given by (7.15a) can be transformed as the following equivalent form:

$$\max\ P_p - \xi_p P_a. \tag{7.16}$$

Therefore, it is a linear program problem, and its optimal solution is

$$P^*_{p,1} = \tfrac{\xi_p + \xi_a \xi_p}{\xi_p + \xi_a + \xi_a \xi_p} P,\ P^*_{a,1} = \tfrac{\xi_a}{\xi_p + \xi_a + \xi_a \xi_p} P. \tag{7.17}$$

2. *When* $P_p \geqslant \frac{\xi_p}{\xi_a}(\xi_a + 1)P_a$: It ensures that $G(\xi_p/\varphi_p) \leqslant G(\xi_a/P_a)$ in this case, and (7.12) is rewritten as

$$\min \ F\left(\frac{\xi_p}{\varphi_p}\right)F\left(\frac{\xi_a}{P_a}\right), \tag{7.18a}$$

$$\text{s.t.} \ \ P_p \geqslant \frac{\xi_p}{\xi_a}(\xi_a+1)P_a, \ (7.12c). \tag{7.18b}$$

To solve (7.18) efficiently, it can be transformed as the following theorem.

Theorem 7.2 *An equivalent form of* (7.18) *can be provided as follows:*

$$\min \ F\left(\frac{\xi_p}{P-(\xi_p+1)P_a}\right)F\left(\frac{\xi_a}{P_a}\right), \tag{7.19a}$$

$$\text{s.t.} \ \ 0 \leqslant P_a \leqslant \frac{\xi_a}{\xi_p+\xi_a+\xi_a\xi_p}P. \tag{7.19b}$$

Proof The feasible region of (7.18) is convex, and thus its optimal solution locates at the boundary of (7.18b). It is proved by contradiction. In particular, we assume that the optimal solution $(P^*_{p,2}, P^*_{a,2})$ is in the feasible region $\mathscr{F}$. Then, another point locating on the boundary of $\mathscr{F}$ can be searched, i.e., $(P^*_{p,2}+\triangle P, P^*_{a,2})$. It ensures that the following inequality can be formulated based on the monotonicity of $G(u)$:

$$F\left(\frac{\xi_p}{\varphi_p(P^*_{p,2}+\triangle P, P^*_{a,2})}\right)F\left(\frac{\xi_a}{P^*_{a,2}}\right) < F\left(\frac{\xi_p}{\varphi_p(P^*_{p,2}, P^*_{a,2})}\right)F\left(\frac{\xi_a}{P^*_{a,2}}\right). \tag{7.20}$$

However, it is contradictory to the optimality assumption of $(P^*_{p,2}, P^*_{a,2})$.

In particular, when $(P^*_{p,2}, P^*_{a,2})$ locates on the boundary $\mathscr{B}_1 = \{(P_p, P_a)|P_p = \frac{\xi_p}{\xi_a}(\xi_a+1)P_a\}$, the objective function of (7.18) can be expressed as follows due to the relationship $P_p = \frac{\xi_p}{\xi_a}(\xi_a+1)P_a$:

$$\min \ F\left(\frac{\xi_a\omega^2}{P_a}\right)^2. \tag{7.21}$$

As introduced previously, $(P^*_{p,1}, P^*_{a,1})$ given by (7.17) is the optimal solution. However, $(P^*_{p,1}, P^*_{a,1})$ locates on another boundary of $\mathscr{F}$, which can be denoted as $\mathscr{B}_2 = \{(P_p, P_a)|P_p + P_a = P\}$. The optimal solution of (7.18) is on $\mathscr{B}_2$, Therefore, (7.18) can be rewritten as (7.19) based on the relationship $P_p + P_a = P$. The theorem has been proved.

As shown in (7.19), it is a programming problem with respect to a single variant, whose constraint is a linear function. Its optimal solution can be obtained straightforwardly by employing descent methods (Bazaraa et al. 1979).

7.3.3 Asymptotic Analysis of Outage Probability

The outage probability $\vartheta_{x_m}^{\mathrm{o,s}}(P_p^*, P_a^*)$ can be upper bounded as follows:

$$\vartheta_{x_m}^{\mathrm{o,s}}(P_p^*, P_a^*) \leqslant \vartheta_{x_m}^{\mathrm{o,s}}(P_{p,1}^*, P_{a,1}^*) = F\left(\frac{\xi_p + \xi_a + \xi_a\xi_p}{\gamma}\right)^2, \tag{7.22}$$

where $\gamma = P/\omega^2$. When the average SNR γ goes infinity, an asymptotic analysis result of $\vartheta_j^{\mathrm{o,s}}(P_{p,1}^*, P_{a,1}^*)$ can be provided as follows.

Theorem 7.3 *The outage probability of NOMA-based content pushing and delivering scheme can be derived as follows when γ goes infinity:*

$$\begin{aligned}\vartheta_{x_m}^{\mathrm{o,s}}(P_p^*, P_a^*) &\leqslant \vartheta_{x_m}^{\mathrm{o,s}}(P_{p,1}^*, P_{a,1}^*) \\ &\overset{\gamma\to\infty}{=} \left(\frac{(\xi_p + \xi_a + \xi_a\xi_p)d^\beta}{\gamma}\right)^2.\end{aligned} \tag{7.23}$$

Proof When γ goes infinity, the argument $(\xi_p + \xi_a + \xi_a\xi_p)/\gamma$ with respect to $G(\cdot)$ in (7.22) approaches 0. As introduced in Gradshteyn et al. (2000), the lower complete gamma function can be expanded as a series of power functions. Then, $\vartheta_{x_m}^{\mathrm{o,s}}(P_{p,1}^*, P_{a,1}^*)$ given by (7.22) can be approximated as

$$\begin{aligned}&\vartheta_{x_m}^{\mathrm{o,s}}(P_{p,1}^*, P_{a,1}^*) \\ &\overset{\gamma\to\infty}{=} \quad \left[1 - \exp\left(-\frac{(\xi_p + \xi_a + \xi_a\xi_p)d^\beta}{\gamma}\right)\right]^2 \\ &\overset{\gamma\to\infty}{=} \quad \left(\frac{(\xi_p + \xi_a + \xi_a\xi_p)d^\beta}{\gamma}\right)^2.\end{aligned} \tag{7.24}$$

And the proof has been finished.

As shown in (7.23), the diversity gain of NOMA-based content pushing and delivering scheme is 2. Therefore, it can achieve the full diversity gains, since each content object is transmitted twice as introduced previously. Moreover, the co-channel interference can be mitigated via the optimization of power allocation.

7.4 Joint Optimization of Cache Management and Resource Allocation

Since the content objects are with different popularity, it should be considered when we study the management of caches and radio resource. Moreover, the cache management and resource allocation are coupled tightly, and thus they should be optimized jointly in F-RANs, which can further improve the performance of

NOMA-based content pushing and caching. However, it is a NP-hard problem. To provide an efficient method to solve it, we decouple it as two independent subproblems, and then a distributed algorithm is designed based on matching theory.

7.4.1 Problem Formulation

Since the users require different content objects, they can be divided into M_a independent groups. Each of them can be denoted as a set, i.e., $\Psi'_1, \ldots, \Psi'_{M_a}$, where Ψ'_m denotes the set that consists of all the users requesting x_m currently, $\Psi'_m \subset \Psi_\mathrm{T}$. Ψ'_m is a random thinning process with respect to Ψ_T. Therefore, the number of users belonging to Ψ'_m also follows Poisson distribution, and its density can be given as $\rho_m^a \lambda_\mathrm{T}$, i.e., $L_m \sim \mathrm{Pois}(\rho_m^a \lambda_\mathrm{T})$, ρ_m^a denotes the popularity of x_m. Moreover, $\rho_{i_m}^p$ is defined similarly, which denotes the popularity of proactive content object y_{i_m}, and the number of users that will request y_{i_m} in the next frame also follows Poisson distribution, i.e., $L'_{y_{i_m}} \sim \mathrm{Pois}(\rho_{i_m}^p \lambda_\mathrm{T})$. Then the average success transmission probabilities of x_m and y_{i_m} can be expressed as follows (7.5):

$$\begin{aligned} \delta_{x_m} &= \rho_m^a \left(1 - \vartheta_{x_m}^\mathrm{o}\right) = \rho_m^a \left(1 - \vartheta_{x_m,p}^\mathrm{o} \vartheta_{x_m,a}^\mathrm{o}\right), \\ \delta_{y_{i_m}} &= \rho_{i_m}^p \left(1 - \vartheta_{y_{i_m}}^p\right), \end{aligned} \tag{7.25}$$

where ϑ_{x_m} denotes the average successful transmission probability of x_m, and $\vartheta_{y_{i_m}}$ is defined similarly for y_{i_m}.

In this part, we aim to maximize the successful transmission probability, and thus the objective function is defined as follows:

$$\bar{\delta}_\mathrm{T} = \sum_{m=1}^{M_a} \bar{\delta}_m = \sum_{m=1}^{M_a} \left(\delta_{x_m} + \delta_{y_{i_m}}\right). \tag{7.26}$$

As illustrated in Fig. 7.1, it spends $\frac{M_a}{M_p} T$ to transmit each proactive content object, where T denotes the length of one time slot. Please note the popularity of content objects is taken into account for the optimization of power allocation, which is different from the discussions in Sect. 7.3.2. Moreover, it is challenging to optimize such a dynamic power allocation scheme, since each active content object may be delivered with two different proactive content objects, i.e., x_2 in Fig. 7.1. To solve this problem, we divide each proactive content object into M_a packets. Hence, in each time slot, M_p individual packets of proactive content objects are transmitted with x_m, and the successful transmission probability $\bar{\vartheta}_m$ in (7.26) is derived as

$$\bar{\delta}_m \left(P_{a,m}, P_{p,m}, \{x_m, \Gamma_m\}\right) = \delta_{x_m} + \bar{\delta}_{\Gamma_m}, \tag{7.27}$$

where $P_{a,m}$ is the transmit power of active content object x_m, and $P_{p,m}$ is defined similarly for the selected packets of proactive content objects in the m-th time slot, Γ_m denotes the set consisting of all the selected proactive content packets, i.e., $\Gamma_m =$

$\{y_{i_1,j_1}, \ldots, y_{i_{M_p},j_{M_p}}\}$, the j_m-th packet of proactive content object y_{i_m} is y_{i_m,j_m}, and the successful transmission probability of all y_{i_m}-s in Γ_m is $\bar{\delta}_{\Gamma_m}$.

As shown in (7.27), the successful transmission probability is mainly determined by power allocation and content management, i.e., the matching of active content objects and proactive content packets, which is called content matching. Therefore, the optimization problem can be established as follows:

$$\max \;\; \bar{\delta}_{\mathrm{T}} = \sum_{m=1}^{M_a} \bar{\delta}_m \big(P_{a,m},\, P_{p,m},\, \{x_m,\, \Gamma_m\}\big), \tag{7.28a}$$

$$\text{s.t.} \;\; (7.12b),\;\; (7.12c). \tag{7.28b}$$

The optimization problem given by (7.28) is neither convex nor linear. In particular, the optimization of content matching is a combinatorial optimization problem, whose objective function is non-linear. As introduced in Papadimitriou et al. (1998), it is a NP-hard problem. To provide an efficient solution, (7.28) is decoupled as the following two independent subproblems.

(1) Power Allocation Subproblem For a pair of active content object and set of selected proactive content packets, i.e., x_m and Γ_m, the subproblem of power allocation can be given as follows:

$$\max \;\; \bar{\delta}_m \big(P_{a,m},\, P_{p,m}\big), \tag{7.29a}$$

$$\text{s.t.} \;\; (7.12b),\;\; (7.12c). \tag{7.29b}$$

And $\bar{\delta}_{\Gamma_m}$ in (7.27) is written as

$$\begin{aligned} \bar{\delta}_{\Gamma_m} &= \frac{1}{M_a} \sum_{y_{m_i,j} \in \Gamma_m} \delta_{y_{i_m,j_m}} = \frac{1}{M_a} \sum_{y_{i_m,j_m} \in \Gamma_m} \delta_{y_{i_m}} \\ &= \bar{\rho}_{\Gamma_m} \big(1 - \vartheta^{p}_{\Gamma_m}\big), \end{aligned} \tag{7.30}$$

where $\bar{\rho}_{\Gamma_m} = \frac{1}{M_a} \sum_{y_{i_m,j_m} \in \Gamma_m} \rho^{p}_{i_m}$. Therefore, based on (7.30), the objective function can be transformed as follows:

$$\min \;\; \rho^{a}_{m} \vartheta^{\mathrm{o}}_{x_m,p} \vartheta^{\mathrm{o}}_{x_m,a} + \bar{\rho}_{\Gamma_m} \vartheta^{p}_{\Gamma_m}. \tag{7.31}$$

Please note that $\vartheta^{\mathrm{o}}_{x_m,p}$ is a constant, since it denotes the transmit power of x_m that has been pushed previously. Recalling (7.7), $\vartheta^{\mathrm{o}}_{x_m,a}$ and $\vartheta^{p}_{\Gamma_m}$ are determined by ξ_p/φ_p and ξ_a/P_a. Then, (7.31) can be rewritten as

$$\min \;\; \rho^{a}_{m} \vartheta^{\mathrm{o}}_{x_m,p} \max \left\{ F\left(\frac{\xi_p}{\varphi_{p,m}}\right),\, F\left(\frac{\xi_a}{P_{a,m}}\right) \right\} + \bar{\rho}_{\Gamma_m} F\left(\frac{\xi_p}{\varphi_{p,m}}\right). \tag{7.32}$$

It shows that (7.29) follows a similar structure with (7.12), and thus it can be solved similarly. Due to the monotonicity of $G(\cdot)$, $(P_{p,1}^*, P_{a,1}^*)$ given by (7.17) is the optimum solution of (7.29) when $P_{p,m} \leqslant \frac{\xi_p}{\xi_a}(\xi_a+1)P_{a,m}$, and it can be denoted as $(P_{p,m1}^*, P_{a,m1}^*) = (P_{p,1}^*, P_{a,1}^*)$. When $P_{p,m} \geqslant \frac{\xi_p}{\xi_a}(\xi_a+1)P_{a,m}$, it can be rewritten as

$$\min \ \ \rho_m^a \vartheta_{x_m,p}^{\mathrm{o}} F\left(\frac{\xi_a}{P_{a,m}}\right) + \bar{\rho}_{\Gamma_m} F\left(\frac{\xi_p}{\varphi_{p,m}}\right), \tag{7.33a}$$

$$\text{s.t.} \ \ P_{p,m} \geqslant \frac{\xi_p}{\xi_a}(\xi_a+1)P_{a,m}, \ (7.12c). \tag{7.33b}$$

Problem (7.33) can be solved by transforming into following form:

$$\min \ \ \rho_m^a \vartheta_{x_m,p}^{\mathrm{o}} F\left(\frac{\xi_a}{P_{a,m}}\right) + \bar{\rho}_{\Gamma_m} F\left(\frac{\xi_p}{P-(\xi_p+1)P_{a,m}}\right), \tag{7.34a}$$

$$\text{s.t.} \ \ 0 \leqslant P_{a,m} \leqslant \frac{\xi_a}{\xi_p+\xi_a+\xi_a\xi_p} P. \tag{7.34b}$$

The solution of (7.34) can be obtained by employing descent methods. Then, optimal solution of (7.29) can be given as

$$(P_{p,m}^*, P_{a,m}^*) = \arg \ \max \left\{\bar{\vartheta}_m(P_{p,m1}^*, P_{a,m1}^*), \ \bar{\vartheta}_m(P_{p,m2}^*, P_{a,m2}^*)\right\}. \tag{7.35}$$

(2) Content Matching Subproblem To solve the content match subproblem efficiently, which is NP-hard, we formulate it as a matching theory-based problem. In particular, the matching between the active and proactive content has the same structure with the hospitals/residents matching problem introduced in Gusfield et al. (1989). In particular, each active content object x_m acts as a hospital. It can provide M_p available places for the proactive content packets, which act as residents in our studied problem. The objective of matching theory is to establish stable matching relationship between each active content object and M_p individual proactive content packets, which can fill all its places.

(a) Preliminary of Stable Matching Theory

As introduced in Gusfield et al. (1989), our considered hospitals/residents matching problem can be treated as an extension of stable marriage matching problem. To provide an efficient method to solve it, we first focus on the stable marriage matching problem. In this problem, all the participants can be divided into two individual sets, which can be named the men and the women, respectively. These two sets are with equal size, and one-one mapping between them can be established via matching procedure. In particular, it can be denoted as $\mathscr{S}(M, W, p_M, p_W)$, where M denotes the set of men, W denotes the set of women of size, the size of M and W is n, and p_M and p_W are the preference lists of M and W, respectively, i.e., $p_M = \{p_M(m_1) \cdots p_M(m_n)\}$, $p_M(m_i)$ denotes an ordered preference list of m_i, and it consists of all the members of W. A one-to-one correspondence between M and W is defined as a matching $\mathscr{R}$, and $\mathscr{R}$ is stable when the following constraints can be satisfied.

(Blocking pair and stable matching, Gusfield et al. (1989).) Under matching $\mathscr{R}$, (m, w) form a blocking pair if and only if

1. Under matching $\mathscr{R}$, m, and w are not matched;
2. m and w prefer each other based on their preferences.

If there exists no such blocking pair, we say that the matching $\mathscr{R}$ is stable.

In Gusfield et al. (1989), Gale-Shapley algorithm is designed to solve the stable marriage matching problem $\mathscr{S}(M, W, p_M, p_W)$. In particular, the elements of M and W are not paired initially. Then, a matching is formulated via proposing and taking proposals in an iterative way. During each iteration, an unpaired man m proposes to the first unproposed woman w of his preference list. They can be matched as a pair if w is not paired. If w is paired with another man, she will compare m with her partner. Based on her preference list, the one with higher priority is chosen as the updated partner of w, and the other one should be rejected. At the end of this algorithm, all the participators are paired. The matching results are stable if the following conditions can be satisfied, as introduced in Gusfield et al. (1989).

Proposition 7.1 *For a stable marriage problem $\mathscr{S}(M, W, p_M, p_W)$, Gale-Shapley algorithm always leads to a final stable matching under the following conditions:*

1. *Each participant has a complete preference list.*
2. *The preference of each participant is strict, i.e., no indifference.*

(b) The Generation of Preference Lists

As shown in Proposition 7.1, a stable matching is established based on the preference lists. For each participator, its preference list is generated based on the utility. In our studied content matching problem, the active content objects and proactive content packets aim to maximize their own successful transmission probability. The preference lists can be generated by solving (7.28). However, it is challenging since each individual preference relationship is jointly decided by other proactive content objects belonging in Γ_m. To calculate the utility of a matching between x_m and y_{i_m,j_m}, the following optimization problem needs to be solved.

$$\max \quad \bar{\delta}_{x_m, y_{i_m,j_m}} = \delta_{x_m} + \delta_{y_{i_m}}, \tag{7.36a}$$

$$\text{s.t.} \;\; (7.12b), \; (7.12c), \tag{7.36b}$$

where δ_m and $\delta_{y_{i_m}}$ follow the notations given by (7.25). (7.36) can be solved by following the same method as (7.29), which has been studied previously. To generate complete preference lists for all the participators, $M_a M_p$ individual optimization problems need to be solved, and all these problems are similar to (7.36).

Please note that all the proactive content packets generated by y_{i_m} have the same utility. Therefore, $y_{i_m,1}, \ldots, y_{i_m,M_a}$ generate a tie of preference list with respect to x_m. Then, the preference list of x_m can be expressed as

$$p_a(x_m) : y_{i_1,1} = \cdots = y_{i_1,M_a} > \cdots > y_{i_{M_p},1} = \cdots = y_{i_{M_p},M_a}, \tag{7.37}$$

and the corresponding successful transmission probability of y_{i_m} follows the inequality

$$\delta_{x_m}(y_{i_1,1}) = \cdots = \delta_{x_m}(y_{i_1,M_a}) > \cdots > \delta_{x_m}(y_{i_{M_p},1}) = \cdots = \delta_{x_m}(y_{i_{M_p},M_a}). \tag{7.38}$$

Similarly, the preference list of a specific proactive content packet y_{i_m,j_m} is expressed as

$$p_p(y_{i_m,j_m}) : x_{k_1} > \cdots > x_{k_{M_a}}, \text{ where } \delta_{y_{i_m,j_m}}(x_{k_1}) > \cdots > \delta_{y_{i_m,j_m}}(x_{k_{M_a}}). \tag{7.39}$$

Based on (7.37) and (7.39), both x_m and y_{i_m,j_m} try to choose a partner/partners to maximize its own successful transmission probability.

(c) Matching Algorithm

As introduced previously, the subproblem of content matching is to establish a many-one matching, which is a typical hospitals/residents matching problem. In particular, each active content object has M_p available places, which means that it can be transmitted with M_p proactive content packets. Therefore, a many-one mapping from the proactive content packets to the active content objects should be generated. To ensure the efficiency of solving this problem, a stable matching is defined as follows.

(Stable matching of many-one matching problem, Gusfield et al. (1989).) A matching $\mathscr{T}$ of a many-one matching problem is stable if there is no x_m and y_{i_m,j_m} such that all the following conditions:

1. x_m and y_{i_m,j_m} are acceptable to each other;
2. Either y_{i_m,j_m} is not matched or x_m is preferred by it compared to its partner in $\mathscr{T}$;
3. Either x_m has unoccupied places or y_{i_m,j_m} is preferred by it compared to at least one of its partners in $\mathscr{T}$.

Our studied problem can be reformed as a classic stable marriage matching problem by replacing x_m by using M_p identical content objects, i.e., $x_{m,1}, \ldots, x_{m,M_p}$. In particular, $x_{m,i}$ has only one available place, $i = 1, \ldots, M_p$. Then, x_m can be replaced by $x_{m,1}, \ldots, x_{m,M_p}$ in the preference list of each proactive content packet. In the preference list of x_m, some proactive content packets are with equivalent priority, and thus the stability of matching results can be guaranteed. But it is a weak-stable matching if the instability is defined as follows: $\mathscr{R}$ is an unstable matching if x_m and y_{i_m,j_m} are not paired with each other, and they both strictly prefers each other rather than their own partners. Next, the ties in the preference list can be broke, and the preference x_m given by (7.37) can be updated as follows, which is with strictly ordered preference:

$$p_a(x_m) : y_{i_1,1} > \cdots > y_{i_1,M_a} > \cdots > y_{i_{M_p},1} > \cdots > y_{i_{M_p},M_a}. \tag{7.40}$$

Based on (7.40), the content matching problem can be solved by employing Gale-Shapley algorithm, which is provided by Algorithm 1. During the matching procedure, the active content objects propose to the proactive content packets. If y_{i_m,j_m} accepts a proposal, it will be unpaired, but a new partner can be chosen. Moreover, it will accept a proposal from another active content object x_k, if and only if x_k is with higher priority than its current partner in the preference list. Therefore, all the proactive content packets that are with lower priority than the current partner of y_{i_m,j_m} should be eliminated from the preference list, and y_{i_m,j_m} can be eliminated from their preference lists. x_m will always make a proposal to the first unmatched candidate y_{i_n,j_n} of its current preference list. Then, x_m and y_{i_m,j_m} can be matched as a pair. If y_{i_n,j_n} has another partner, y_{i_n,j_n} needs to break up with it. In the end of this algorithm, the active content objects and proactive content packets are matched with each other. The stability can be verified by the following theorem.

Theorem 7.4 *Algorithm 1 can obtain a stable matching result $\mathscr{T}$, and each active content object x_m is with the best stable partners in $\mathscr{T}$.*

Proof First, we need to prove that all the M_p available places of each active content objects will be taken by the proactive content packets. It can be proved by contradiction. We assume that x_m is rejected by the last candidate of its preference list. Then it must have at least one available place, while all the proactive content packets are matched. However, it is conflict with the fact that all the active content objects have $M_a M_p$ available places.

Next, we will prove the stability of matching $\mathscr{T}$. Please note that a pair is not stable if it is not in the preference list due the following reasons: If (x_m, y_{i_k,j_k}) is the first stable pair that has been removed, it must be eliminated when another proposal is made to y_{i_k,j_k}, and y_{i_k,j_k} prefers another active content object to x_m. Moreover, the number of proactive content packets y_{i_l,j_l}, which has high priority than y_{i_k,j_k} in the preference list of x_n, is less than M_p. Otherwise, one of them must have been removed before deleting (x_m, y_{i_k,j_k}), which conflicts with the assumption that (x_m, y_{i_k,j_k}) is the first removed stable pair. Therefore, x_n either has been matched with one proactive content packet that is with higher priority than y_{i_k,j_k} or has one available place. Therefore, (x_m, y_{i_k,j_k}) is not stable, since there exists a blocking pair (x_n, y_{i_k,j_k}). Since the unlisted pairs are not stable, they are not block pairs. In addition, there does not exist any block pair in $\mathscr{T}$. Hence, $\mathscr{T}$ is stable.

Due to the stability of $\mathscr{T}$, each active content object x_m is matched with at least M_p partners. Please note that $\mathscr{T}$ consists of all the stable matching results, and thus x_m can with its first M_p be preferred partners, which coincide with the proposing order. Therefore, x_m can be paired with the best stable partners. Then the proof has been finished.

Algorithm 1 The content matching algorithm

```
Step 1. Preference list generation

The preference lists $p_a(x_m)$ and $p_p\ (y_{i_m,j_m})$ are generated due to (7.39) and (7.40).
Step 2. The content matching procedure
Initialization: all the content objects are set to be unpaired.
Repeat: For $x_m$, $m = 1, \ldots, M$,

  • If There exist unproposed proactive content packets in the preference list of $x_m$
        – $y_{i_n,j_n}$ is the first candidate of the preference list of $x_m$.
        – If $y_{i_n,j_n}$ is in the matched status
            * The matching between $y_{i_n,j_n}$ and its current partner $x_n$ can be broken,
              and a new matching with $x_m$ should be established.
            * The active content objects, which are with lower priority than $x_m$ in
              the preference list of $y_{i_n,j_n}$, should be deleted, and $y_{i_n,j_n}$ needs to be
              removed from their preference lists.
        – Else
            * A matching between $y_{i_n,j_n}$ and $x_m$ should be established.
        – End
  • End

Termination: When all the active content objects and proactive content packets are matched.
```

7.5 Computation Complexity Analysis and Simulation Results

To evaluate the computation efficiency and performance of content caching, especially the NOMA-based content pushing and delivering scheme, both the computational complexity analysis of joint optimization algorithm and simulation results are provided in this part.

7.5.1 Computation Complexity Analysis

The computational complexity of joint optimization is determined by the decoupled subproblems. In particular, the main computation task of power allocation is solving (7.34) by employing descent method. Following Cartis et al. (2010), the corresponding computational complexity is $O(\epsilon_1^{-2})$, where ε_1 is a constant with respect to the norm of gradient. The power allocation strategy can be obtained by solving M_a independent problems. Then, the computational complexity of power allocation can be provided as $O(M_a\epsilon_1^{-2})$.

Moreover, content matching problem is solved by the following two steps: First, the preference lists should be established for all the active content objects and proactive content packets. M_aM_p optimization problems, which are provided by (7.36), have to be solved to obtained the utility of each potential matching. Similar to (7.34), these problems are solved by gradient descent-based method. Then, the corresponding computational complexity is $O(M_aM_p\epsilon_2^{-2})$, where ε_2 is a constant. Next, all the participators can be paired via the matching procedure, and at most M_aM_p proposals can be made. The computation cost of a single time proposal is denoted as y_p, and then the computational complexity of entire matching procedure can be written as $O(M_aM_py_p)$.

Therefore, the computational complexity of joint optimization algorithm is $O(M_a\epsilon_1^{-2} + M_aM_p(\epsilon_2^{-2} + y_p))$, which is a polynomial function with respect to the number of content objects.

7.5.2 Simulation Results

In this part, the simulation results are plotted, which can demonstrate the analytical results and evaluate the performance of optimization algorithm. The key parameters are set as Zhao et al. (2017).

The outage probability of content delivery with content caching in F-RANs has been evaluated in Fig. 7.2, where both the conventional and NOMA-based content pushing and delivering schemes are considered. Moreover, compared with the conventional scheme, the NOMA-based scheme can achieve better outage

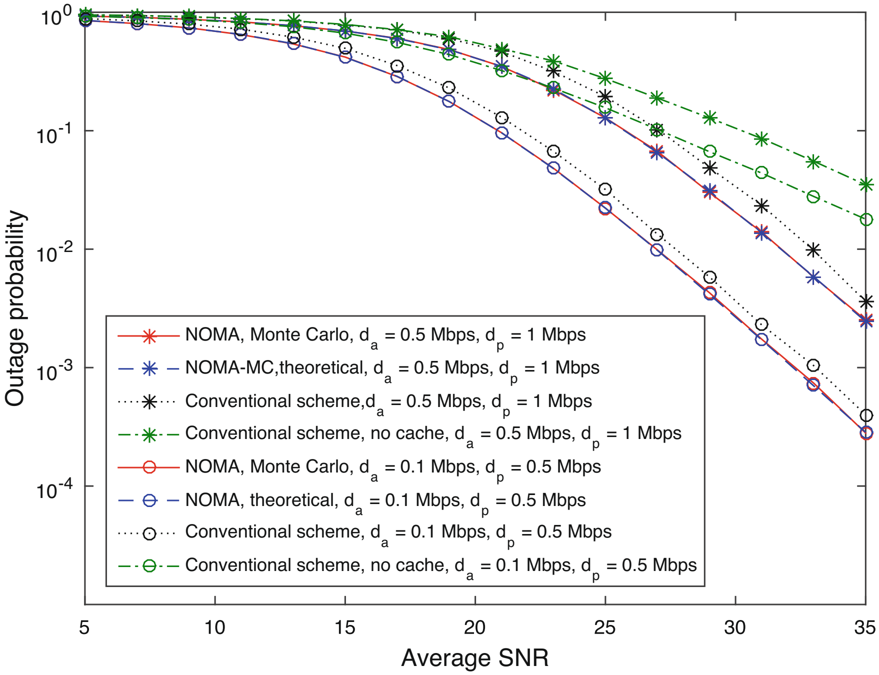


Fig. 7.2 Outage probability vs. SNR γ of NOMA-MC scheme (Zhao et al. 2017)

performance with optimized power allocation schemes. The curves of outage probability with respect to the conventional and NOMA-based schemes are with the same slop, which means that the NOMA-based scheme can obtain full diversity gains without occupying extra radio resource for content pushing.

In Fig. 7.3, the average data rate is provided, which is defined as the production of the target data rate and successful transmission probability. Compared with the conventional scheme, the NOMA-based scheme can achieve better data rate performance, since it is with higher spectrum efficiency by avoiding taking extra radio resource for content pushing. Moreover, the NOMA-based scheme with content caching can higher transmission reliability than the content delivery scheme without caching in F-RANs, which shows that it can improve the data rate with little cost of transmission reliability.

To evaluate the performance gains of joint optimization algorithm, the outage probability is plotted in Fig. 7.4, where the numbers of active and proactive content objects are $M_a = 5$ and $M_a = 3$, respectively. In particular, the popularity of content objects is assumed to follow Zipf distribution. As shown in the figure, the performance gains of NOMA-based content pushing and delivering scheme can

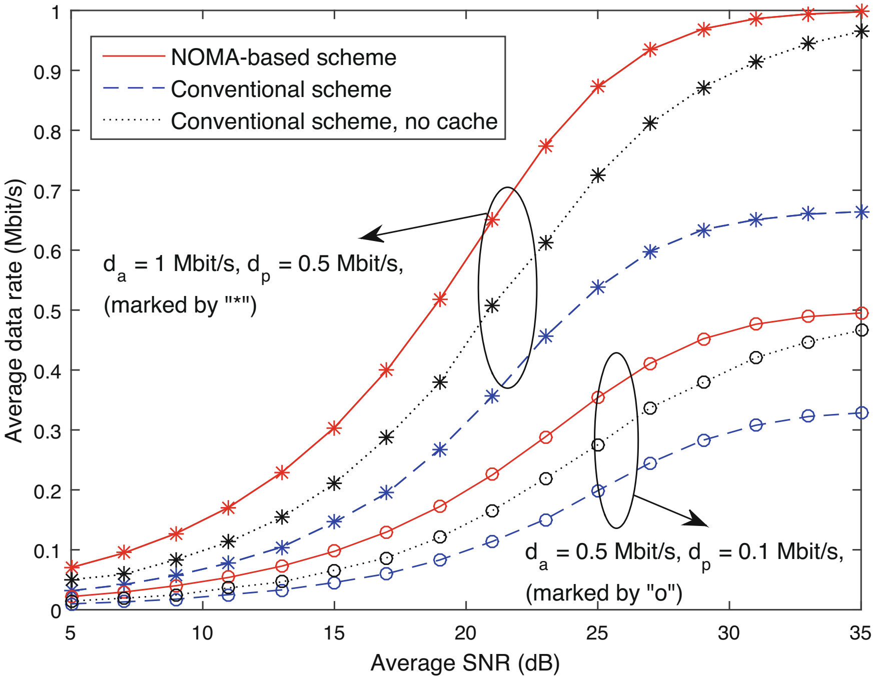


Fig. 7.3 Average data rate vs. SNR γ of NOMA-MC scheme

be improved via joint optimization. It can be enlarged by using our proposed joint optimization algorithms. In particular, when the average SNR is $\gamma = 20$ dB, the outage probability of algorithm can be reduced by 0.003.

7.6 Summary

In this chapter, we studied content caching in F-RANs, especially the tradeoff between the cost and utility of content caching. First, the performance of content pushing and delivering with edge caching in F-RANs has been studied, it shows that the spectrum efficiency can be improved significantly with little cost of reliability by employing NOMA-based scheme. Then, the joint optimization of radio resource and content caching management is designed to further improve the performance gains.

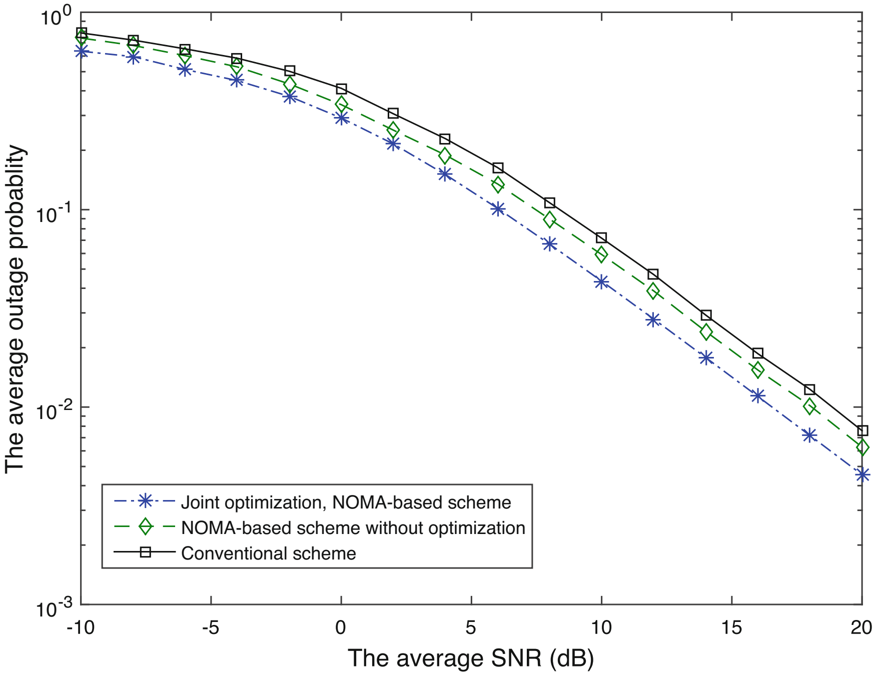


Fig. 7.4 Outage performance of joint optimization of power allocation and content matching

References

Bazaraa M et al (1979) Nonlinear programming: theory and algorithms. Wiley, New York

Cartis C et al (2010) On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization. SIAM J Optim 20:2833–2852

Gradshteyn I et al (2000) Table of integrals, series, and products. Academic, San Diego

Gusfield D et al (1989) The stable marriage problem: structure and algorithms. The MIT Press, Cambridge

Papadimitriou C et al (1998) Combinatorial optimization-algorithms and complexity. Dover, Mineola

Stoyan D et al (1995) Stochastic geometry and its applications. Wiley, New York

Zhao Z et al (2017) A non-orthogonal multiple access-based multicast scheme in wireless content caching networks. IEEE J Sel Areas Commun 35:2723–2735

Chapter 8
Computation Offloading in Fog Radio Access Networks

The rest of this chapter is organized as follows: In Sect. 8.1, the performance analysis of hierarchical cloud-fog computation offloading is provided. Then, the joint optimization of computation offloading in F-RANs is introduced in Sect. 8.2. Section 8.3 provides the simulation results, and the conclusion is given in Sect. 8.4.

8.1 Performance Analysis of Hierarchical Cloud-Fog Computation Offloading

In F-RANs, sufficient computation resources are provided by both the cloud and fog-computing layers, which can be employed to execute various computation tasks. However, there exist a tradeoff between the cost and execution efficiency of offloading the computation tasks to different layers. To fully explore the potential of F-RANs, a hierarchical cloud-fog computation offloading framework should be established in F-RANs, and the latency performance is studied in this part, which aims to provide some insights.

8.1.1 System Model

We consider of deploying computation offloading in F-RANs, where the users can offload their computation tasks to improve the execution efficiency. To model the amorphous coverage feather of F-RANs, the locations of cloud computing centers are modeled as a homogeneous PPP Φ_c, while the locations of F-APs are formed as another homogeneous PPP Φ_f, and their densities are denoted as ρ_c and ρ_f, respectively. Each F-AP can connect with only one cloud computing center via wired backhaul. Please note that the users are always close to their associated F-

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0_8

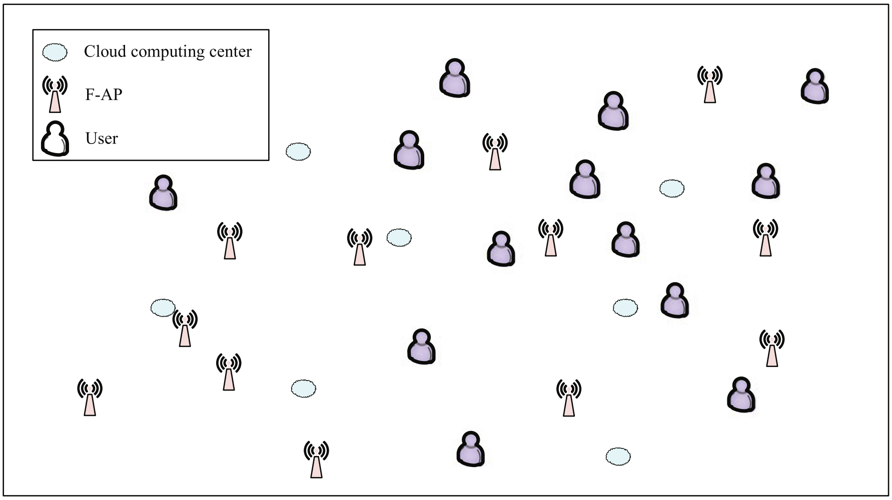


Fig. 8.1 Stochastic geometry-based system model of F-RANs

APs, and thus Thomas cluster process is employed to model the locations of users, which can be denoted as Ψ_u (Haenggi 2013). In a specific user cluster G_T, it contains N users, which associate with an F-AP F_T. For a typical user U_n in G_T, the distance between U_n and F_T, which is denoted as d_n, follows Gaussian distribution, and its PDF can be expressed as

$$p(d_n) = \frac{d_n}{\omega^2} \exp\left(- \frac{d_n^2}{2\omega^2} \right), d_n \geqslant 0, \tag{8.1}$$

where ω^2 denotes the variance of d_n. To remove the co-channel interference in G_T, the users in G_T use orthogonal subchannels to offload their computation tasks, and all the computation tasks are with equivalent data volume (Fig. 8.1).

8.1.2 *Hierarchical Cloud-Fog Computation Offloading Framework*

As shown in Fig. 8.2, computation tasks are generated by U_n, which can be formulated as a queue r_n. It can be modeled a Poisson process with a given arrival rate ς_n. Moreover, the data volume of each computation task follows independently identically exponential distribution, i.e., S_n denotes the data volume of a specific computation task τ_m, S_n ~Exp(ϕ), and ϕ denotes the expectation of S_n. ζ_n is the output rate of computation task queue r_n, which is a constant. Then, the service time of each computation task follows exponential distribution, and the corresponding

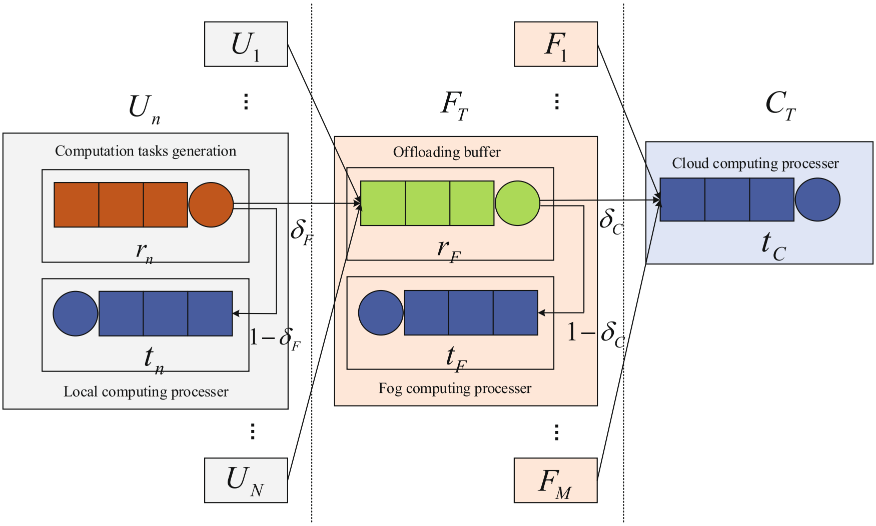


Fig. 8.2 Illustration of hierarchical cloud-fog computation offloading framework

expected service rate can be expressed as $\nu_n = \zeta_n/\phi$. A computation task τ_m in r_n can be executed by the following modes.

(1) Local Computation Mode $\mathscr{M}_1$ A local computing processer $\mathscr{P}_n$ is equipped with U_n, which can be used to execute τ_n with a fixed processing speed s_n. Denoting δ_F as the offloading probability, the probability that τ_n is processed by $\mathscr{P}_n$ can be written as $(1-\delta_F)$. Since a part of computation tasks of r_n are processed locally, these computation tasks can be modeled as a queue t_n. In the local computation mode, the latency is caused by the waiting in r_n and t_n, which can be expressed as

$$D_L = D(r_n) + D(t_n), \tag{8.2}$$

where $D(r_n)$ and $D(t_n)$ are the time of τ_m spending in r_n and t_n, respectively.

(2) Fog Computation Offloading Mode $\mathscr{M}_2$ As illustrated in Fig. 8.2, N users are associated with F_T, and a fog-computing processer $\mathscr{P}_F$ is equipped with F_T, whose processing speed is s_F. All the associated users can offload their computation tasks to F_T through the wireless channels, and all the offloaded computation tasks formulate a queue r_F. Please note that both the cloud and fog-computing processors can be employed for computation task execution, and the corresponding offloading probability of computation tasks in r_F can be given as δ_C and $(1-\delta_C)$, respectively. The offloaded computation tasks that are processed by $\mathscr{P}_F$ can be modeled as a queue t_F. Then, the latency caused by fog computation offloading mode is

$$D_F = D(r_n) + D(r_F) + D(t_F), \tag{8.3}$$

where $D(r_n)$ follows the notation given by (8.2), $D(r_F)$ is the latency of τ_m spending in r_F, and $D(t_F)$ is defined similarly for τ_m in t_F. We assume that the data volume of execution results are so small that the latency caused by results feedback can be overlooked.

3) Cloud Computation Offloading Mode $\mathscr{M}_3$: When τ_m is offloaded to the cloud computing center, it should be first transmitted to the offloading queue r_F at F_T via the wireless channel, and then τ_m is forwarded to the cloud computing center C_T. A cloud processer $\mathscr{P}_C$ is equipped with C_T, and its processing speed s_C. All the offloaded computation tasks at $\mathscr{P}_C$ is modeled as a queue $\mathscr{P}_C$. Then, the total latency of processing computation tasks via cloud computation offloading mode is given as follows:

$$D_C = D(r_n) + D(r_F) + D(t_C), \tag{8.4}$$

where $D(t_C)$ is the time of τ_m spending in queue t_C, and the notations of $D(r_n)$ and $D(r_F)$ are defined by (8.3).

8.1.3 Computation Offloading Probability and Latency Analysis

In this part, we focus on the performance of opportunistic offloading strategy, where the computation tasks can be offloaded if and only if the transmission capability of wireless channels can support the requirements of computation tasks. Hence, the failure of computation offloading can be avoided, and its efficiency can be guaranteed.

The Computation Offloading Probability

The computation offloading probability can be defined as the occurrence probability that the data rate of wireless channel between U_n and F_T is larger than the departure rate of computation task queue of U_n, i.e.,

$$\delta_F = \Pr\left\{W \log(1 + \varphi_n) \geq \zeta_n\right\}, \tag{8.5}$$

where W is the bandwidth of wireless channel, ζ_n is the departure rate of computation task queue r_n of U_n, φ_n denotes the SIR. In this part, we focus on an interference limited scenario, and φ_n can be written as

$$\varphi_n = \frac{|g_n|^2 d_n^{-\alpha}}{\sum_{U_k \in \Psi_a / U_n} |g_k|^2 d_k^{-\alpha}}, \tag{8.6}$$

where Ψ_a denotes the set of interfere users that occupy the same frequency channel with U_n, g_n and g_k capture the channel fading of the wireless channels for U_n and U_k, respectively, d_n is the distance between U_n and F_T, d_k is defined similarly for U_k and F_T, and the exponent of path loss is denoted as α. In this part, the flat Rayleigh channel fading model is employed, i.e., g_n, $g_k \sim CN(0, 1)$. Then, a tractable expression of δ_F is given by the following theorem.

Theorem 8.1 *When the opportunistic offloading strategy is employed, the computation offloading probability is written as*

$$\delta_F = \frac{\sqrt{\zeta_n^2 + 16\pi\varsigma_n\phi\rho_f J(\alpha)\varepsilon^{2/\alpha}\omega^2\zeta_n} - \zeta_n}{8\pi\varsigma_n\phi\rho_f J(\alpha)\varepsilon^{2/\alpha}\omega^2}, \tag{8.7}$$

where $J(\alpha) = \frac{\pi}{\alpha}\csc\left(\frac{2\pi}{\alpha}\right)$ *and* $\varepsilon = 2^{\frac{\zeta_n}{W}} - 1$.

Proof Based on (8.6) and (8.5), δ_F can be further derived as follows:

$$\delta_F = \Pr\left\{|g_n|^2 \geq \varepsilon d_n^{\alpha}\left(\sum_{U_k\in\Psi_a/U_n} |g_k|^2 d_k^{-\alpha}\right)\right\}. \tag{8.8}$$

Please note that $|g_n|^2$ and $|g_k|^2$ are identically independently distributed, and thus (8.8) is derived as follows:

$$\begin{aligned}\delta_F &= \mathbb{E}\left\{\prod_{U_k\in\Psi_a/U_n}\exp\left(-\varepsilon|g_k|^2 d_n^{\alpha} d_k^{-\alpha}\right)\right\}\\ &= \mathbb{E}_{d_n}\underbrace{\left\{\mathbb{E}_{\Psi_a}\left\{\prod_{U_k\in\Psi_a/U_n}\frac{1}{1+\varepsilon d_k^{-\alpha} d_n^{\alpha}}\right\}\right\}}_{\mathcal{T}}.\end{aligned} \tag{8.9}$$

Based on the PGFL of Thomas cluster process given by Ganti et al. (2009), (8.9) is derived as follows:

$$\mathcal{T} \geq \exp\left\{-2\pi\rho_a\int_0^{\infty}\left[1-\left(\frac{1}{1+\varepsilon d_k^{-\alpha} d_n^{\alpha}}\right)^{\bar{c}}\right]d_k \mathrm{d}d_k\right\}, \tag{8.10}$$

where $\bar{c}$ is the number of interfere users of each cluster. Please note the orthogonal channels are employed by the users in the same cluster, and thus there exists only one interfere user in each cluster, i.e., $\bar{c} = 1$. Then the equality of (8.10) can be established, and (8.9) is expressed as follows:

$$\begin{aligned}\mathscr{T} &= \exp\left(-2\pi\rho_a \int_0^\infty \frac{\varepsilon d_k^{-\alpha} d_n^{\alpha}}{1+d_k^{-\alpha} d_n^{\alpha}} d_k \mathrm{d}d_k \right),\\ &= \exp\left(-2\pi\rho_a \varepsilon^{2/\alpha} d_n^2 \int_0^\infty \frac{s}{s^{\alpha}+1} \mathrm{d}s \right),\\ &= \exp\left[-2\pi\rho_a J(\alpha) \varepsilon^{2/\alpha} d_n^2 \right].\end{aligned} \tag{8.11}$$

Equation (8.11) can be delivered by variable substitution $s = (\varepsilon d_n^{\alpha})^{-1/\alpha} d_k$. Based on (8.1), (8.11), and (8.9), δ_F is expressed as follows:

$$\delta_F = \frac{1}{1+4\pi\rho_a J(\alpha)\varepsilon^{2/\alpha}\omega^2}. \tag{8.12}$$

Please note the generation rate of r_n follows Poisson distribution, and its departure rate follows exponential distribution. Therefore, r_n is an $M/M/1$ queue, and its non-idle probability is $\frac{\varsigma_n}{\nu_n}$. Then, Ψ_a is a Poisson process, and its density can be given as $\rho_a = \frac{\varsigma_n \phi}{\zeta_n \delta_F}$. Finally, (8.7) is derived based on ρ_a and (8.12). The proof has been finished.

The Latency Performance

As introduced previously, the latency of computation offloading consists of waiting latency, offloading latency, and execution latency. Then, the expected latency of τ_m is defined as follows:

$$\bar{T} = \vartheta_L \bar{D}_L + \vartheta_F \bar{D}_F + \vartheta_C \bar{D}_C, \tag{8.13}$$

where $\bar{D}_L$, $\bar{D}_F$, and $\bar{D}_C$ are the expected delay of local computation, fog computation offloading, and cloud computation offloading modes, respectively, and ϑ_L, ϑ_F and ϑ_C are the corresponding probability that the aforementioned three modes are employed for processing τ_m, and they can be written as follows:

$$\vartheta_L = 1-\delta_F,\ \vartheta_F = \delta_F(1-\delta_C),\ \vartheta_C = \delta_F\delta_C. \tag{8.14}$$

Since the computation task generation queue r_n is an $M/M/1$ queue, the arrival process of t_n is a Poisson process based on the Burke's Theorem, and the arrival rate is $(1-\delta_F)\varsigma_n$. The processing speed of t_n is fixed, and thus its service time is exponentially distributed with a given expectation $\frac{\phi}{s_n}$. Therefore, t_n is an $M/M/1$

queue, and so are r_F, t_F, and t_C. As introduced in Balter (2013), the hierarchical cloud-fog computation offloading framework can be modeled as a Jackson network.

Remark 8.1 The hierarchical cloud-fog computation offloading framework in F-RANs is a queuing system, which consists of $M/M/1$ queues r_n, t_n, r_F, t_F, and t_C. It can be treated as a Jackson network, and can be decoupled as independent queues.

Due to Remark 8.1, the expected latency of computation offloading in F-RANs can be provided by the following theorem.

Theorem 8.2 *The expected latency defined by* (8.13) *can be given as*

$$\begin{aligned}\bar{T} =& \frac{\phi}{\zeta_n - \varsigma_n\phi} + \frac{(1-\delta_F)\phi}{s_n - (1-\delta_F)\varsigma_n\phi} + \frac{\delta_F\phi}{\zeta_n - \eta\phi} + \frac{\delta_F(1-\delta_C)\phi}{s_F - (1-\delta_C)\eta\phi} \\ &+ \frac{\delta_F\delta_C\phi}{s_C - \mu\phi - \delta_C\eta\phi},\end{aligned} \tag{8.15}$$

where $\eta = \sum_{n=1}^{N} \delta_F \varsigma_n$ *and* $\mu = \sum_{F_i \in \Psi_f/F_T} \delta_{C,i}\eta$. *To guarantee the stability of all the queues, the following restricts have to be followed:*

$$\begin{aligned}\zeta_n - \varsigma_n\phi > 0,\ s_n - (1-\delta_F)\varsigma_n\phi > 0,\ \zeta_n - \eta\phi > 0, \\ s_F - (1-\delta_C)\eta\phi > 0,\ s_C - \mu\phi - \delta_C\eta\phi > 0.\end{aligned} \tag{8.16}$$

Proof Based on Remark 8.1, the expected latency of all the three computation modes can be derived independently, which can be analyzed as follows.

(1) The Expected Latency of Local Computation Mode

Recalling (8.2), the latency of local computation mode is determined by the time spending in r_n and t_n need to be analyzed, which are both $M/M/1$ queues. We first focus on the expected latency of r_n, which can be written as follows based on Balter (2013):

$$\bar{L}(r_n) = \frac{\varsigma_n/\nu_n}{1 - \varsigma_n/\nu_n} = \frac{\varsigma_n\phi}{\zeta_n - \varsigma_n\phi}. \tag{8.17}$$

Due to Little's law, the expected latency of r_n is derived as follows:

$$\bar{D}(r_n) = \frac{\bar{L}(r_n)}{\varsigma_n} = \frac{\phi}{\zeta_n - \varsigma_n\phi}. \tag{8.18}$$

Moreover, the probability that a computation task τ_m of r_n arrives t_n with probability $(1-\delta_F)$, which is provided by (8.14). Then, its arrival rate is $\varsigma_{t_n} = (1-\delta_F)\varsigma_n$, while its service rate can be given as $\nu_{t_n} = \frac{s_n}{\phi}$. Therefore, the expected latency of t_n is

$$\bar{D}(t_n) = \frac{\phi}{s_n - (1-\delta_F)\varsigma_n\phi}. \tag{8.19}$$

Then, the expected latency of local computation mode $\bar{D}_L$ is derived as

$$\bar{D}_L = \frac{\phi}{\zeta_n - \varsigma_n \phi} + \frac{\phi}{s_n - (1 - \delta_F)\varsigma_n \phi}. \tag{8.20}$$

(2) The Expected Latency of Fog Computation Offloading Mode

Recalling (8.3), the expected latency of fog computation offloading mode is decided by the time of τ_m spending r_n, r_F, and t_F. As introduced previously, the arrival rate of r_F is $\varsigma_{r_F} = \eta$, while its service rate is $\nu_{r_F} = \frac{\zeta_n}{\phi}$. Since r_F is an $M/M/1$ queue, the expected latency is derived as follows:

$$\bar{D}(r_F) = \frac{\phi}{\zeta_n - \eta\phi}. \tag{8.21}$$

Similarly, the expected latency of t_F is derived as

$$\bar{D}(t_F) = \frac{\phi}{s_F - (1 - \delta_C)\eta\phi}, \tag{8.22}$$

where $\nu_{t_F} = \frac{s_F}{\phi}$ and $\varsigma_{t_F} = (1 - \delta_C)\eta$ denote the service and arrival rates, respectively. Moreover, the probability that τ_m is offloaded to t_F is $(1-\delta_C)\eta$, and the expected latency of fog computation offloading mode can be expressed as follows based on (8.18), (8.21), (8.22), and (8.3):

$$\bar{D}_F = \frac{\phi}{\zeta_n - \varsigma_n \phi} + \frac{\phi}{\zeta_n - \eta\phi} + \frac{\phi}{s_F - (1 - \delta_C)\eta\phi}. \tag{8.23}$$

(3) The Expected Latency of Cloud Computation Offloading Mode

Similarly to the fog computation offloading mode, the offloading procedure of cloud computation offloading mode can be modeled as a tandem queueing system. Please note that multiple F-APs connects with a cloud computing center C_T, and all of them can offload their computation tasks to C_T. The corresponding offloading probability can be denoted as δ_C. Moreover, the expected number of F-APs that need to offload to C_T can be given as $\bar{M} = \rho_f/\rho_c$. Then, the arrival process of offloaded computation tasks queue t_C can be modeled as a Poisson process, whose arrival rate can be expressed as t_C is $\varsigma_{t_C} = \mu\phi + \delta_C\eta\phi$. Moreover, the service rate of C_T can be given as$\nu_{t_C} = \frac{s_C}{\phi}$. Therefore, the expected latency of t_C is derived as

$$\bar{D}(t_C) = \frac{\phi}{s_C - \mu\phi - \delta_C\eta\phi}. \tag{8.24}$$

Based on (8.18), (8.21), (8.24), and (8.4), the expected latency of cloud computation offloading mode is expressed as follows:

$$\bar{D}_C = \frac{\phi}{\zeta_n - \varsigma_n\phi} + \frac{\phi}{\zeta_n - \eta\phi} + \frac{\phi}{s_C - \mu\phi - \delta_C\eta\phi}. \tag{8.25}$$

Then, the theorem can be proved based on (8.20), (8.23), and (8.25).

8.1.4 Further Discussion of Offloading Strategy

Since different processing modes of computation tasks can be employed in F-RANs, a sophisticated coordination strategy should be designed to make full use of computation resources, which can be managed by optimizing δ_C in (8.13). To minimize the expected latency, the following optimization problem can be formulated:

$$\min_{\delta_C} \bar{T} \text{ given by (8.15)}, \; s.t. \, 0 \leqslant \delta_C \leqslant 1. \tag{8.26}$$

Equation (8.26) is a single-variant optimization problem with respect to δ_C. To achieve the minimum latency, the first and second-order derivatives of $\bar{T}$ are provided as follows:

$$\nabla_{\delta_C}\bar{T} = \frac{\delta_F\phi(s_C - \mu\phi)}{[s_C - (\mu + \eta\delta_C)\phi]^2} - \frac{\delta_F s_F\phi}{[s_F - \eta(1-\delta_C)\phi]^2}, \tag{8.27a}$$

$$\nabla^2_{\delta_C}\bar{T} = \frac{2\delta_F(s_C - \mu\phi)\eta\phi^2}{[s_C - (\mu + \eta\delta_C)\phi]^3} + \frac{2\delta_F s_F\eta\phi^2}{[s_F - \eta(1-\delta_C)\phi]^3}. \tag{8.27b}$$

To obtain the optimal solution of (8.26), we first formulate an equation $\nabla_{\delta_C}\bar{T} = 0$, whose solutions can be expressed as follows:

$$\chi_1 = \frac{(2\eta\phi + s_C - \mu\phi) + \sqrt{(2\eta\phi + s_C - \mu\phi)^2 - 4\eta^2\phi^2}}{2}, \tag{8.28a}$$

$$\chi_2 = \frac{[s_C - (\mu + \eta)\phi]^2}{s_C - \mu\phi}. \tag{8.28b}$$

Due to the relationship of s_F, χ_1, and χ_2, the optimization of computation offloading strategy can be divided into three different cases.

(1) When $s_F \geqslant \chi_1$ In this case, the inequality $\nabla_{\delta_C}\bar{T} \geqslant 0$ can be established, which means that $\bar{T}$ is an increasing function. Therefore, the probability of cloud computation offloading mode should be set as $\delta_C^* = 0$ to minimize the expected latency. It indicates that the cloud computation offloading mode should not be

employed, and all the offloaded computation tasks in r_F are processed by the fog-computing processors.

(2) When $\chi_2 < s_F < \chi_1$ The optimal ratio of employing cloud computation offloading mode is between 0 and 1 when $\chi_2 < s_F < \chi_1$, and thus both the cloud and fog computation offloading modes should be implemented. Please note that (8.26) is a convex problem, and the optimal solution can be derived by solving $\nabla_{\delta_C} \bar{T} = 0$, which can be written as

$$\delta_{C,2}^* = \frac{(s_C - \mu\phi)\sqrt{\frac{s_F}{s_C - \mu\phi}} - s_F + \eta\phi}{\eta\phi\left(\sqrt{\frac{s_F}{s_C - \mu\phi}} + 1\right)}. \tag{8.29}$$

(3) When $s_F \leqslant \chi_2$ Unlike the first case, $\bar{T}$ keeps decreasing as δ_C increases in this case, and thus the optimal solution is $\delta_C^* = 1$. Therefore, it indicates that only the cloud computation offloading mode should be employed to handle the offloaded computation tasks.

8.2 Joint Optimization of Computation Offloading in F-RANs

The performance of computation offloading in F-RANs is jointly determined by computation and communication capabilities, which should be optimized jointly. In this part, the joint optimization of resource allocation and offloading decision is considered, which aims to further improve the performance gains of computation offloading in F-RANs.

8.2.1 Problem Formulation

Based on the aforementioned hierarchical cloud-fog paradigm, we focus on the optimization of computation offloading in F-RANs, which consists of multiple F-APs and a single cloud computing center. In particular, the wireless backhaul links are employed to establish the connections between the F-APs and cloud computing center. First, a specific F-AP F_{T} is studied to formulate the optimization problem, which has M associated users, i.e., $U_1, \ldots, U_M$. For each user U_m, it generates N computing tasks that needs to be executed, which can be denoted as $\tau_{m,1}, \ldots, \tau_{m,N}$. For a specific computation task $\tau_{m,n}$, the total latency and energy consumption can be modeled as follows when different processing modes are employed.

(1) Local Computation Mode In this mode, the computation task can be handled locally at the user. Therefore, its performance and cost are mainly determined by

the processing capability of local computing processor. In particular, by denoting the CPU-cycle frequency of U_m as $\psi_{\mathrm{L},m}$, the energy consumption is modeled as follows:

$$G^{\mathrm{L}}_{m,n} = \kappa_{\mathrm{L},m} N_{m,n} \psi^2_{\mathrm{L},m}, \tag{8.30}$$

where $\kappa_{\mathrm{L},m}$ denotes switched capacitance, it is determined by the chip architecture (Burd et al. 1996), and $N_{m,n}$ is the data volume of $\tau_{m,n}$. Moreover, $\zeta_{\mathrm{L},m}$ denotes the CPU cycles that are needed to execute one bit, and the computation latency caused by handling $\tau_{m,n}$ is

$$T^{\mathrm{L}}_{m,n} = \frac{\zeta_{\mathrm{L},m} N_{m,n}}{\psi_{\mathrm{L},m}}. \tag{8.31}$$

(2) Fog Computation Offloading Mode U_m needs to offload $\tau_{m,n}$ to its associated F-AP F_{T} through the wireless channel, and the corresponding channel capacity is

$$S_m = B_m \log_2 \left(1 + \frac{h_m \rho_m}{B_m \omega_0}\right), \tag{8.32}$$

where B_m is the bandwidth, h_m is the channel coefficient, ρ_m denotes the transmit power, and the density of noise power can be denoted as ω_0. The latency caused by offloading $\tau_{m,n}$ to F_{T} is expressed as

$$T^{\mathrm{F},1}_{m,n} = \frac{N_{m,n}}{S_m} = \frac{N_{m,n}}{B_m \log_2 \left(1 + \frac{h_m \rho_m}{B_m \omega_0}\right)}. \tag{8.33}$$

The energy consumption caused by offloading can be given as

$$G^{\mathrm{F},1}_{m,n} = T^{\mathrm{F},1}_{m,n} \rho_m. \tag{8.34}$$

When $\tau_{m,n}$ is executed by fog-computing processor of F_{T}, the corresponding energy consumption and latency are derived as follows:

$$G^{\mathrm{F},2}_{m,n} = \kappa_{\mathrm{F}} N_{m,n} \psi^2_{\mathrm{F}}, \; T^{\mathrm{F},2}_{m,n} = \frac{\zeta_{\mathrm{F}} N_{m,n}}{\psi_{\mathrm{F}}}, \tag{8.35}$$

where κ_{F} and ζ_{F} are the switched capacitance and CPU cycles of executing one bit, and the CPU-cycle frequency is denoted as ψ_{F}, respectively.

(3) Cloud Computation Offloading Mode $\tau_{m,n}$ is first forwarded to the F-AP, and then offloaded to the cloud computing center, when the cloud computation offloading mode is employed. When $\tau_{m,n}$ is forwarded to its associated F-AP F_{T}, the energy consumption and latency is the same as the offloading procedure in the fog

computation offloading mode, and thus they can be modeled as (8.33) and (8.34), respectively.

When $\tau_{m,n}$ is offloaded to the cloud computing center, the corresponding latency is given as follows:

$$T_{m,n}^{\mathrm{C},2} = \frac{N_{m,n}}{S_\mathrm{T}} = \frac{N_{m,n}}{B_\mathrm{T}\log_2\left(1+\frac{h_\mathrm{T}\rho_\mathrm{T}}{B_\mathrm{T}\omega_0}\right)}, \tag{8.36}$$

where S_T is rate of wireless backhaul, B_T is the bandwidth of wireless backhaul, h_T is the channel coefficient, and the transmit power is denoted as ρ_T. In addition, its energy consumption is expressed as follows:

$$G_{m,n}^{\mathrm{C},2} = T_{m,n}^{\mathrm{C},2}\rho_\mathrm{T}. \tag{8.37}$$

When $\tau_{m,n}$ is executed by the cloud computing processor, the energy consumption and latency caused by computation task processing are written as follows:

$$G_{m,n}^{\mathrm{C},3} = \kappa_\mathrm{C} N_{m,n}\psi_\mathrm{C}^2,\ T_{m,n}^{\mathrm{C},3} = \frac{\zeta_\mathrm{C} N_{m,n}}{\psi_\mathrm{C}}, \tag{8.38}$$

where κ_C and ζ_C are the switched capacitance and CPU cycles of processing one bit, respectively, and ψ_C denotes the CPU-cycle frequency of cloud computing processor.

In this part, the target is to jointly minimize the cost and latency of computation offloading in F-RANs, which can fully explore the computation capability in an efficient approach. Therefore, the objective function is established as a linear combination of energy consumption and latency. Based on (8.30)–(8.38), it can be expressed as follows:

$$\begin{aligned}\Phi_{m,n} &= \delta_T T_{m,n} + \delta_G G_{m,n}\\ &= \delta_T\big[w_{m,n}^\mathrm{L} T_{m,n}^\mathrm{L} + w_{m,n}^\mathrm{F}\big(T_{m,n}^{\mathrm{F},1} + T_{m,n}^{\mathrm{F},2}\big) + w_{m,n}^\mathrm{C}\big(T_{m,n}^{\mathrm{C},1} + T_{m,n}^{\mathrm{C},2} + T_{m,n}^{\mathrm{C},3}\big)\big]\\ &\quad +\delta_G\big[w_{m,n}^\mathrm{L} G_{m,n}^\mathrm{L} + w_{m,n}^\mathrm{F}\big(G_{m,n}^{\mathrm{F},1} + G_{m,n}^{\mathrm{F},2}\big) + w_{m,n}^\mathrm{C}\big(G_{m,n}^{\mathrm{C},1} + G_{m,n}^{\mathrm{C},2} + G_{m,n}^{\mathrm{C},3}\big)\big],\end{aligned} \tag{8.39}$$

where $w_{m,n}^\mathrm{L}$ is a Boolean variable to indicate the implementation status of local computation mode, i.e., when $w_{m,n}^\mathrm{L} = 1$, $\tau_{m,n}$ is processed by local computing processor, otherwise ${}_{m,n}^\mathrm{L} = 0$, the fog and cloud computation offloading modes are indicated in a similar way by $w_{m,n}^\mathrm{F}$ and $w_{m,n}^\mathrm{C}$.

Due to the limitation of communication and computation capability of each equipment in F-RANs, the following constraints should be considered for the joint optimization of computation offloading in F-RANs.

First, since each computation task is not splittable, the following constraint with respect the offloading decision indicators should be established:

$$w_{m,n}^{\mathrm{L}} + w_{m,n}^{\mathrm{F}} + w_{m,n}^{\mathrm{C}} = 1,\ w_{m,n}^{\mathrm{L}}, w_{m,n}^{\mathrm{F}}, w_{m,n}^{\mathrm{C}} \in \{0, 1\}. \tag{8.40}$$

Second, both the users and the F-AP should follow the individual constraints of transmit power, i.e.,

$$\rho_{\mathrm{T}}^{\min} \leqslant \rho_{\mathrm{T}} \leqslant \rho_{\mathrm{T}}^{\max},\ \rho_{m}^{\min} \leqslant \rho_{m} \leqslant \rho_{m}^{\max}, \tag{8.41}$$

where the minimum and maximum transmit power of F_{T} are denoted as $\rho_{\mathrm{T}}^{\min}$ and $\rho_{\mathrm{T}}^{\max}$, respectively, and $\rho_{m}^{\min}$ and $\rho_{m}^{\max}$ are the minimum and maximum transmit power of U_m.

Third, orthogonal radio frequency channels are occupied by the users. Due to the limitation of spectrum resource, the following constraint is formulated:

$$\sum_{m=1}^{M} B_m \leqslant B,\ B_m \geqslant 0,\ m = 1, \ldots, M. \tag{8.42}$$

Fourth, due to the processing restrict of each computing processor, the following constraint with respect to the processing speed should be established:

$$\begin{aligned} \psi_{\mathrm{L},m}^{\min} \leqslant \psi_{\mathrm{L},m} &\leqslant \psi_{\mathrm{L},m}^{\max}, \\ \psi_{\mathrm{F}}^{\min} \leqslant \psi_{\mathrm{F}} &\leqslant \psi_{\mathrm{F}}^{\max}, \\ \psi_{\mathrm{C}}^{\min} \leqslant \psi_{\mathrm{C}} &\leqslant \psi_{\mathrm{C}}^{\max}, \end{aligned} \tag{8.43}$$

where the minimum and maximum processing speed of $\psi_{\mathrm{L},m}$ are denoted as $\psi_{\mathrm{L},m}^{\min}$ and $\psi_{\mathrm{L},m}^{\max}$, respectively, $\psi_{\mathrm{F}}^{\min}$ and $\psi_{\mathrm{F}}^{\max}$ are the minimum and maximum processing speed of ψ_{F}, and $\psi_{\mathrm{C}}^{\min}$ and $\psi_{\mathrm{C}}^{\max}$ follow similar notations for ψ_{C}.

Finally, the processing capability of F-APs is also determined by its storage volume, i.e.,

$$\sum_{n=1}^{N} \sum_{m=1}^{M} w_{m,n}^{\mathrm{F}} N_{m,n} \leqslant N_{\mathrm{F}}, \tag{8.44}$$

where N_{F} denotes the maximum storage volume of the associated F-AP F_{T}.

Then, the joint optimization problem of computation offloading in F-RANs can be established:

$$\min_{\substack{w_{m,n}^{\mathrm{L}},\, w_{m,n}^{\mathrm{F}},\, w_{m,n}^{\mathrm{C}},\, \rho_{\mathrm{T}},\\ \rho_m,\, B_m,\, \psi_{\mathrm{L},m},\, \psi_{\mathrm{F}},\, \psi_{\mathrm{C}}}} \quad Q = \sum_{n=1}^{N} \sum_{m=1}^{M} \Phi_{m,n} \tag{8.45a}$$

$$s.t.\ (8.40), (8.41), (8.42), (8.43), (8.44). \tag{8.45b}$$

8.2.2 Resource Management and Offloading Decision Optimization

Recalling (8.45), the joint optimization problem is neither linear nor convex. To provide an efficient method to approach a stable optimal solution, we first decouple it into four independent subproblems, whose optimal solution can be obtained. Then, an iterative joint optimization algorithm is designed to solve (8.45).

CPU Frequency Optimization Subproblem

Since the processing and offloading procedures of computation tasks are independent, the CPU frequency optimization subproblem is not related to the radio resource management. Therefore, it can be reformed as follows:

$$\min_{\psi_i} \ Q_{1,i}(\psi_i) = \frac{\delta_T \zeta_i}{\psi_i} + \delta_G \kappa_i \psi_i^2 \tag{8.46a}$$

$$s.t.\ \psi_i^{\min} \leqslant \psi_i \leqslant \psi_i^{\max},\ i \in \left\{\tau_1, \ldots, \tau_M, \mathrm{F}, \mathrm{C}\right\}. \tag{8.46b}$$

Please note that (8.46) is a convex problem, and its convexity can be verified as follows:

$$\nabla_{\psi_i}^2 Q_{1,i}(\psi_i) = \frac{2\delta_T \zeta_i}{\psi_i^3} + 2\delta_G \kappa_i \geqslant 0. \tag{8.47}$$

Therefore, its optimal solution is tractable, which can be expressed as follows:

$$\psi_i^{\mathrm{opt}} = \begin{cases} \psi_i^{\min}, \ \psi_i^{\min} > \psi_i^* \\ \psi_i^*, \ \psi_i^{\min} \leqslant \psi_i^* \leqslant \psi_i^{\max}, \\ \psi_i^{\max}, \ \psi_i^* > \psi_i^{\max} \end{cases} \tag{8.48}$$

where ψ_i^* is the solution of equation $\nabla_{\psi_i} Q_{1,i}(\psi_i) = 0$, and it can be expressed as

$$\psi_i^* = \sqrt[3]{\frac{\delta_T \zeta_i}{2\delta_G \kappa_i}}. \tag{8.49}$$

Bandwidth Allocation Subproblem

Recalling (8.39), only the cost and performance of offloading procedure are related to results of bandwidth allocation. Then, this subproblem can be expressed as follows:

$$\min_{B_1,\ldots,B_M} \; Q_2\big(B_1,\ldots,B_M\big) = \sum_{m=1}^{M} \frac{T_m(\delta_T + \delta_G p_m)}{B_m \log_2\big(1 + \frac{h_m p_m}{B_m \omega_0}\big)} \tag{8.50a}$$

$$s.t. \; \sum_{m=1}^{M} B_m \leqslant B, \;\; B_m \geqslant 0, \tag{8.50b}$$

where $T_m = \sum_{n=1}^{N} N_{m,n}(w_{m,n}^{\mathrm{F}} + w_{m,n}^{\mathrm{C}})$.

To testify the convexity of (8.50), we first derive its Hessian matrix, which is formulated by the second-order derivatives with respect to $B_1, \ldots, B_M$, i.e.,

$$\mathbf{H} = \mathrm{diag}\big\{\nabla^2_{B_1} Q_2(B_1,\ldots,B_M), \ldots, \nabla^2_{B_M} Q_2(B_1,\ldots,B_M)\big\}. \tag{8.51}$$

When we focus on $\nabla^2_{B_m} Q_2(B_1,\ldots,B_M)$, it can be treated as a single-variant function with respect to B_m, i.e., $\nabla^2_{B_m} Q_2(B_1,\ldots,B_M) = \nabla^2_{B_m} Q_{2,m}(B_m)$. Recalling (8.50), it is a composition of the following two functions:

$$Q_{2,m}(B_m) = s_2(s_1(B_m)), \tag{8.52}$$

where

$$s_1(B_m) = B_m \log_2\left(1 + \frac{h_m p_m}{B_m \omega_0}\right), \;\; s_2(\nu) = \frac{T_m(\delta_T + \delta_G p_m)}{\nu}. \tag{8.53}$$

We can demonstrate that $s_1(B_m)$ is a concave function, while $s_2(\nu)$ is a convex function that keeps decreasing as ν increases. Based on Boyd et al. (2004), the convexity of $Q_{2,m}(B_m)$ can be verified. Therefore, $\nabla^2_{B_m} Q_{2,m}(B_m) > 0$, and the Hessian matrix given by (8.51) is a positive definite matrix, which means that the optimization problem given by (8.50) is convex. Hence, the optimal solution of (8.50) can be obtained straightforwardly by employing numerical methods, such as the interior point method introduced in Boyd et al. (2004).

Transmit Power Optimization Subproblem

When we focus on the transmit power optimization problem, (8.45) can be transformed into the following equivalent form:

$$\min_{\rho_i} \ Q_{3,i}(\rho_i) = \frac{\Phi_i(\delta_T + \delta_G \rho_i)}{B_i \log_2\left(1 + \frac{h_i \rho_i}{B_i \omega_0}\right)} \tag{8.54a}$$

$$s.t. \ \rho_i^{\min} \leqslant \rho_i \leqslant \rho_i^{\max}, \ i \in \left\{1, \ldots, M, \mathrm{T}\right\}, \tag{8.54b}$$

where Φ_i can be defined as

$$\Phi_i = \begin{cases} 1, \ i \in \{\mathrm{T}\} \\ \sum_{n=1}^{N} N_{i,n}(w_{i,n}^{\mathrm{F}} + w_{i,n}^{\mathrm{C}}) \end{cases}. \tag{8.55}$$

Equation (8.54) indicates that the optimization of transmit power is a nonlinear fractional of problem. To obtain its optimal solution, we first focus on the following transformed problem:

$$\min_{\rho_i} \ r(\rho_i) = \Phi_i(\delta_T + \delta_G \rho_i) - \beta^* B_i \log_2\left(1 + \frac{h_i \rho_i}{B_i \omega_0}\right) \tag{8.56a}$$

$$s.t. \ \rho_i^{\min} \leqslant \rho_i \leqslant \rho_i^{\max}, \ i \in \left\{1, \ldots, M, T\right\}, \tag{8.56b}$$

where $\beta*$ denotes the value of $Q_{3,i}(\rho_i)$ given by (8.54). The second-order derivative of object function given by (8.56) can be expressed as

$$\nabla_{\rho_i^2} r(\rho_i) = \frac{\beta^* B_i}{\ln 2\left(\frac{B_i \omega_0}{h_i} + \rho_i\right)^2} \geqslant 0. \tag{8.57}$$

Therefore, (8.56) is a convex optimization problem, and its optimal solution can be given as

$$\rho_i^* = \begin{cases} \rho_i^{\min}, \ \rho_i^{\min} > \bar{\rho}_i \\ \bar{\rho}_i, \ \rho_i^{\min} \leqslant \bar{\rho}_i \leqslant \bar{\rho}_i \\ \rho_i^{\max}, \ \bar{\rho}_i > \rho_i^{\max} \end{cases}, \tag{8.58}$$

where $\bar{\rho}_i$ is the solution of function $\nabla_{\rho_i} r(\rho_i) = 0$, and can be expressed as

$$\bar{\rho}_i = \frac{\beta^* B_i}{\Phi_i \ln 2(\delta_T + \delta_G \rho_i)} - \frac{B_i \omega_0}{h_i}. \tag{8.59}$$

The objective function of (8.56) is defined as a difference between the numerator and denominator of $Q_{3,i}(\rho_i)$. As introduced in Zhao et al. (2019), the minimum value of (8.54), which is denoted as $\beta*$, can be achieved when the following equation can be established:

$$\min_{\rho_i} \; g_1(\rho_i) - \beta^* g_2(\rho_i) = g_1(\rho_i^*) - \beta^* g_2(\rho_i^*) = 0, \tag{8.60}$$

where ρ_i^* is given by (8.58), $g_1(\rho_i) = \Phi_i(\delta_T + \delta_G \rho_i)$, $g_2(\rho_i) = B_i \log_2(1 + \frac{h_i \rho_i}{B_i \omega_0})$.

Based on (8.60), Algorithm 2 can be designed to approach the optimal solution of (8.54) in an iterative way. Without loss of generality, we focus on the k-th iteration. First, $\rho_i^{*(k)}$, which is the optimal solution of (8.56) for the k-th iteration, is obtained based on (8.58). Next, the value of $Q_{3,i}(\rho_i)$, which is denoted as $\beta*^{(k)}$, is updated as follows:

$$\beta^{*(k)} = \frac{g_1(\rho_i^{*(k)})}{g_2(\rho_i^{*(k)})}. \tag{8.61}$$

Finally, to testify (8.60), $h(\rho_i^{*(k)}, \beta^{*(k)})$ is denoted as the value of objective function of (8.56), and it can be updated as follows in the k-th iteration:

$$\begin{aligned} h(\rho_i^{*(k)}, \beta^{*(k)}) &= \min_{\rho_i^{(k)}} \; g_1\big(\rho_i^{(k)}\big) - \beta^{*(k-1)} g_2\big(\rho_i^{(k)}\big) \\ &= g_1\big(\rho_i^{*(k)}\big) - \beta^{*(k-1)} g_2\big(\rho_i^{*(k)}\big). \end{aligned} \tag{8.62}$$

The convergence of Algorithm 2 is proved by Theorem 2 in Zhao et al. (2019), which can approach the optimal solution of (8.54).

Algorithm 2 Transmit power allocation algorithm

```
**Initialize**: $\rho_i^{*(0)}$, $\beta*^{(0)}$, $K$, and the accuracy requirement $\varepsilon$.
For the $k$-th iteration:
    • Update $\rho\,_i^{*(k)}$ based on (8.58);
    • Update $\beta*^{(k)}$ based on (8.61);
    • Update $h(\rho\,_i^{*(k)},\ \beta\,^{*(k)})$ based on (8.62);
**Terminate**:$|h(\rho\,_i^{*(k)},\ \beta\,^{*(k)}) - h(\rho\,_i^{*(k-1)},\ \beta\,^{*(k-1)})| \leqslant \epsilon$ , or $k > K$.
```

Offloading Decision Subproblem

Based on (8.45), the offloading decision subproblem can be formulated as the following 0-1 programming problem:

$$\min_{w_{m,n}^{\mathrm{L}}, w_{m,n}^{\mathrm{F}}, w_{m,n}^{\mathrm{C}}} \quad \sum_{v \in \{\mathrm{L,F,C}\}} \sum_{n=1}^{N} \sum_{m=1}^{M} \Phi_{m,n}^{i} w_{m,n}^{i} \tag{8.63a}$$

$$s.t. \;\; w_{m,n}^{\mathrm{L}} + w_{m,n}^{\mathrm{F}} + w_{m,n}^{\mathrm{C}} = 1, \tag{8.63b}$$

$$w_{m,n}^{\mathrm{L}},\, w_{m,n}^{\mathrm{F}},\, w_{m,n}^{\mathrm{C}} \in \{0,\, 1\}, \tag{8.63c}$$

$$\sum_{n=1}^{N} \sum_{m=1}^{M} w_{m,n}^{\mathrm{F}} N_{m,n} \leqslant N_{\mathrm{F}}, \tag{8.63d}$$

where

$$\begin{aligned}
\Phi_{m,n}^{\mathrm{L}} &\quad \delta_T T_{m,n}^{\mathrm{L}} + \delta_G G_{m,n}^{\mathrm{L}}, \\
\Phi_{m,n}^{\mathrm{F}} &= \delta_T \left(T_{m,n}^{\mathrm{F},1} + T_{m,n}^{\mathrm{F},2}\right) + \delta_G \left(G_{m,n}^{\mathrm{F},1} + G_{m,n}^{\mathrm{F},2}\right), \\
\Phi_{m,n}^{\mathrm{C}} &= \delta_T \left(T_{m,n}^{\mathrm{C},1} + T_{m,n}^{\mathrm{C},2} + T_{m,n}^{\mathrm{C},3}\right) + \delta_G \left(G_{m,n}^{\mathrm{C},1} + G_{m,n}^{\mathrm{C},2} + G_{m,n}^{\mathrm{C},3}\right).
\end{aligned}$$

The existing works usually employ implicit enumeration methods to solve the 0-1 problems as (8.63), which causes high computational complexity that increases exponentially with respect to the number of computation tasks (Korte et al. 2008). To solve it efficiently, the following relaxed problem is formulated:

$$\min_{w_{m,n}^{\mathrm{L}}, w_{m,n}^{\mathrm{F}}, w_{m,n}^{\mathrm{C}}} \quad \sum_{v \in \{\mathrm{L,F,C}\}} \sum_{n=1}^{N} \sum_{m=1}^{M} \Phi_{m,n}^{i} w_{m,n}^{i} \tag{8.64a}$$

$$s.t. \;\; w_{m,n}^{\mathrm{L}} + w_{m,n}^{\mathrm{F}} + w_{m,n}^{\mathrm{C}} = 1, \tag{8.64b}$$

$$w_{m,n}^{\mathrm{L}},\, w_{m,n}^{\mathrm{F}},\, w_{m,n}^{\mathrm{C}} \in \{0,\, 1\}. \tag{8.64c}$$

Equation (8.64) is obtained by removing the F-AP storage volume constraint, and it can be solved by processing each computation task via the mode that can minimize the cost, i.e.,

$$\bar{w}_{m,n}^{v} = \begin{cases} 1, & \Phi_{m,n}^{v} = \min\left\{\Phi_{m,n}^{\mathrm{L}},\; \Phi_{m,n}^{\mathrm{F}},\; \Phi_{m,n}^{\mathrm{C}}\right\} \\ 0, & \text{else} \end{cases}, \tag{8.65}$$

where $v \in \{\mathrm{L,\ F,\ C}\}$. As shown in (8.65), based on the processing modes, all the computation tasks can be divided into three sets, which can be denoted as $\mathscr{S}_{\mathrm{L}}$, $\mathscr{S}_{\mathrm{F}}$, and $\mathscr{S}_{\mathrm{C}}$, respectively. In particular, $\mathscr{S}_{\mathrm{L}}$ denotes the set of computation tasks that are handled by local computation mode, $\mathscr{S}_{\mathrm{F}}$ and $\mathscr{S}_{\mathrm{C}}$ are defined similarly for the fog and cloud computation offloading modes, respectively, i.e.,

$$\mathscr{S}_{v} = \{\tau_{m,n} | \bar{w}^{v}_{m,n} = 1 \text{ in } (8.65)\}, \tag{8.66}$$

where $m = 1, \ldots, M$, $n = 1, \ldots, N$ and $v \in \{\mathrm{L}, \mathrm{F}, \mathrm{C}\}$.

When the storage volume constraint (8.63d) is followed, (8.65) is also the optimal solution of original computation offloading problem given by (8.63), i.e., $w^{v*}_{m,n} = \bar{w}^{v}_{m,n}$. If (8.63d) is not satisfied, the optimal solution can be obtained by adjusting the elements in $\mathscr{S}_{\mathrm{F}}$, which aims to ensure that the data volume of $\mathscr{S}_{\mathrm{F}}$ can satisfy the constraint (8.63) with the minimum increment of costs after adjustment. To capture the movement of adjustment, a Boolean indicator is employed for each computation task in $\mathscr{S}_{\mathrm{F}}$, which can be defined as

$$x_{m,n} = \begin{cases} 1, & \tau_{m,n} \text{ is moved to } \mathscr{S}_{\mathrm{L}} \textit{ or } \mathscr{S}_{\mathrm{C}} \\ 0, & \tau_{m,n} \text{ stays in } \mathscr{S}_{\mathrm{F}} \end{cases}. \tag{8.67}$$

To obtain the optimal offloading decision to minimize the cost, the following optimization problem is established to adjust the solution of relaxed problem given by (8.65):

$$\min_{x_{m,n}} \sum_{\tau_{m,n} \in \mathscr{S}_{\mathrm{F}}} C_{m,n} x_{m,n} \tag{8.68a}$$

$$s.t. \sum_{\tau_{m,n} \in \mathscr{S}_{\mathrm{F}}} N_{m,n} x_{m,n} \geqslant \sum_{\tau_{m,n} \in \mathscr{S}_{\mathrm{F}}} N_{m,n} - N_{\mathrm{F}}, \tag{8.68b}$$

$$x_{m,n} \in \{0, 1\}, \tag{8.68c}$$

where $C_{m,n} = \min\{\Phi^{\mathrm{L}}_{m,n} - \Phi^{\mathrm{F}}_{m,n}, \Phi^{\mathrm{C}}_{m,n} - \Phi^{\mathrm{F}}_{m,n}\}$. By replacing $y_{m,n}$ by $y_{m,n} = 1 - x_{m,n}$, (8.68) can be rewritten as

$$\max_{y_{m,n}} \sum_{\tau_{m,n} \in \mathscr{S}_{\mathrm{F}}} C_{m,n} y_{m,n} \tag{8.69a}$$

$$s.t. \sum_{\tau_{m,n} \in \mathscr{S}_{\mathrm{F}}} N_{m,n} y_{m,n} \leqslant 2 \sum_{\tau_{m,n} \in \mathscr{S}_{\mathrm{F}}} N_{m,n} - N_{\mathrm{F}}, \tag{8.69b}$$

$$y_{m,n} \in \{0, 1\}. \tag{8.69c}$$

Equation (8.69) is a knapsack problem, and it can be solved by dynamic programming. Then the final offloading decision results can be expressed as follows:

$$\begin{cases} w_{m,n}^{\mathrm{L}*} = 0,\ w_{m,n}^{\mathrm{F}*} = 1,\ w_{m,n}^{\mathrm{C}*} = 0,\ \text{if } y_{m,n}^{*} = 0 \\ w_{m,n}^{\mathrm{L}*} = 0,\ w_{m,n}^{\mathrm{F}*} = 0,\ w_{m,n}^{\mathrm{C}*} = 1,\ \text{if } y_{m,n}^{*} = 1 \text{ and } \Phi_{m,n}^{\mathrm{C}} < \Phi_{m,n}^{\mathrm{L}} \\ w_{m,n}^{\mathrm{L}*} = 1,\ w_{m,n}^{\mathrm{F}*} = 0,\ w_{m,n}^{\mathrm{C}*} = 0,\ \text{if } y_{m,n}^{*} = 1 \text{ and } \Phi_{m,n}^{\mathrm{C}} \geqslant \Phi_{m,n}^{\mathrm{L}} \end{cases} ,\ \tau_{m,n} \in \mathscr{S}_{\mathrm{F}}. \tag{8.70}$$

As introduced in Zhao et al. (2019), (8.70) is the optimal solution of the original offloading decision subproblem given by (8.63), which can minimize the cost of computation offloading in F-RANs.

Joint Optimization Algorithm

Algorithm 3 is designed to approach the optimal solution of the joint optimization problem given by (8.45). In particular, the optimization variables with respect to CPU frequency, bandwidth and transmit power allocation, and offloading decision are updated based on solutions of decoupled subproblems. Since the optimal solutions of all the subproblems can be obtained, it guarantees that the cost defined by the objective function of (8.45) keeps decreasing as the time of iterations increases. Meanwhile, there exists a finite lower bound of the costs. Therefore, Algorithm 3 can converge to a stable optimal solution of (8.45).

Algorithm 3 Joint optimization of resource management and offloading decision

```
Initialization: $w_{m,n}^{\mathrm{L}}(0)$, $w_{m,n}^{\mathrm{F}}(0)$, $w_{m,n}^{\mathrm{C}}(0)$, $\rho_{\mathrm{T}}(0)$, $\rho_m(0)$, $B_m(0)$,
$\psi_{\mathrm{L},m}(0)$, $\psi_{\mathrm{F}}(0)$, $\psi_{\mathrm{C}}(0)$, $Q(0)$.
For the $l$-th iteration:
    • Update $\psi_{\tau\ m}(l)$, $\psi_{\mathrm{F}}(l)$, and $\psi_{\mathrm{C}}(l)$ by following (8.48);
    • Update $B_m(l)$ based on (8.50);
    • Update $\rho_{\mathrm{T}}(l)$ and $\rho_m(l)$ by following Algorithm 1;
    • Update $w_{m,n}^{L}(l)$, $w_{m,n}^{F}(l)$, and $w_{m,n}^{C}(l)$ based on (8.70);
    • Update $Q(l)$ based on (8.45).
Termination: When $l > L$ or $|Q(l) - Q(l-1)| \leqslant \varepsilon$ .
```

8.3 Simulation Results

To evaluate the performance of computation offloading in F-RANs, the performance of theoretical expected latency and joint optimization algorithm is provided in this part.

8.3.1 Simulation Results of Latency Performance

In this part, the densities of F-APs and cloud computing centers are $\rho_f = 10^{-4}$ and $\rho_c = 10^{-5}$, respectively. Moreover, the number of served users is set as 4 for each F-AP. The processing speeds can be given as $s_n = 2.5$ Mbit/s, $s_F = 7.5$ Mbit/s, and $s_C = 45$ Mbit/s for the local, the fog, and the cloud computing processors, respectively (Zhao et al. 2019). The average data volume is set as $\phi = 5$ Kbits, the bandwidth is $W = 4$ MHz, and the exponent of path loss is set $\alpha = 3$.

The offloading probability is plotted in Fig. 8.3. As shown in the figure, the offloading probability decreases as ρ_f increases, since the severe inference is caused when the F-APs are deployed densely. Moreover, the theoretical results coincide with the Monte-Carlo results, and thus its accuracy can be demonstrated. Finally, when the output rate of computation tasks generation queue is high, which implies that each user has a lot of computation tasks, the offloading probability is low since it is challenging to satisfy the strict requirement of computation offloading.

To evaluate the utility of computation offloading in F-RANs, the expected latency is proved in Fig. 8.4. Since the theoretical results can match with the simulation results, the accuracy of derived theoretical latency performance can be testified. In addition, it causes long latency when the number of served user is large, since the loadings of processing offloaded computation tasks are heavy. For the same reason, the latency performance is worsen when the generation rate of r_n is high.

8.3.2 Simulation Results of Joint Optimization Algorithm

In this part, the simulation results of joint optimization algorithm are provided. The key parameters are set as follows: The data volume of computation tasks is generated based on uniform distributions, and volume of storage of each F-AP is $N_{\mathrm{F}} = 4$ Gbits. Based on Mao et al. (2017), the CPU cycles that are used to process one bit is $\zeta_{\mathrm{L},m} = \zeta_{\mathrm{F}} = \zeta_{\mathrm{C}} = 737.5$ cycles/bit, and the switched capacitance of the user, the fog, and the cloud computing processors are $\kappa_{\mathrm{L},m} = 2 \times 10^{-26}$, $\kappa_{\mathrm{F}} = 7 \times 10^{-28}$, $\kappa_{\mathrm{C}} = 5 \times 10^{-29}$, respectively. The bandwidth of wireless access and backhaul links can be given as $B = 10$ MHz and $B_{\mathrm{T}} = 20$ MHz, and the range of transmit power of F-AP is [30,40] dBm, and [20,30] dBm for the user devices.

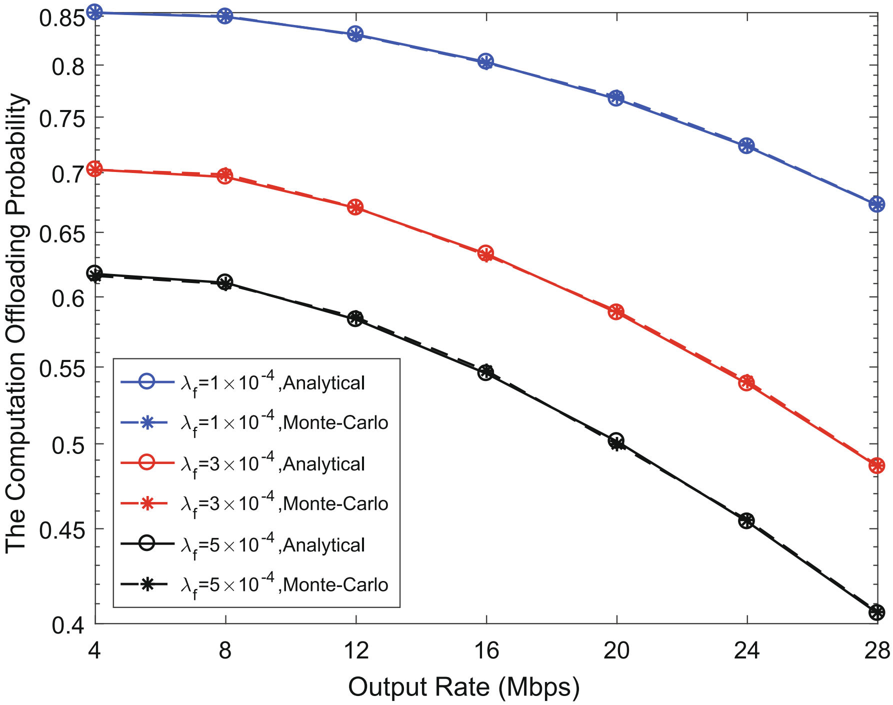


Fig. 8.3 The computation offloading probability in F-RANs

The convergence of joint optimization algorithm is verified by Fig. 8.5 when the user number is $M = 3$. It shows that the cost keeps decreasing as the iteration time increases, and finally converges to the optimal performance. In particular, the simulation results show that the algorithm converges within 12 iterations.

In Figs. 8.6 and 8.7, the performance of joint optimization algorithm is provided. In particular, the benchmarks are set as follows: (1) The random offloading scheme (Benchmark 1), where the offloading decision is made randomly without resource allocation optimization. (2) Optimization scheme of offloading mode selection and CPU frequency (Benchmark 2). (3) Optimization scheme of CPU frequency, and power and bandwidth allocation (Benchmark 3). As shown in the figures, compared with the benchmarks, the joint optimization can achieve the lowest cost and latency of computation offloading in F-RANs.

To evaluate the performance of computation offloading in F-RANs, the comparisons with the single-tier MCC and MEC systems are plotted in Figs. 8.8 and 8.9. In Fig. 8.8, the total cost performance of computation offloading is provided. In particular, the total cost can be lowered by 76 by employ F-RANs, compared with the MCC systems, while it can be reduced by 39 compared with the MEC systems. The latency performance is provided by Fig. 8.9. As shown in the figure, the latency of computation offloading is reduced by 43 s and 27 s, respectively.

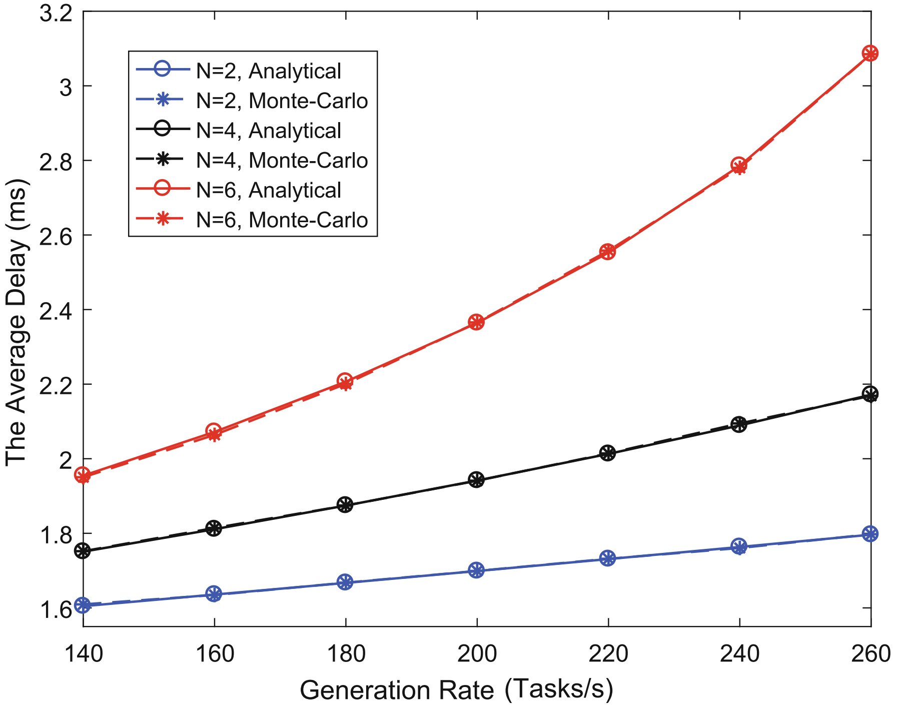


Fig. 8.4 The average delay performance of computation offloading in F-RANs

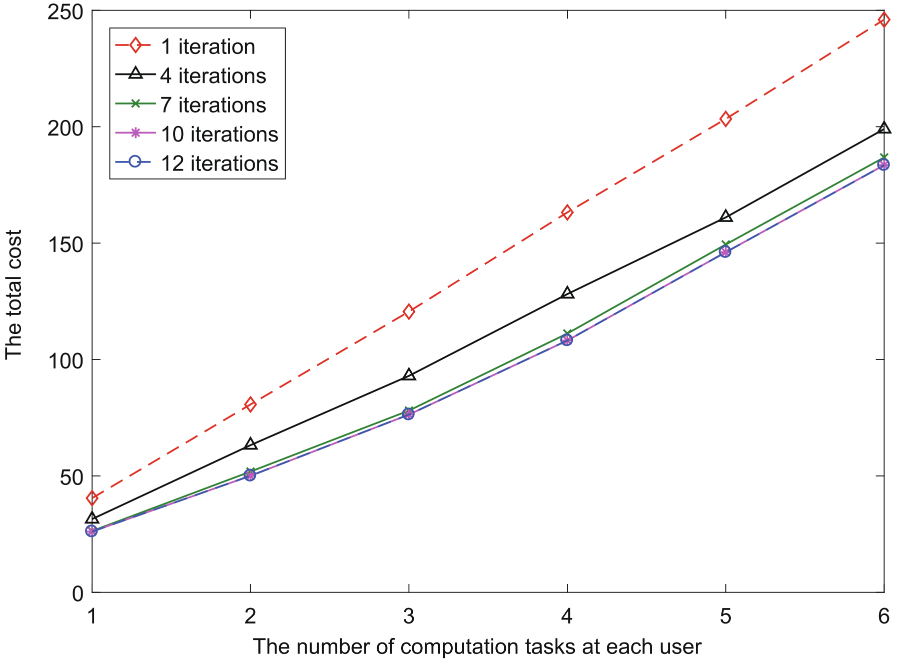


Fig. 8.5 The convergence performance of joint optimization algorithm

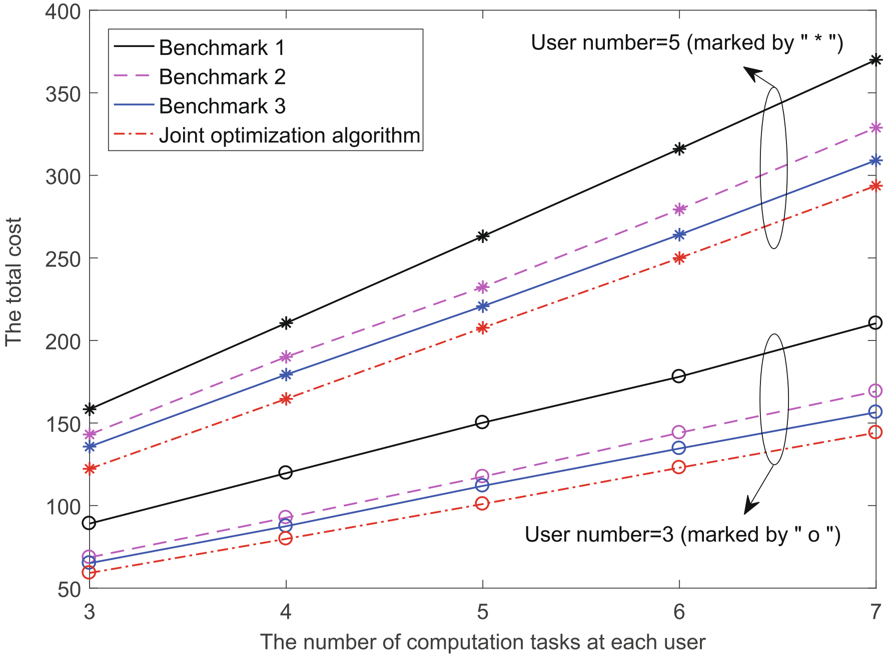


Fig. 8.6 The evaluation of total cost performance (Zhao et al. 2019)

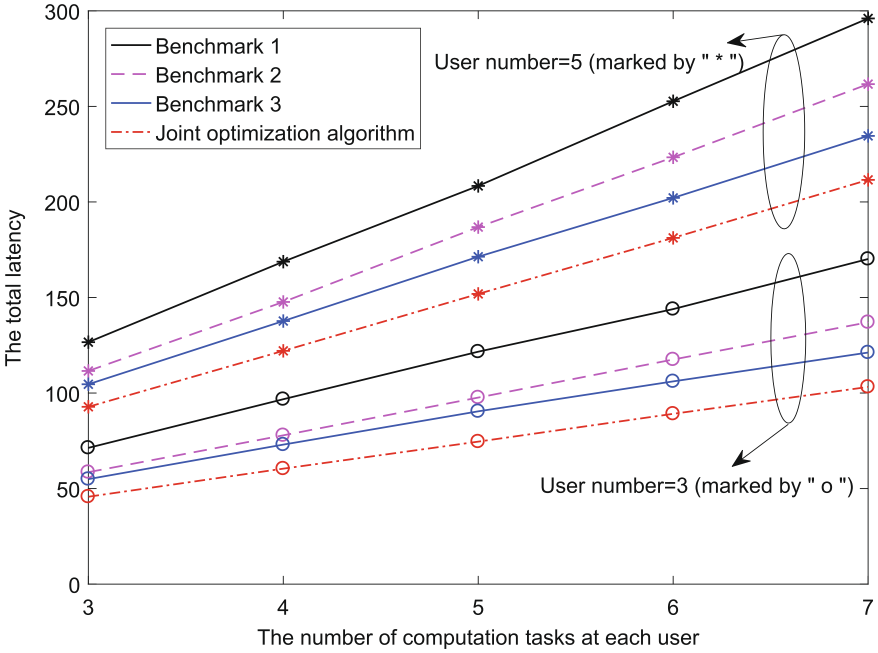


Fig. 8.7 The evaluation of the total latency performance (Zhao et al. 2019)

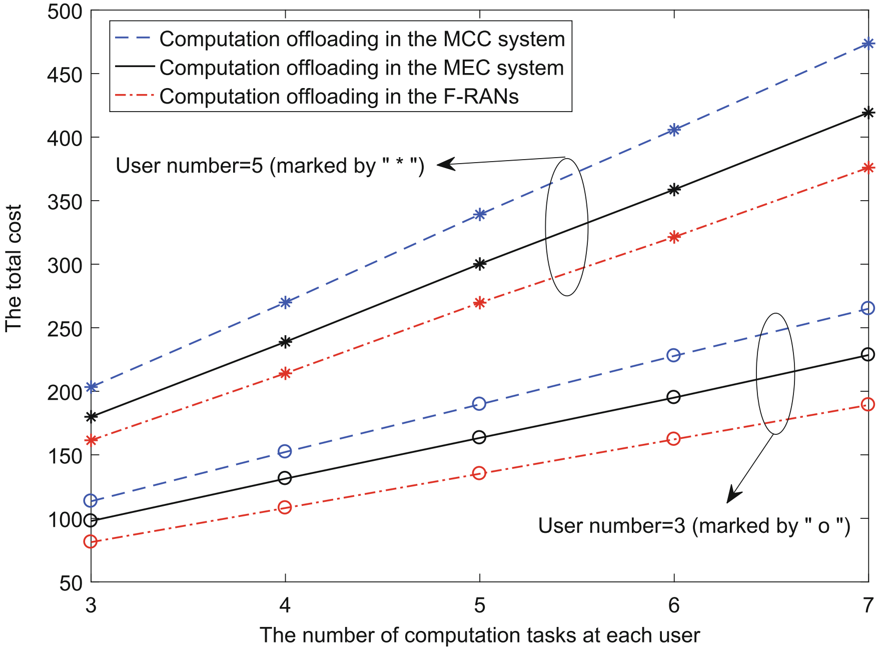


Fig. 8.8 The total cost performance compared to MCC and MEC systems (Zhao et al. 2019)

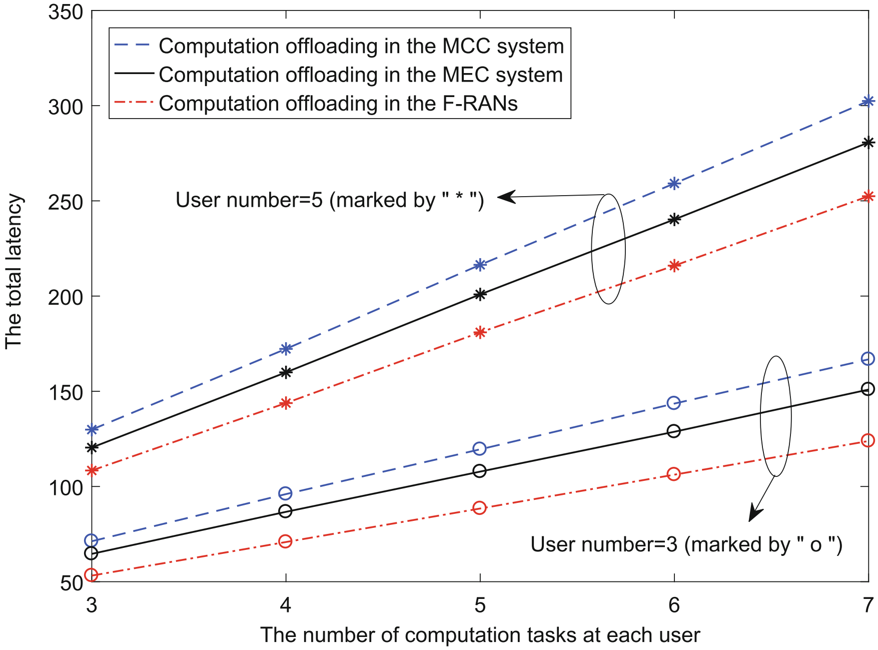


Fig. 8.9 The total latency performance compared to MCC and MEC systems

8.4 Summary

In this part, the joint computation offloading in F-RANs has been discussed, where the expected latency performance and joint optimization algorithm are provided. The analytical performance shows that the offloading decision should be made based on the processing capability of computing processors, which can keep a balance between the execution efficiency and offloading cost. Moreover, to further explore the potential of computation capability of F-RANs, the joint optimization of offloading decision and radio resource management has been designed, and an efficient algorithm has been proposed to approach the stable optimal result.

References

Balter M (2013) Performance modeling and design of computer systems: queuing theory in action. Cambridge University Press, New York

Boyd S et al (2004) Convex optimization. Cambridge University Press, Cambridge

Burd T et al (1996) Processor design for portable systems. J VLSI Signal Process Syst 13:203–221

Ganti R et al (2009) Interference and outage in clustered wireless ad hoc networks. IEEE Trans Inf Theory 55:4067–4086

Haenggi M (2013) Stochastic geometry for wireless networks. Cambridge University Press, New York

Korte B et al (2008) Combinatorial optimization: theory and algorithms. Springer, Germany

Mao Y et al (2017) Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems. IEEE Trans Wireless Commun 16:5994–6009

Zhao Z et al (2019) On the design of computation offloading in fog radio access networks. IEEE Trans Veh Technol 68:7136–7149

Chapter 9
Prototype Design of Fog Radio Access Networks

In Sect. 9.1, we discuss the basics for implementing an F-RAN test bed, including the realization of fog computing and the network controller. In Sect. 9.2, useful development tools will be briefly introduced, such as software-defined radio software and database software. In Sect. 9.3, an F-RAN test bed example will be elaborated, in terms of hardware platforms, client software, and featured functionalities. At last, the performance of the designed test bed will be evaluated in Sect. 9.4.

9.1 Design Basics

In this section, the design basics of F-RAN test beds are elaborated, mainly focusing on fog computing enablers and the network controller.

9.1.1 Enabling Fog Computing

According to Peng et al. (2016), a fog access point (F-AP) has capabilities of communication and fog computing. From the perspective of communication, the FAP needs to execute a four-layer stack, including PHY, MAC, RLC, and PDCP, which can follow current related 3GPP standards. PHY is responsible for coding/decoding, modulation/demodulation, and so on. MAC and RLC process MAC protocol data unit and RLC service data unit, respectively, while header compression and decompression are performed at PDCP.

For traditional base stations, like eNodeB, a GTP-U header will be added to user packets after PDCP processing and will be then sent to S-GW using UDP/IP protocols. However, to be benefited from fog computing at F-APs, user packets

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0_9

requesting an application locally available should be able to be re-forwarded to the proper edge application server. Intuitively, there are two ways to achieve this goal. The first one is to introduce an additional forwarding module between PDCP processing and the GTP stack, which requires a modification in base stations (Huang et al. 2017). The second one is to develop a middlebox over the S1-U interface between a base station and S-GW without disturbing current network equipment implementations (Li et al. 2018).

The principle of the first solution is summarized in Fig. 9.1. When a user requests an application for the first time, an application-aware function, integrated within the F-AP, acquires user data processed at PDCP layer and then extracts the IP address and port number of the requested application server. A list is maintained containing pre-stored pieces of $\langle$IP – port number – application name$\rangle$. If the requested $\langle$IP A – port number B – application $C\rangle$ is in this list, a notification will be sent to the fog computing platform at the F-AP, which then creates desired applications based on virtualization techniques with a local IP address and port number allocated. After that, the local IP address and port number are reported to the remote application server that sends them back to the user. Next, the user requests the application with the received information. If the destination IP extracted by the application-aware function is a local IP, the forwarding module will send user packets to the application located in the fog computing platform and just makes user data go to the GTP-U stack otherwise.

The principle of the second solution is summarized in Fig. 9.2. In this solution, the F-AP first sends user packets on the user's GTP tunnel to a middlebox, which clarifies received packets into one of three categories, i.e., packets going to the DNS server, local application servers, and the core network, based on the destination port number and destination IP addresses. Then packets are forwarded according to the classification result. To achieve communication between the F-AP and the middlebox at the link layer, address resolution protocol (ARP) is utilized to respond to the ARP request from the F-AP, which links the MAC address of the middlebox to the S-GW's IP address. Once user packets are received, the re-forwarding procedure

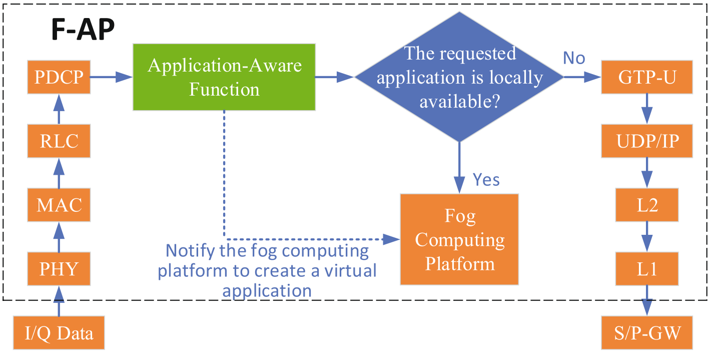

F-AP
PDCP
RLC
MAC
PHY
I/Q Data
Application-Aware Function
The requested application is locally available?
No
Yes
GTP-U
UDP/IP
L2
L1
S/P-GW
Fog Computing Platform
Notify the fog computing platform to create a virtual application

Fig. 9.1 The first solution to enable fog computing in literatures

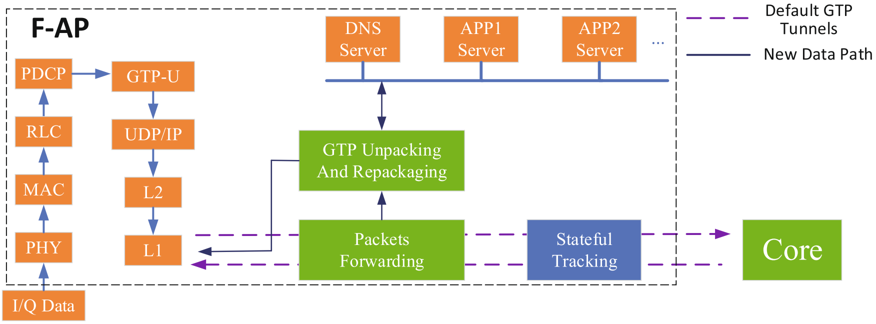


Fig. 9.2 The second solution to enable fog computing in literatures

is achieved with the help of an edge DNS server. Specifically, if the domain name of the requested application server is locally available, the DNS server feeds the local IP address of the application back to the user, while the DNS server forwards DNS queries to other remote servers otherwise. For the packets going to the local application server, the GTP header is removed and user IP packets are finally served. When the edge application server sends data to users, packets are repackaged with the corresponding GTP header according to the mapping between user IP and its tunnel ID, and such a mapping can be got via stateful tracking.

9.1.2 Network Controller

The network controller is to facilitate better management and performance optimization of F-RANs, which includes a network state monitoring module, a policy management module, a self-optimization module, and a network configuration module.

The network state monitoring module is able to be aware of the operation status of the whole F-RAN, including traffic load, coverage, throughput, handover rate, reference signal receiving power, and so on. In addition, this module can detect different network faults like cell outage and analyzes the root cause. The policy management module stores various resource management, mobility management, and routing policies. In terms of radio resource management, common policies are those based on round-robin, proportional fairness, and signal-to-interference-plus-noise ratio. For cache resource management, one can consider policies like caching mostly requested contents and random caching, while computation resource policies decide the rules at F-APs to serve user's edge computing requests, such as the first-in-first-served policy. The mobility management policies provide guidelines on the conditions under which a user should be handed over from the current F-AP to another F-AP. At last, the routing policies refer to various protocols to forward user's content or edge computation requests among connected FAPs.

Once the network state monitoring module detects F-RAN performance degradation, a network optimization procedure will be triggered, which is handled by the self-optimization module. The parameters to be optimized include spectrum allocation among FAPs, handover related parameters, transmit power of FAPs, the cache update duration, and so on. The optimized parameter values will then go to the network configuration module. Similar to the software-defined-networking paradigm, there is an interface between the network configuration module and each F-AP, facilitating flexible parameter or policy adjustment.

9.2 Useful Development Tools

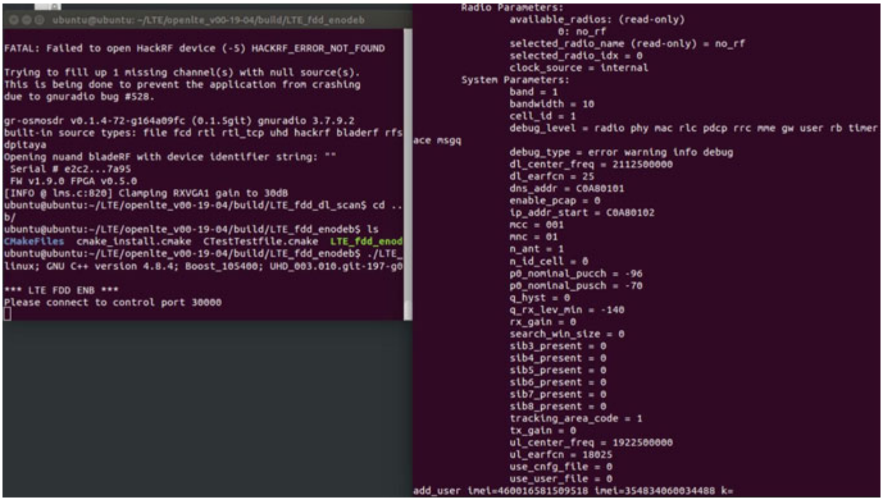
In this section, several development tools to build F-RAN test beds are introduced, including soft-defined-radio (SDR) software, database software, and so on.

9.2.1 SDR Software

Traditional communication platform is composed of dedicated hardware equipment. Compared with conventional solutions, communication platforms based SDR provides low-cost and flexible solutions, which implements functionalities of user equipment, base stations, and core networks in a softwarization manner. The most popular SDR solutions include OpenAirInterface, SRS, and OpenLTE.

OpenLTE

OpenLTE is an open-source project of 3GPP communication protocols implemented in Linux systems. It was initiated by Ben Wojtowicz, a former Motorola engineer, in 2011, and it mainly accomplishes the functions of a simple 4G base station. In PHY layer, OpenLTE supports all physical channels that follow the LTE release 10 standard protocol. Encoding, decoding, modulation, demodulation, multi-antenna mapping are realized. However, further development is still needed to support multiple-input-multiple-output and multimedia broadcast multicast services. As for MAC layer, which is responsible for controlling HARQ retransmission and scheduling of uplink and downlink, OpenLTE uses a simple polling scheduling algorithm. Overall, OpenLTE has some imperfections in the MAC layer processing, which affects the robustness of the software. OpenLTE has fully implemented the functions above MAC layer. Figure 9.3 is a snapshot of the software user interface.



```
FATAL: Failed to open HackRF device (-5) HACKRF_ERROR_NOT_FOUND

Trying to fill up 1 missing channel(s) with null source(s).
This is being done to prevent the application from crashing
due to gnuradio bug #528.

gr-osmosdr v0.1.4-72-g164a09fc (0.1.5git) gnuradio 3.7.9.2
built-in source types: file fcd rtl rtl_tcp uhd hackrf bladerf rfs
dpitaya
Opening nuand bladeRF with device identifier string: ""
 Serial # e2c2...7a95
 FW v1.9.0 FPGA v0.5.0
[INFO @ lms.c:820] Clamping RXVGA1 gain to 30dB
ubuntu@ubuntu:~/LTE/openlte_v00-19-04/build/LTE_fdd_dl_scan$ cd ..
b/
ubuntu@ubuntu:~/LTE/openlte_v00-19-04/build/LTE_fdd_enodeb$ ls
CMakeFiles  cmake_install.cmake  CTestTestfile.cmake  LTE_fdd_enod
ubuntu@ubuntu:~/LTE/openlte_v00-19-04/build/LTE_fdd_enodeb$ ./LTE_
linux; GNU C++ version 4.8.4; Boost_105400; UHD_003.010.git-197-g0

*** LTE FDD ENB ***
Please connect to control port 30000
```

```
        Radio Parameters:
                available_radios: (read-only)
                        0: no_rf
                selected_radio_name (read-only) = no_rf
                selected_radio_idx = 0
                clock_source = internal
        System Parameters:
                band = 1
                bandwidth = 10
                cell_id = 1
                debug_level = radio phy mac rlc pdcp rrc mme gw user rb timer
ace msgq
                debug_type = error warning info debug
                dl_center_freq = 2112500000
                dl_earfcn = 25
                dns_addr = C0A80101
                enable_pcap = 0
                ip_addr_start = C0A80102
                mcc = 001
                mnc = 01
                n_ant = 1
                n_id_cell = 0
                p0_nominal_pucch = -96
                p0_nominal_pusch = -70
                q_hyst = 0
                q_rx_lev_min = -140
                rx_gain = 0
                search_win_size = 0
                sib3_present = 0
                sib4_present = 0
                sib5_present = 0
                sib6_present = 0
                sib7_present = 0
                sib8_present = 0
                tracking_area_code = 1
                tx_gain = 0
                ul_center_freq = 1922500000
                ul_earfcn = 18025
                use_cnfg_file = 0
                use_user_file = 0
add_user imei=460016581509518 imei=354834060034488 k=
```

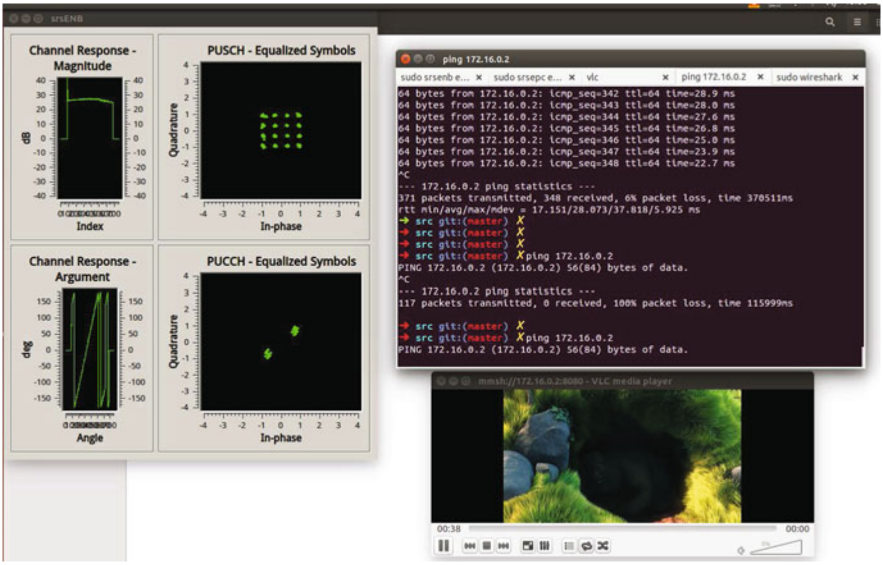
Fig. 9.3 The OpenLTE user interface

SrsLTE

SrsLTE is a free and high-performance LTE library for software defined radio applications, which is developed by software radio systems (SRS) in Ireland and uses the relevant functions of OpenLTE in the implementation. It currently provides interfaces to the universal hardware driver, supporting the Ettus USRP family of devices. The common features provided by the srsLTE library are:

- LTE Release 10 aligned.
- Detailed log system with per-layer log levels and hex dumps.
- MAC, RLC, PDCP, RRC, NAS, S1AP, and GW layers.
- Supported bandwidths: 1.4, 3, 5, 10, 15, and 20 MHz.
- Highly optimized Turbo decoder.
- MAC layer Wireshark packet capture.
- Channel simulator for EPA, EVA, and ETU 3GPP channels.

SrsLTE code is mainly divided into three modules, namely srsUE, srsENB, and srsEPC. The srsLTE features FDD and TDD configuration and carrier aggregation support. With 20 MHz bandwidth, downlink data rate can achieve 150 Mbps under MIMO TM3/TM4, while it can reach 75 Mbps under SISO configuration. SrsENB is a complete software radio LTE eNodeB, building upon the srsLTE library, and it includes L1, L2, and L3 layers. In L3 layer, SrsENB implements RRC, S1-AP, and GTP-U protocols, which manage the control plane between eNodeB and connected UEs as well as the core network and provide the data plane connection between eNodeB and the core network. The SrsENB also supports FDD configuration, Round-Robin MAC scheduling, periodic and aperiodic channel quality indicator feedback, user plane encryption, and so on. SrsEPC is a light-weight and complete


srsENB
Channel Response - Magnitude
dB
Index
PUSCH - Equalized Symbols
Quadrature
In-phase
Channel Response - Argument
deg
Angle
PUCCH - Equalized Symbols
ping 172.16.0.2
sudo srsenb e...
sudo srsepc e...
vlc
ping 172.16.0.2
sudo wireshark
mmsh://172.16.0.2:8080 - VLC media player
00:38
00:00

```
64 bytes from 172.16.0.2: icmp_seq=342 ttl=64 time=28.9 ms
64 bytes from 172.16.0.2: icmp_seq=343 ttl=64 time=28.0 ms
64 bytes from 172.16.0.2: icmp_seq=344 ttl=64 time=27.6 ms
64 bytes from 172.16.0.2: icmp_seq=345 ttl=64 time=26.8 ms
64 bytes from 172.16.0.2: icmp_seq=346 ttl=64 time=25.0 ms
64 bytes from 172.16.0.2: icmp_seq=347 ttl=64 time=23.9 ms
64 bytes from 172.16.0.2: icmp_seq=348 ttl=64 time=22.7 ms
^C
--- 172.16.0.2 ping statistics ---
371 packets transmitted, 348 received, 6% packet loss, time 370511ms
rtt min/avg/max/mdev = 17.151/28.073/37.818/5.925 ms
➜ src git:(master) ✗
➜ src git:(master) ✗
➜ src git:(master) ✗
➜ src git:(master) ✗ping 172.16.0.2
PING 172.16.0.2 (172.16.0.2) 56(84) bytes of data.
^C
--- 172.16.0.2 ping statistics ---
117 packets transmitted, 0 received, 100% packet loss, time 115999ms

➜ src git:(master) ✗
➜ src git:(master) ✗ping 172.16.0.2
PING 172.16.0.2 (172.16.0.2) 56(84) bytes of data.
```

Fig. 9.4 The SrsLTE user interface

realization of an LTE core network (EPC), including a mobility management entity, a home subscriber server, and S/P-GWs. Figure 9.4 is a snapshot of the software user interface.

OpenAirInterface

As a continuously updating open-source SDR project, OpenAirInterface (OAI) is organized, developed, and maintained by EURECOM in France. This project has realized network functions spanning all the layers using C programming language in Linux systems, which aims to provide an open and integrated development environment as well as the flexibility to architect, instantiate, and configure the network components at the edge or cloud. In total, the OAI software is divided into five parts as follows.

- OpenAir0: This folder mainly describes hardware modules and corresponding hardware drivers for FPGA.
- OpenAir1: This folder contains function modules at physical layers along with specific parameters and initialization. Processing baseband signal and providing interfaces of MAC layer are two key features implemented in this code. Further, it also includes simulation environments and channel models to test the code or do performance simulations.
- OpenAir2: This folder includes upper protocol layers, such as MAC, RLC, and PDCP, and hence it is primarily responsible for wireless access control, wireless

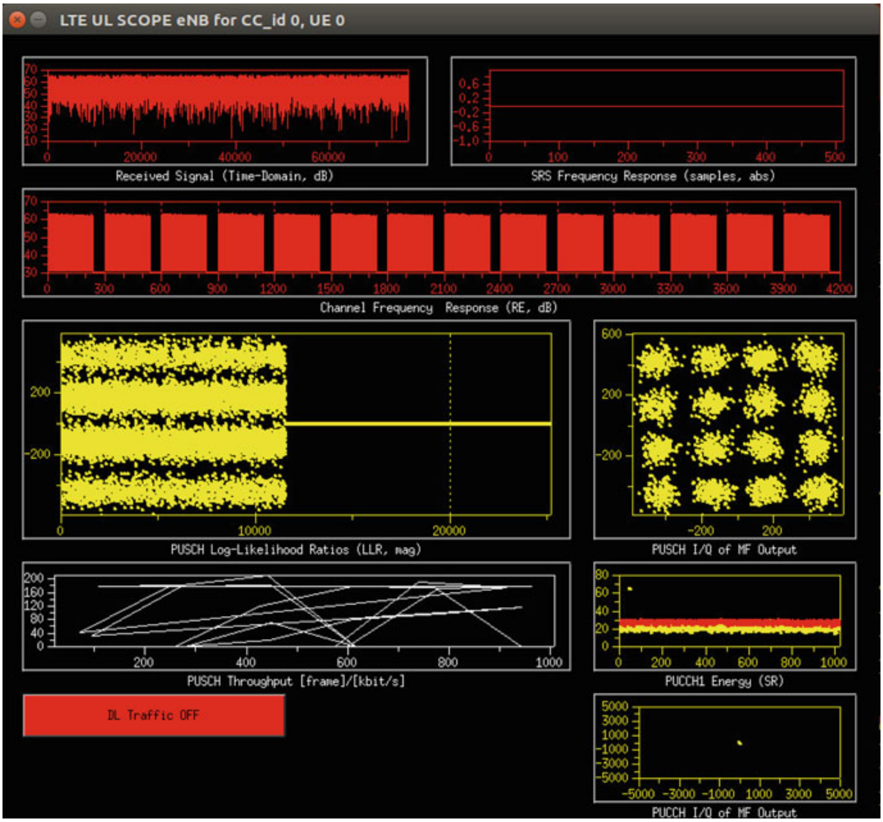

LTE UL SCOPE eNB for CC_id 0, UE 0
Received Signal (Time-Domain, dB)
SRS Frequency Response (samples, abs)
Channel Frequency Response (RE, dB)
PUSCH Log-Likelihood Ratios (LLR, mag)
PUSCH I/Q of MF Output
PUSCH Throughput [frame]/[kbit/s]
PUCCH1 Energy (SR)
DL Traffic OFF
PUCCH I/Q of MF Output

Fig. 9.5 The OpenAir user interface

resource management, and relevant schedules of the protocol process. Via the interface of physical layer, OpenAir1 and OpenAir2 become the fundamental part of a wireless communication system.

- OpenAir3: This folder carries out further development of upper layer modules based on IP networks. Moreover, it provides application interfaces as well.
- OpenAir-CN: This folder consists of various function modules for core networks, including home subscriber servers, mobility management entities, and serving gateways.

Benefited from full softwarization design, the OAI based communication platform can be flexibly partitioned according to specific requirements and this feature can be useful in handling edge computing or low latency use cases. Figure 9.5 is a snapshot of the software user interface.

9.2.2 QT

Qt is a C++ based graphical user interface (GUI) application development software. It provides all the functions needed by application developers to build an art level GUI. At the same time, it also supports non-GUI program development, such as console. Qt has excellent cross platform performance and supports cross platform application development of windows, Mac OS, Android, and embedded systems. Qt supports multiple compilation methods. One can use Qt qmake tool, Qt creator, visual studio, and other third-party compilation tools to compile. Qt creator is a cross platform IDE developed by Qt. It provides a unified development environment for many different platforms. As shown in Fig. 9.6, Qt creator integrates the GUI layout and design, qmake, Qt linguist application localization, and other functions of Qt designer. Qt has a good encapsulation mechanism and a high degree of modularity. Based on object-oriented characteristics, it enables efficient coordination among various components. In addition, Qt also provides a wealth of APIs, including a large number of C++ classes and templates and other processing functions. It supports OpenGL graphics rendering, database access, XML, JSON parsing, and parallel management.

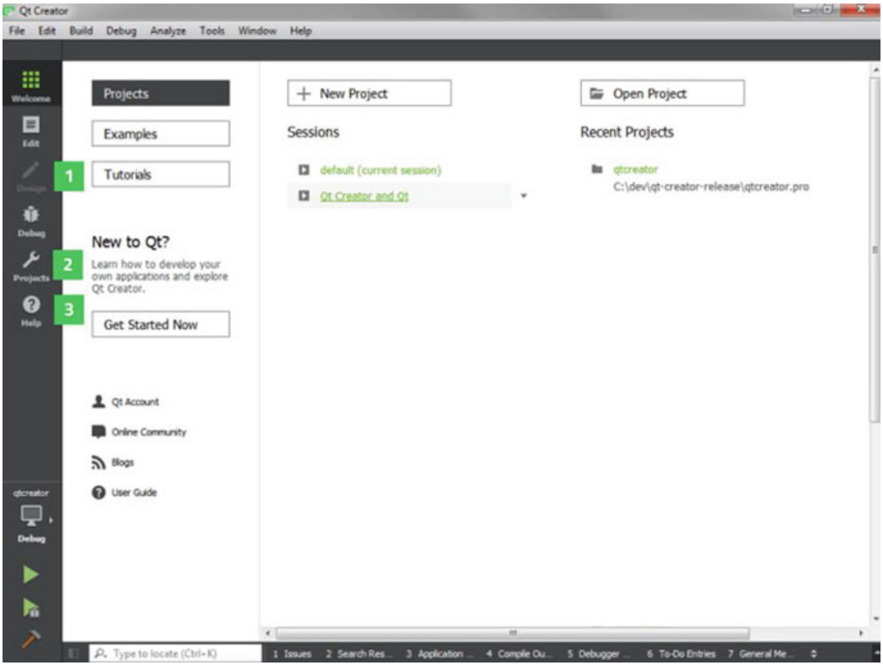
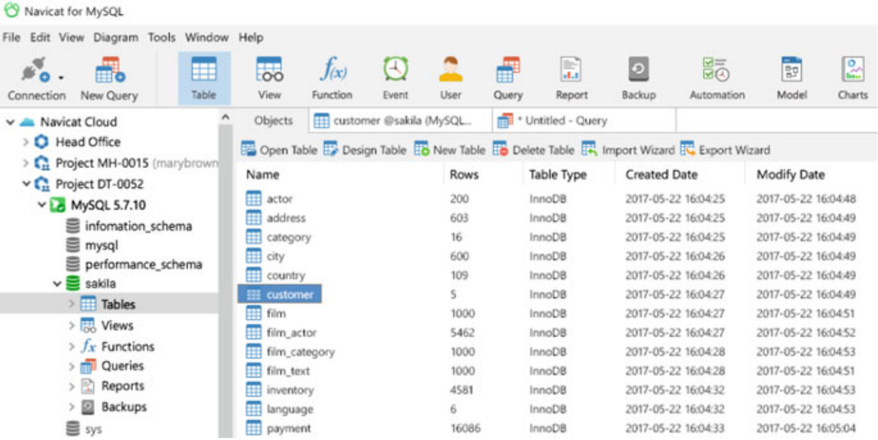

Qt Creator
File Edit Build Debug Analyze Tools Window Help
Welcome
Edit
Design
Debug
Projects
Help
1
2
3
Projects
Examples
Tutorials
New to Qt?
Learn how to develop your own applications and explore Qt Creator.
Get Started Now
Qt Account
Online Community
Blogs
User Guide
+ New Project
Sessions
default (current session)
Qt Creator and Qt
Open Project
Recent Projects
qtcreator
C:\dev\qt-creator-release\qtcreator.pro
qtcreator
Debug
Type to locate (Ctrl+K)
1 Issues 2 Search Res... 3 Application ... 4 Compile Ou... 5 Debugger ... 6 To-Do Entries 7 General Me...

Fig. 9.6 The QT user interface



Name	Rows	Table Type	Created Date	Modify Date
actor	200	InnoDB	2017-05-22 16:04:25	2017-05-22 16:04:48
address	603	InnoDB	2017-05-22 16:04:25	2017-05-22 16:04:49
category	16	InnoDB	2017-05-22 16:04:25	2017-05-22 16:04:49
city	600	InnoDB	2017-05-22 16:04:26	2017-05-22 16:04:49
country	109	InnoDB	2017-05-22 16:04:26	2017-05-22 16:04:49
customer	5	InnoDB	2017-05-22 16:04:27	2017-05-22 16:04:49
film	1000	InnoDB	2017-05-22 16:04:27	2017-05-22 16:04:51
film_actor	5462	InnoDB	2017-05-22 16:04:27	2017-05-22 16:04:52
film_category	1000	InnoDB	2017-05-22 16:04:28	2017-05-22 16:04:53
film_text	1000	InnoDB	2017-05-22 16:04:28	2017-05-22 16:04:51
inventory	4581	InnoDB	2017-05-22 16:04:32	2017-05-22 16:04:53
language	6	InnoDB	2017-05-22 16:04:32	2017-05-22 16:04:53
payment	16086	InnoDB	2017-05-22 16:04:33	2017-05-22 16:05:04

Fig. 9.7 The MySQL user interface

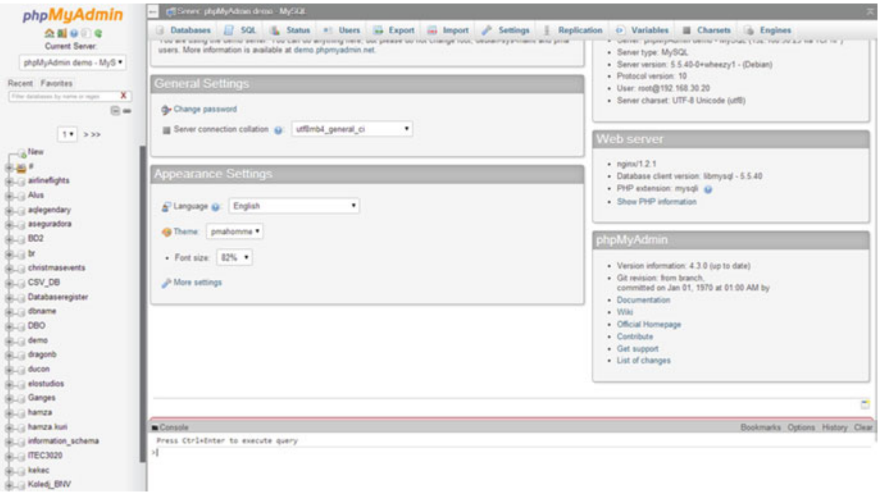
9.2.3 Database Software

MySQL

MySQL is an open-source relational database management system, which stores data in a table format to improve the speed of lookup and the flexibility of management. The standard language to access databases is SQL that supports various operation systems. Using MySQL, data creation, access, management, search, and replication can be easily realized via the APIs provided by each database. MySQL can be deployed either as a separate application or a library of another software. Currently, it is used by many database-driven web applications, including Drupal, Joomla, and WordPress. Figure 9.7 is a snapshot of the software user interface.

PhpMyAdmin

PhpMyAdmin is a free software developed based on PHP, which intends to manage and manipulate MySQL databases over the web. Frequently used operations can be performed via the user interface, while users still have the ability to directly execute any SQL statement. It could import data in CSV and SQL format and meanwhile data can be exported to different formats with the help of a set of predefined functions. The export formats support CSV, SQL, XML, PDF, Word, LATEX, and others. Moreover, it can create graphics of users’ database layout and provides the function to search globally in a database or a subset of it. Figure 9.8 is a snapshot of the software user interface.



users. More information is available at demo.phpmyadmin.net.

General Settings

Change password

Server connection collation: utf8mb4_general_ci

Appearance Settings

Language: English

Theme: pmahomme

- Font size: 82%

More settings

- Server type: MySQL
- Server version: 5.5.40-0+wheezy1 - (Debian)
- Protocol version: 10
- User: root@192.168.30.20
- Server charset: UTF-8 Unicode (utf8)

Web server

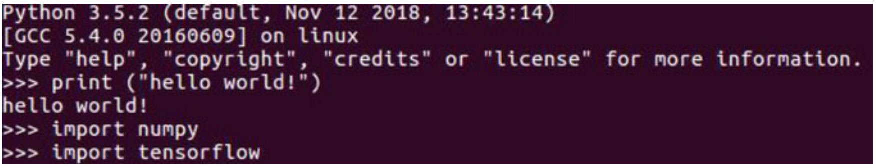
- nginx/1.2.1
- Database client version: libmysql - 5.5.40
- PHP extension: mysqli
- Show PHP information

phpMyAdmin

- Version information: 4.3.0 (up to date)
- Git revision: from branch, committed on Jan 01, 1970 at 01:00 AM by
- Documentation
- Wiki
- Official Homepage
- Contribute
- Get support
- List of changes

```
Press Ctrl+Enter to execute query
>
```

Fig. 9.8 The PhpMyAdmin user interface

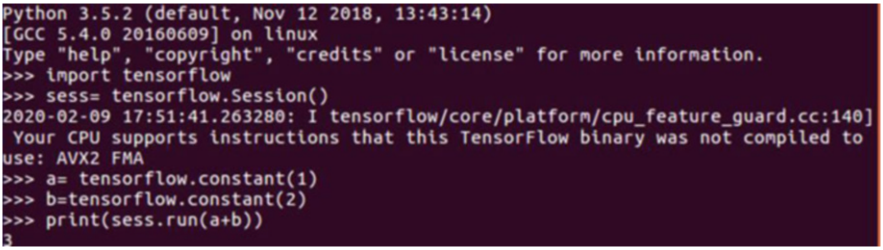


```
Python 3.5.2 (default, Nov 12 2018, 13:43:14)
[GCC 5.4.0 20160609] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> print ("hello world!")
hello world!
>>> import numpy
>>> import tensorflow
```

Fig. 9.9 The Python user interface

9.2.4 Python

Python is an object-oriented and cross-platform computer programming language. Objects include function, module, number, string, and so on. It also supports inheritance, overloading, derivation, and multi-inheritance, which enhances the reusability of source codes. Python's design philosophy is "elegant," "clear," and "simple," enabling developers to focus on solving problems rather than understanding the language itself. In addition, the author of Python designs a very restrictive syntax, so that bad programming habits will lead to compilation errors. For example, Python's indentation rules will report an error if the next line of the 'if' statement is not indented to the right. Moreover, Python features good scalability, which allows that users can easily use C, C++, and other programming languages to expand and write modules. The python compiler itself can also be integrated into other programs. Figure 9.9 is a snapshot of the user interface.



```
Python 3.5.2 (default, Nov 12 2018, 13:43:14)
[GCC 5.4.0 20160609] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import tensorflow
>>> sess= tensorflow.Session()
2020-02-09 17:51:41.263280: I tensorflow/core/platform/cpu_feature_guard.cc:140]
 Your CPU supports instructions that this TensorFlow binary was not compiled to
use: AVX2 FMA
>>> a= tensorflow.constant(1)
>>> b=tensorflow.constant(2)
>>> print(sess.run(a+b))
3
```

Fig. 9.10 The TensorFlow user interface

9.2.5 TensorFlow

TensorFlow is an open-source platform developed by Google Brain. With TensorFlow, users can build mathematical models for artificial intelligence applications through python or C++ programming. Recently, owing to the high-level Keras APIs, it is more convenient for users to get started. TensorFlow supports multiple operation systems, such as Linux, Windows, MAC, and even those embedded in mobile phones. For common machine learning algorithms, one can easily implement them by importing constantly updated algorithm libraries provided by TensorFlow. As for large-scale machine learning tasks, the Distribution Strategy API can be applied to perform distributed training on different hardware configurations without changing the model definition. The official website of TensorFlow provides detailed documents, introduction to various APIs, and basic application examples. Some basic theories of deep learning are also summarized. To install the latest version, users with CUDA-enabled GPU cards type the pip command "pip install tensorflow," and other users type "pip install tensorflow-cpu." If TensorFlow has already been installed, one can attach the "upgrade" tag to the end of commands to update it with the latest version. Figure 9.10 is a snapshot of the user interface.

9.2.6 Docker

As a popular virtualization software, Docker packages an application with its dependencies in a virtual container running in Linux, making the application deployment very flexible. In traditional virtualization solutions, each virtualized application includes not only the necessary binaries and libraries but also a complete guest operation system, which consumes a lot of storage space. As for Docker, owing to Linux kernel and the union-capable file system, a single Linux instance can hold multiple containers and virtual machines are not needed anymore. Hence, a virtualized application only contains the application itself and dependencies, which is more light-weight and portable. Detailed structures of the virtual machine solution and Docker solution are demonstrated in Figs. 9.11 and 9.12, respectively.

Fig. 9.11 The virtual machine solution

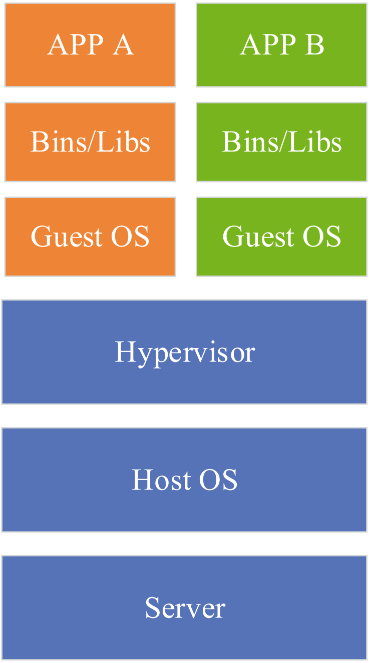
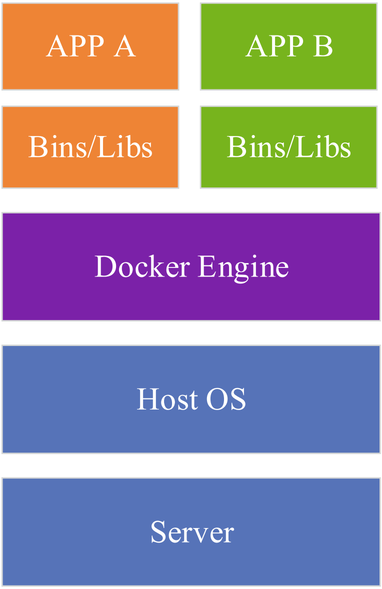


Fig. 9.12 The Docker solution



9.3 An Example of F-RAN Prototypes

In this section, an example implementation of F-RAN prototypes is elaborated, in terms of hardware configuration, client software design, featured functionalities, and performance validation.

9.3.1 Hardware Platforms

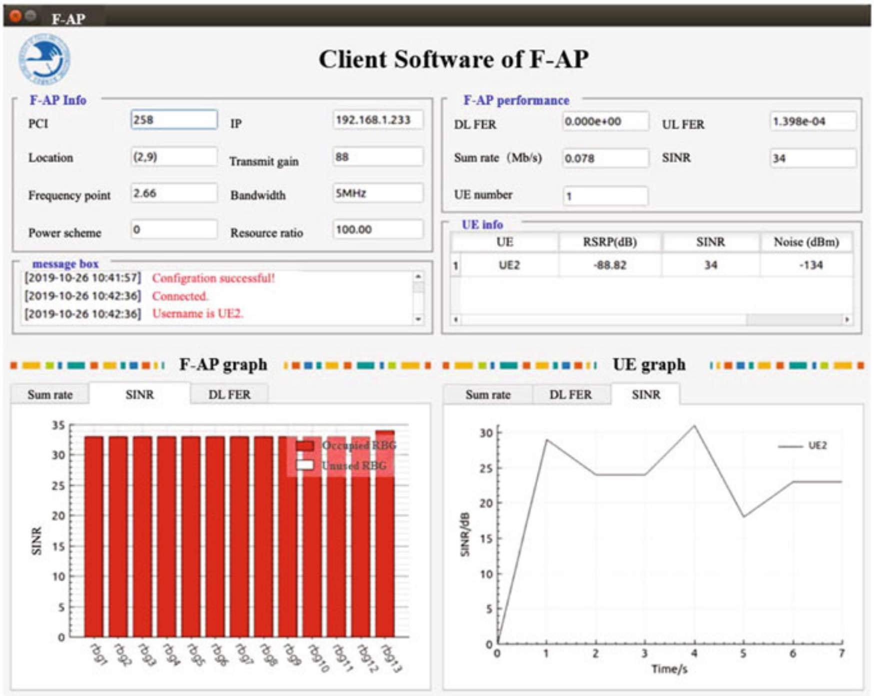
To run SDR software, there are three potential choices, namely FPGA-based SDR systems, DSP-based SDR systems, and GPP-based SDR systems. FPGA-based solutions possess significant processing power while requiring vast expenses. Although spending as much as FPGA, DSP-based solutions have poorer processing capability. GPPs means general purpose processors, also known as personal computers, which can utilize different programming languages and libraries. Compared to the former solutions, GPP-based solutions have advantages like less cost, shorter development cycles, convenient debugging, and so on. Therefore, the F-RAN prototype example uses GPPs with Intel Core i7-7700, 16 GB RAM, and Ubuntu 14.04. As for SDR software, since OpenLTE does not have UE software while SrsLTE only accomplishes the functions of eNBs in downlink, OAI is utilized here to build communication modules of UEs and F-APs. In addition, to achieve radio communication between UEs and FAPs, GPPs should be equipped with radio frontends. Produced by Ettus company, USRP B210 offers a completely integrated board to researchers, belonging to USRP (Universal Software Radio Peripheral) series. It supports wide frequency range from 70 MHz to 6 GHz, which can fully meet the experiment requirements of research community. When a UE transmits data to an F-AP, the data is first received by a USRP through the radio interface and then the radio signal is transformed into I/Q data. The connection link between a GPP and a USRP B210 can sustain massive data streams through high speed USB 3.0 interface. After installing corresponding hardware drivers, a GPP can also send commands to control USRP B210. Moreover, equipped with professional antennas supplied by Ettus, the F-RAN prototype is able to cover various frequency bands, like band 3, band 7, and so on.

9.3.2 Client Software

To intuitively demonstrate UE and FAP performance as well as facilitate flexible network configuration and network information collection, client software are developed for the UE, FAP, and network controller, which are based on Qt and MySQL.

Client Software at FAPs

The client software for each FAP includes three basic function modules, namely the performance monitoring module, the data transmission module, and the command interaction module. The performance monitoring module can acquire information related to the F-AP from OAI directly, including cell ID, IP address, antenna gain, system bandwidth, downlink sum rate, and so on. In addition, performance monitoring module can also collect UE's performance indicator values with the help of the data transmission module based on user datagram protocol that can achieve low delay. The command interaction module is responsible for delivering specific demands between client software. For example, the module of an FAP can receive the commands about adjusting transmit power and resource block allocation from that of the network controller. In order to guarantee the reliability of command delivery, the software uses transmission control protocol at this time, which is a connection-oriented, reliable, and stream-based transport layer communication protocol. The user interface of the client software at FAPs is shown in Fig. 9.13.



F-AP Info

PCI	258	IP	192.168.1.233
Location	(2,9)	Transmit gain	88
Frequency point	2.66	Bandwidth	5MHz
Power scheme	0	Resource ratio	100.00

F-AP performance

DL FER	0.000e+00	UL FER	1.398e-04
Sum rate（Mb/s）	0.078	SINR	34
UE number	1		

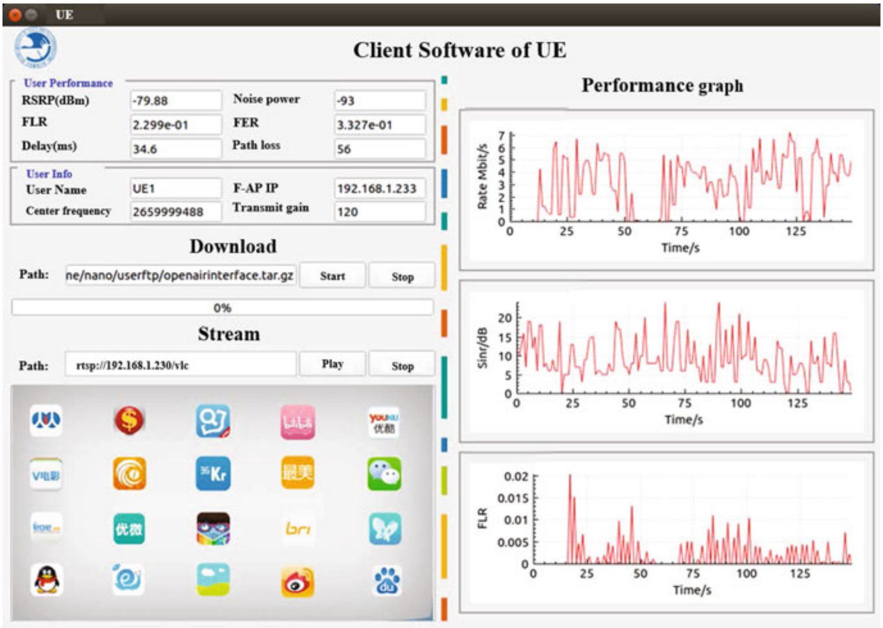
message box

```
[2019-10-26 10:41:57]  Configration successful!
[2019-10-26 10:42:36]  Connected.
[2019-10-26 10:42:36]  Username is UE2.
```

UE info

	UE	RSRP(dB)	SINR	Noise (dBm)
1	UE2	-88.82	34	-134

Fig. 9.13 The user interface of the client software at FAPs



User Performance

RSRP(dBm)	-79.88	Noise power	-93
FLR	2.299e-01	FER	3.327e-01
Delay(ms)	34.6	Path loss	56

User Info

User Name	UE1	F-AP IP	192.168.1.233
Center frequency	2659999488	Transmit gain	120

Path: ne/nano/userftp/openairinterface.tar.gz

Path: rtsp://192.168.1.230/vlc

Fig. 9.14 The user interface of the client software at UEs

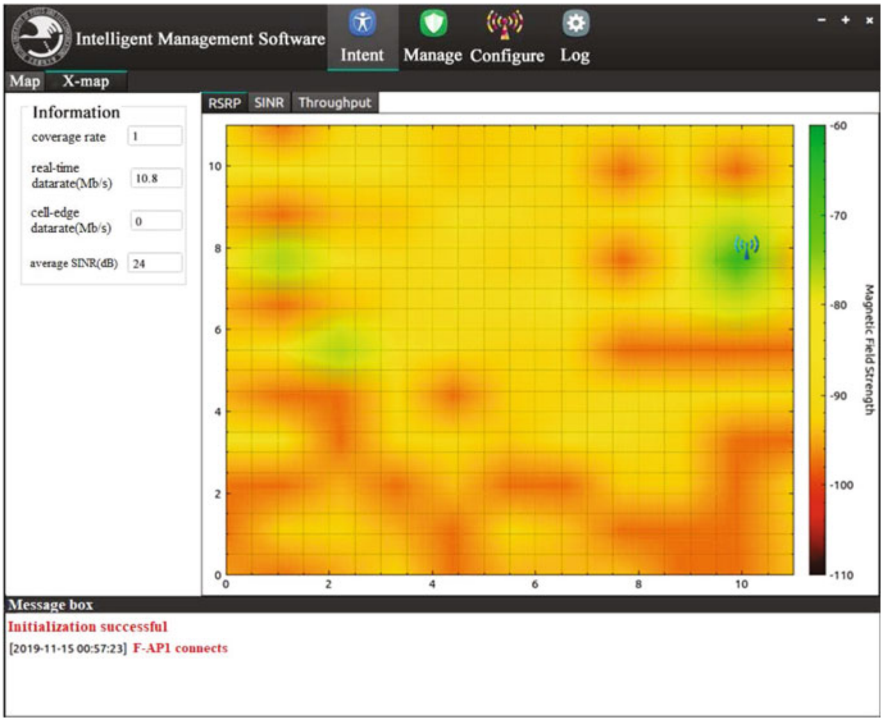
Client Software at UEs

Client software at UEs provide vivid graphs to demonstrate user performance such as receiving SINR, receiving rate, and receiving frame error ratio. In addition, it integrates VLC and FTP to realize video streaming and file downloading. In Fig. 9.14, the client user interface includes three main components, namely user information display, service workplace, and performance curves. In the left side, user information like RSRP, delay, and frame lose ratio is demonstrated, and the rest of the place is the service workplace. By clicking download button, one can start file downloading from a specified IP address. Similarly, play and stop button of VLC control video streaming service. In the right side of the user interface, UE performance curves are drawn, which vary with time. In this way, it is convenient to observe the peaks and variations of UE performance.

Client Software at the Network Controller

The client software at the network controller is used to help monitor and manage network in near real-time. By obtaining information reported from F-APs and UEs based on UDP, the network controller has a global view of the whole network. In addition, this client software integrates network fault diagnose algorithms, network performance self-optimization algorithms, and semantic recognition algorithms. Therefore, once some network failures happen, the client software can detect and do optimization as soon as possible. In addition, the software allows network operators input their service demands in the user interface using human languages, which can be then transformed to specific network configuration parameters. These parameters are sent to F-APs by the client software via TCP. Moreover, the user interface can demonstrate the current network topology and radio environment map, making the operation of the prototype more convenient.

Figure 9.15 shows the user interface, where it is divided into three functional areas, namely the control area, the information display area, and the message box from top to bottom. In the control area, there are four buttons. The “Intent” button gives a bridge to obtain operator’s various requirements and transform them into specific network configurations. The “Management” button is designed to select various advanced algorithms helping network operation and management. The “Configuration” button is responsible for distributing configuration parameters. For example, via this button, one can change the antenna gain of each F-AP. The “Log” button enables to review the past operations we have performed. The information display area shows the global network topology and radio environment condition.



Information

coverage rate 1

real-time datarate(Mb/s) 10.8

cell-edge datarate(Mb/s) 0

average SINR(dB) 24

Initialization successful

[2019-11-15 00:57:23] F-AP1 connects

Fig. 9.15 The user interface of the client software at the network controller

In the bottom of the user interface, message box shows the working process of the client software.

9.3.3 Featured Functionalities

Based on the network information fed back by UEs and F-APs, the network controller of the F-RAN prototype realizes several advanced functionalities, including network-wise radio environment map (REM) construction, cell outage detection, and performance self-optimization.

REM Construction

Radio environment maps (REMs) (Zhao et al. 2006) are promising for diverse techniques in wireless networks, such as cognitive radio and self-organizing networks (SONs). Particularly, SONs regard them as crucial bases to automatically adjust network and radio frequency parameters (Peng et al. 2013). According to the measurements reported by UEs, REMs can be constructed based on different information, such as radio link failure rates and Reference Signal Received Power (RSRP). In literatures, there have been considerable researches building REMs by measured data. The general methods are mainly based on interpolation techniques, such as inverse distance weighted interpolation (Renka 1988) and Kriging interpolation based on spatial correlation (Konak 2010). Further, improved Bayesian Kriging and fixed rank Kriging are applied in Sayrac et al. (2012), Braham et al. (2017). The former takes into account various uncertainties in the path-loss models, while the latter reduces the computational complexity of prediction. However, these methods require a large number of UEs reporting their geo-located measurements, which may not be available in practice.

To improve the REM construction accuracy when the number of RSRP measurements is limited, the F-RAN prototype implements an algorithm containing two steps in the network controller. In the first step, historical RSRP measurements are fully utilized. According to the historical RSRP values, the whole area, which the prototype serves, is divided into multiple grids that is further grouped into multiple clusters via K-means clustering based on the large-scale fading model only with distance related parameters. After clustering, the grids in a cluster have the same propagation characteristics. In the second step, current measurement data is reported by UEs located in each cluster and is utilized to estimate and update the parameters of the shadow fading part of the fading model by the Expectation Maximization algorithm. Based on the above two steps, the RSRP values of the unmeasured locations can be determined by calculating the optimal estimation. Figure 9.16 shows an example of the constructed RSRP map.

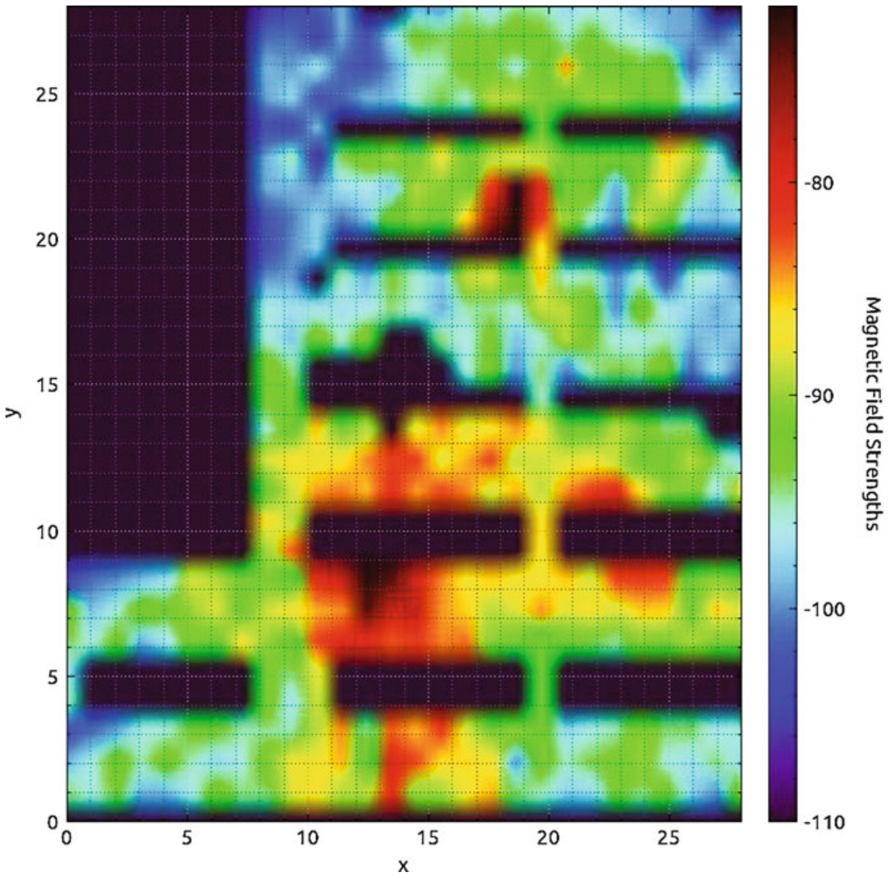


Fig. 9.16 An example of the constructed RSRP map

Cell Fault Diagnosis

Traditional supervised learning based fault diagnosis algorithms need manually labeling historical network data with corresponding operation state, which increases considerable cost. In the network controller of the prototype, by referring to Gómez-Andrades et al. (2016), an unsupervised learning approach with self-organizing map (SOM) and hierarchical clustering is adopted, which only needs to label much less number of data. As a type of unsupervised neural networks, SOM can extract features from unlabeled datasets. The main function of SOM is to convert input network state vectors with arbitrary dimensions into low-dimensional maps without discretizing the original data thus avoiding information loss. Therefore, SOM is adopted to classify the cell states based on network KPIs uploaded by UEs. The entire algorithmic process begins with building a network of neurons. Typically, the elements of SOM are organized in a 2-D grid. The dimension of each neuron's weight vector is determined by the number of collected network KPIs in the training set. The number of neurons and topological structure are generally determined in the following ways:

$$M = 5\sqrt{N}, \tag{9.1}$$

$$\frac{\mathrm{n}_1}{n_2} = \sqrt{\frac{e_1}{e_2}}, \tag{9.2}$$

where N is the number of training samples, M is the number of neurons, e_1 and e_2 are the two largest eigenvalues of the input spatial sample matrix, n_1 and n_2 are the number of rows and columns of neurons, respectively.

After determining the topology and the number of neurons, the weight vector of each neuron needs to be initialized. There are two methods to initialize SOM, namely randomized initialization and linear initialization. Here, linear initialization is selected, which can contribute to more rapid convergence of the diagnosis algorithm. The training process is divided into two stages, rough-tuning and fine-tuning. The main purpose of the rough-tuning phase is to achieve a gradual ordering of neurons and the fine-tuning phase allows the weight vector of each neuron to be fine-tuned to their optimal values.

The next step is to cluster the neurons with similar characteristics. The purpose is to group the neurons into appropriate clusters, and each cluster represents a specific cell state, such as normal state, excessive inter-cell interference, and weak coverage. Then, probability density function (PDF) of each KPI in each cluster is utilized to analyze the relationship between the cluster and network states, i.e., identifying appropriate labels for each cluster. Furthermore, this statistic analysis facilitates to verify that the cluster is correctly partitioned. Note that one of the clusters must belong to the normal state. Figure 9.17 shows an example of the neuron clustering result, where three different network states are considered. Once the diagnose model is well trained, at the fault diagnosis stage, according to the Euclidean distance, the matched neuron of the current network KPI vector is found at first, and the current cell state can be determined by the cluster to which this neuron belongs. To improve diagnostic accuracy, a border detection algorithm is utilized to determine the state of an input vector when its matched neuron locates at a cluster boundary. The algorithm calculates the Euclidean distance between the input vector and all neurons of each border cluster and re-determines the cluster to which it belongs.

9.4 Performance Evaluation of F-RAN Prototypes

In this chapter, an F-RAN experimental prototype is implemented, which is shown in Fig. 9.18. The local DC is a small data center that is for edge caching and

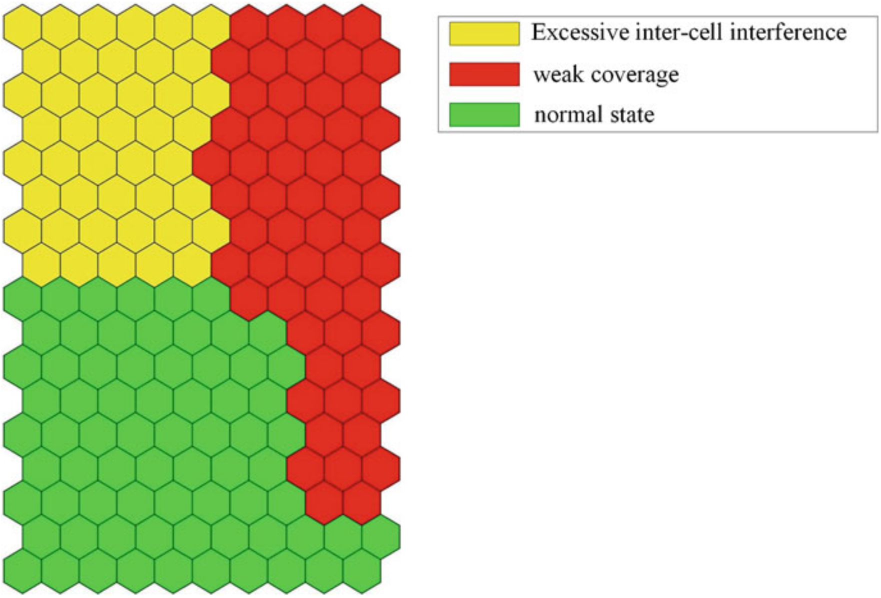


Fig. 9.17 An example of neuron clustering

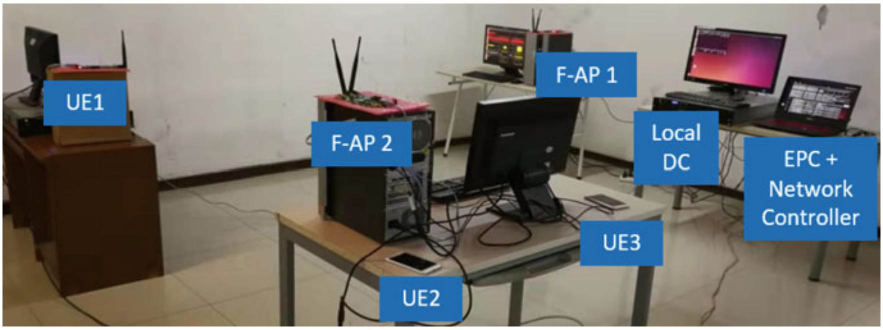


Fig. 9.18 The developed real F-RAN prototype

computing, and the network controller is co-located in the same GPP server with the EPC. To show the superiority of edge caching in F-RANs, a high-definition video (HDV) service is deployed in the prototype. Specifically, five different video transmission modes are evaluated, which are shown in Table 9.1.

The HDV content request of a UE will be transmitted to the F-AP antenna via wireless channels first, and then the radio peripheral device forwards the data in I/Q form to F-AP's protocol stack functions. Then the complete protocol stack functionalities from PHY layer to PDCP layer will be further processed. When added the GTP-U header, the request data packets will be sent to the EPC, and the controller then acquires user data from UDP/IP flows. In the controller, we have

Table 9.1 HDV transmission modes

		Locations caching HDV files		
Trans. mode	Sch. Num.	F-AP	Local DC	Remote DC
F-AP mode	1	√		
	2		√	
	3			√
C-RAN mode	4		√	
	5			√

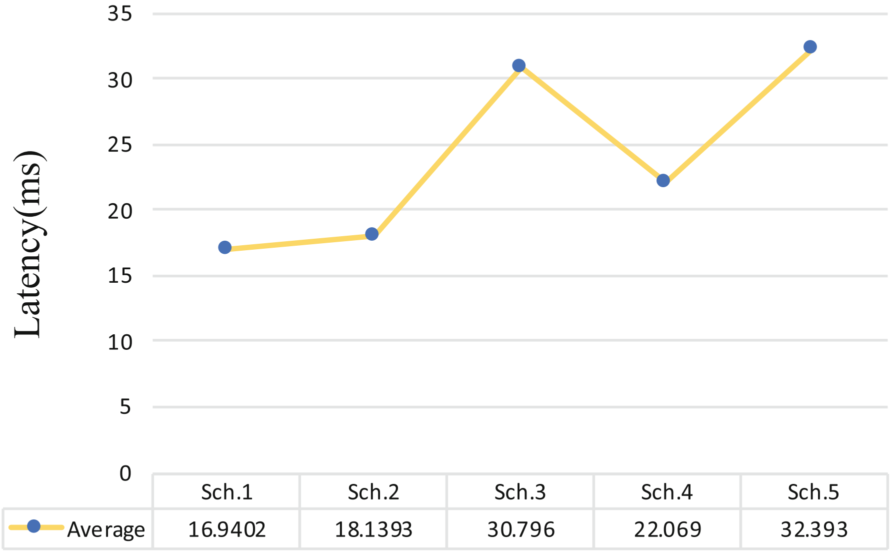


Fig. 9.19 The experimental end-to-end latency

implemented a traffic-aware function module that inspects and parses the Uniform Resource Location (URL) contained in the Hypertext Transfer Protocol request data packets. Next, the request redirection module checks whether the URL is stored in the redirection list maintained by a database. If yes, the module performs network address translation and redirects the request data packets to the local DC or F-APs according to the cache states.

The performance evaluation indexes include end-to-end latency, frame loss rate, and video quality. Particularly, the end-to-end latency is defined as the time duration starting from when the UE's content request packet arrives at the controller until the HDV packets arrives at UEs. The experimental results are illustrated in Fig. 9.19. It can be seen that pre-caching the HDV file at the F-AP or local DC can lower the end-to-end latency considerably. Moreover, due to the usage of general gigabit twisted pair cable, which is used for the communication between the control unit and the data unit, some extra delay is incurred in C-RAN mode (around 16.65% performance loss) when compared to the typical F-AP mode.

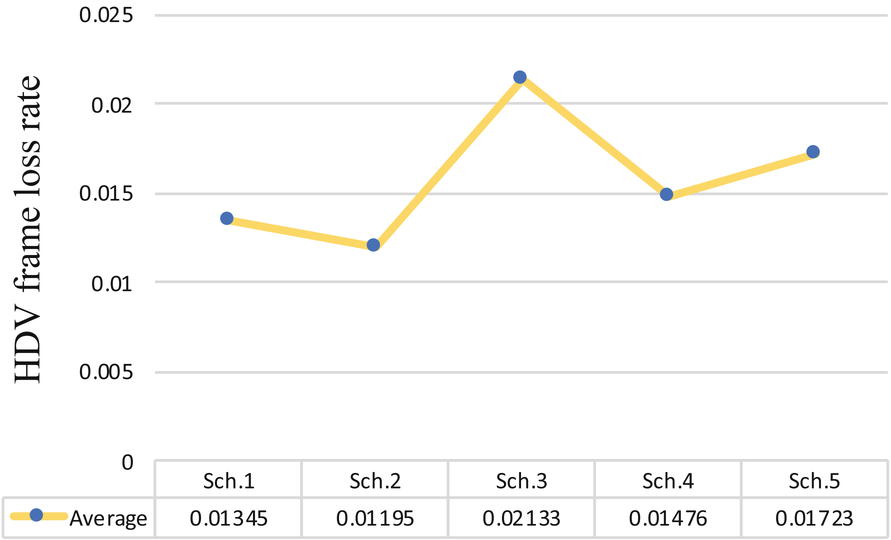


Fig. 9.20 The experimental frame loss rate

In Fig. 9.20, real-time HDV frame loss rate results are presented under various transmission schemes, which are obtained by the VLC media player. Compared to caching in the Cloud level, caching the HDV file in the Fog level can significantly lower the frame loss rate by 30.57% on average. In addition, for Sch. 2, computing tasks like video coding are offloaded from the F-AP by caching HDV file at the local DC, and hence Sch. 2 possesses the best performance. Compared with Sch. 2, Sch. 4 suffers from transmitting high speed baseband data between the control unit and data unit, which impairs the frame loss rate.

9.5 Summary

In this chapter, the prototype and test bed design of fog radio access networks (F-RANs) have been discussed. Specifically, the details about the implementation of fog computing and the network controller have been introduced. In addition, useful development tools have been investigated, based on which an example F-RAN prototype has been built. By experimental evaluation, it has been shown that the quality of experience can be significantly enhanced for a high-definition video service with F-RANs.

References

Braham H et al (2017) Fixed rank kriging for cellular coverage analysis. IEEE Trans Veh Technol 66(5):4212–4222

Gómez-Andrades A et al (2016) Automatic root cause analysis for LTE networks based on unsupervised techniques. IEEE Trans Veh Technol 65(4):2369–2386

Huang S et al (2017) Application-aware traffic redirection: a mobile edge computing implementation toward future 5G networks. In: 2017 IEEE 7th international symposium on cloud and service computing (SC2). IEEE, Piscataway, pp 17–23

Konak A (2010) Estimating path loss in wireless local area networks using ordinary kriging. In: Proceedings of the 2010 winter simulation conference. IEEE, Piscataway, pp 2888–2896

Li C et al (2018) Mobile edge computing platform deployment in 4G LTE networks: a middlebox approach. In: USENIX workshop on hot topics in edge computing (HotEdge 18), USENIX Association

Peng M et al (2013) Self-configuration and self-optimization in LTE-advanced heterogeneous networks. IEEE Commun Mag 51(5):36–45

Peng M et al (2016) Fog-computing-based radio access networks: issues and challenges. IEEE Netw 30(4):46–53

Renka RJ (1988) Multivariate interpolation of large sets of scattered data. ACM Trans Math Softw 14(2):139–148

Sayrac B et al (2012) Improving coverage estimation for cellular networks with spatial Bayesian prediction based on measurements. In: Proceedings of the 2012 ACM SIGCOMM workshop on cellular networks: operations, challenges, and future design. ACM, New York, pp 43–48

Zhao Y et al (2006) Radio environment map enabled situation-aware cognitive radio learning algorithms. In: Proceedings of software defined radio (SDR) technical conference, pp 13–17

Chapter 10
Future Trends and Open Issues in Fog Radio Access Networks

As a promising network architecture to satisfy the whole package of performance requirements of 5G, the paradigm of F-RANs has drawn a lot of attentions from both academia and industry. As introduced in Chap. 2, a fog computing layer, which comprise enhanced network edge equipment, is formulated in F-RANs. It provides great convenience to handle some user requirements locally. The advanced signal processing techniques, such as NOMA and other cooperative techniques introduced in Chap. 4, can be employed in the physical layer of F-RANs. Moreover, the edge caching and computing techniques, which have been discussed in Chaps. 7 and 8, respectively, can fully explore the potential of edge processing capability of F-RANs. Therefore, as introduced in Chaps. 5 and 6, it is flexible to achieve the KPIs of various 5G scenarios by taking full advantages of distributed caching and centralized processing. The performance gains of both theoretical analysis and prototyping tests are provided by Chaps. 3 and 9, respectively.

From the Network Architecture Perspective The existing network architecture and management strategies of F-RANs still orient to the conventional wireless service, whose key objective is to improve the performance of spectrum efficiency, energy efficiency, and transmission latency. To satisfy the highly dynamic QoS requirements of different application scenarios, i.e., ultra high-speed broad bandwidth, massive connections, and ultra reliability and low latency, the adaptive management strategies should be implemented in F-RANs. However, it cannot be supported in the existing paradigm of F-RANs due to the following reasons: First, to mitigate the loadings and latency caused by backhaul transmissions, the F-APs are deployed in F-RANs to enhance the computing and communication capability at the edge of networks. However, the distributed processing and transmissions only can be implemented in the data plan. Meanwhile, the centralized management mechanisms are employed in the control plan. Unlike the distributed paradigms of network management, the centralized management schemes require a huge amount of signaling exchange between the cloud computing centers and the F-APs, which cannot support dynamic network management strategies due to the low efficiency.

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0_10

Moreover, all the signaling messages are aggregated at the cloud computing center, and these high-dimension network parameters are quite different in the time scale. Due to the high computational complexity and slow convergence, the optimization strategies are difficult to approach, and the adaptive strategies cannot be made in real-time. Therefore, it becomes the bottleneck of exploring the potential of F-RANs. Second, the existing distributed network management schemes cannot satisfy the diverse requirements of QoS. Although the AI techniques have been employed to implement distributed network management, they cannot keep a balance between the performance and efficiency. The first category of distributed network management strategies is based on the local collected information, which is based on the uncooperative paradigm. It will cause conflicts among the F-APs, and cannot achieve global optimal results. Therefore, it cannot satisfy the requirements of network management in F-RANs. Another category is to implement model-level distributed paradigms in F-RANs, where the decisions of network management can be made locally. However, the employed models are generated in centralized approach. When the status of F-RANs changes, they need long time to sense and make adjustments accordingly, and thus cannot adapt with the high dynamics.

From the Data Support Perspective Due to the wide extension of the application scenarios, the future 6G systems should bear different services simultaneously. Since the bearer services always change, the networking schemes should be adjusted adaptively. However, the existing configuration of wireless networks is so complicated that cannot support all these applications in a cost-efficient way. To implement flexible orchestration and automatic configuration, data-driven networking paradigm should be employed in F-RANs. Due to the dense deployment of F-APs, a huge volume of data can be collected in F-RANs. Moreover, unlike the conventional cellular networks, the communication and computation capabilities are tightly coupled in F-RANs, and thus each data element is with extremely high-dimension. Even though enough data can be obtained by F-RANs to implement data-driven networking paradigm, it still has some critical issues: First, the collected data in F-RANs is not in a unified format. In F-RANs, various categories of data can be sensed and collected. In particular, heterogenous access nodes can act as F-APs in F-RANs. Due to the implementations of different protocols, it is challenging to fuse all the collected data of F-APs. Moreover, different signaling messages are exchanged in the radio access networks, transport networks, and core networks. Since these signaling messages are not in the same format, it cannot formulate a unified data-driven networking framework in F-RANs. Various application scenarios, i.e., eMBB, URLLC, and mMTC, should be supported in F-RANs. In the future 6G systems, it attempts to automatically instantiate and terminate the network slices to bear all these services in efficient approaches, which requires to implement intelligent network orchestration. However, the QoS requirements and user behavior patterns are widely different, which are difficult to fuse and analyze to generate learning-based models and strategies. Second, the quality of collected data cannot guaranteed: In F-RANs, the network data is mainly generated by the network edge devices, such as the F-APs and the user equipment. Due to the limited storage and processing capability, the quality of generated data cannot be guaranteed, i.e., the values of some data elements might be abnormal, and even absent. In addition, the transmissions of data are via the wireless channels. The detection error exists due to the reliable transmission circumstances, which will affect the quality of data during the aggregation and distribution procedures. All these non-ideal features of network data will cause complicated error propagations during the intelligent networking model generation process, which block the intelligent models to approach the optimal networking strategies. Finally, the data volume is too large to manage. Due to explosive development of IoT devices and applications, which are the mainstream services provided by F-RANs, a huge amount of data are generated all the time. However, due to the low cost and limited volume, the network edge equipment, i.e., IoT devices, cannot store all the generated network data. Although some advanced techniques, such as distributed storage, can be employed to manage the data of F-RANs, they will cause high cost and power consumption for data distribution. Therefore, it is challenging to manage the huge amount of network data in a cost-efficient way.

From the Learning Paradigm Perspective To satisfy the diverse QoS requirements of various application scenarios in the future 6G, the resource management of F-RANs should approach the optimum results, since the radio resource is scarce. However, the existing resource optimization schemes are quasi-static due to the following reasons: First, most of the existing resource optimization schemes are based on the ideal assumption that the network parameters are unchanged, and thus they cannot adapt the changes of network status. Second, all these schemes are based on the theories of convex and numerical optimizations. Since a lot of iterations need to be taken to converge to stationary results, these schemes are time consuming, and cannot implement resource allocations efficiently. To provide a feasible solution to overcome all these critical issues, the concept of network edge intelligence has been proposed, which can support real-time intelligent network and resource management at the edge of F-RANs. However, due to the restricts of learning techniques and dispersive computation capability, the deployment of network edge intelligence in F-RANs still faces some technical issues, which are introduced as follows: First, the conventional centralized learning methods cannot guarantee the efficiency and privacy requirements of network edge intelligence: In the existing centralized learning-based paradigms, the generated data at the edge of F-RANs has to be aggregated by the cloud computing servers. To formulate a comprehensive framework for network and resource management, a high-dimension network parameter vector, which are employed to capture the operation status of F-RANs, should be established. Moreover, a huge amount of data should be collected to guarantee the accuracy performance of learning-based methods. Therefore, the costs of centralized learning-based approaches, especially caused by data exchanging, are extremely high. Moreover, the centralized learning paradigms require the network edge devices to upload their raw data via the wireless channels, which is privacy-sensitive. Due to the existence of eavesdroppers and interfering nodes, there exist severe security issues in the centralized learning methods. Second, it is challenging to generate high quality learning models by employing dispersive computation resource in F-RANs: The model training procedure is computation intensive, which has high demands of computation capability. Although there exists a lot of fog computing nodes at the edge of F-RANs, such as the F-APs and users. These nodes locate distributively, and it is difficult to integrate them in a cost-efficient way. In particular, the computation tasks of model generation cannot be split into simple tasks that can be executed by each node independently. Therefore, novel collaboration mechanisms should be designed to handle the model generation procedures efficiently.

As a data-level distributed learning paradigm, federated learning can provide a potential approach to implement network edge intelligence in F-RANs, which can solve all these critical issues. Therefore, the federated learning-enabled paradigm of F-RANs shows the future evolution path of 6G network architecture, which is discussed in details in the following sections of this chapter.

The rest of this chapter is organized as follows: In Sect. 10.1, we proposed federated learning-enable F-RANs, which is considered as a potential evolution path of F-RANs. The fundamentals and key enabling techniques of future F-RANs are studied in Sects. 10.2 and 10.3, respectively. The open issues are discussed in Sect. 10.4, and this chapter is concluded in Sect. 10.5.

10.1 Future Trends of F-RANs: Federated Learning-Based Paradigms

Consider the F-RAN scenario shown in Fig.10.1. As introduced in Chap. 2, a fog computing layer can be formulated at the edge of F-RANs, which consists of the F-APs and users, while the cloud computing center can act as the cloud computing layer. In particular, the functions of local processing and network management are supported in the fog computing layer in a distributed way, and the global centralized processing and management can be implemented in the cloud computing layer. The

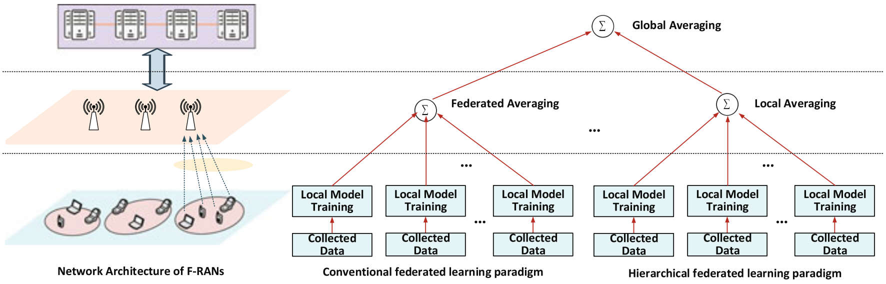

Global Averaging
Federated Averaging
Local Averaging
Local Model Training
Collected Data
...
Network Architecture of F-RANs
Conventional federated learning paradigm
Hierarchical federated learning paradigm

Fig. 10.1 Illustrations of federated learning paradigms in F-RANs

connections between the F-APs and cloud computing centers are established via the backhaul links.

As we discussed previously, the AI-based frameworks should be deployed to implement signal processing and network management in F-RANs, which can support various services with highly dynamic diverse requirements. The performance of AI-based frameworks is determined by the employed learning techniques and collected network data. Since the conventional centralized learning methods cannot keep a balance between the performance and efficiency, a model-level collaborative learning paradigm, named federated learning, is employed to implement network edge intelligence in F-RANs.

Federated learning can generate high quality learning models without aggregating the raw data to the computing servers. The key idea of federated learning is to encourage each client to generate a local model based on its own data. The update results of all these local models should be sent back to the server, which can be employed to generate a global model.

Algorithm 1 Federated learning procedure

```
Initialization: Global learning model $\mathbf{v}_F$ and local learning models $\mathbf{v}_1, \ldots, \mathbf{v}_M$.
For each round of model update:
  Local model update:

    The users update the local models as $\mathbf{v}_m \rightarrow \mathbf{v}_m + F(\mathbf{v}_m, S_m)$.
  Update results feedback:
    $\mathbf{v}_m$ is sent to the associated F-AP through the wireless channels.
  Federated averaging:

    The update result of global model can be expressed as $\mathbf{v}_F \rightarrow \sum_{m=1}^{M} \frac{N_m}{N} \mathbf{v}_m$ .
    $\mathbf{v}_F$ is transmitted to the users, and their local models can be updated accordingly.
```

10.1.1 The Conventional Federated Learning Paradigm

In F-RANs, the conventional federated learning paradigm can be deployed among between single F-APs and multiple servers. In particular, the fog computing processor equipped with F-APs can act as the server, while the local processors of users can be treated as clients. As introduced in Algorithm 1, a federated learning model can be generated based on the interactions between the F-AP and users via the wireless channels. During each round of interaction, the local and global models can be updated as follows.

- **Local model update:** The network data can be sensed and collected by the users, which can be formulated as a local dataset, and implemented to generate local learning model. In particular, the local models are updated independently by the users. As shown in Goodfellow et al. (2016), the gradient descent-based methods can be employed to optimize the update results, whose objective is to minimize the empirical risk of loss function.

- **Update results feedback:** When the procedure of local model update has been finished, the update results need to be sent to the associated F-AP through the wireless channels. To fit with the wireless transmission circumstances, the model update results should be transformed into radio signals via modulation, coding, and compression.
- **Federated averaging:** When the update results of all the users can be received by the F-AP, the global model can be updated by its computing processor. In Algorithm 1, the result of federated averaging can be given as a linear combination of local models, where the averaging weights are the proportions of employed training data. Then, the updated global model is sent back to all the users, and their local models should be updated accordingly.

10.1.2 A Hierarchical Federated Learning Paradigm

Due to the hierarchical cloud-fog computing-based architecture of F-RANs, the conventional single-layer federated learning paradigms cannot implement adaptive processing and network management in F-RANs, the reasons can be explained as follows: First, the mobile devices are usually in the idle status, and thus they spend a long time to generate enough data to training high quality learning models. Moreover, to support global centralized processing and network management, a comprehensive learning model should be formulated to capture the full view of networks, which needs to aggregate the information of all the F-APs.

To provide a feasible solution, the conventional federated learning paradigm should be enhanced into a hierarchical structure, where different levels of aggregations can be implemented at both the F-APs and cloud clouding centers. As illustrated in Fig. 10.1, two tiers of federated averaging should be deployed in the hierarchical federated learning paradigm. The local averaging is firstly implemented at the F-APs, which is identical with the conventional federated learning paradigm. In the cloud computing centers, a higher tier of model aggregation, which is named as global averaging, should be implemented. In particular, each F-AP should upload the update result of its local averaged model to the cloud computing center, and then the global averaged model can be updated accordingly. The F-APs can be treated as the medium of formulating such hierarchical framework, which act as the aggregators during the local averaging procedure, while can be treated as the clients in the global averaging phase.

10.1.3 Potential Applications of Federated Learning in F-RANs

As the foundation of implementing network edge intelligence in F-RANs, federated learning should be fully integrated in all layers, and its potential applications can be introduced as follows.

- **In physical layer:** Some signal processing techniques are with high computational complexity, such as channel decoding and transceiver design for MIMO, which cannot satisfy the QoS requirements of URLLC and mMTC scenarios. To approach the optimal performance more efficiently, these techniques can be rethought from federated learning perspectives. Moreover, to cope with the complicated interference circumstance, federated learning can be implemented to approach, or even achieve better performance than the state-of-art methods, as long as suitable learning models can be used, and the high quality training data can be provided. It can also be deployed for the enhancement of spectrum sensing and physical layer security.
- **In network layer:** The bearer services in F-RANs are various, which are quite different in the time scale. Therefore, it is challenging to manage all the traffics in a centralized approach, which requires to aggregate all the related data to the cloud computing center in F-RANs. An alternative scheme is to implement distributed scheduling strategy based on federated learning, which can support adaptive management with respect to the dynamic traffics in F-RANs.
- **In application layer:** In the future 6G, AI-based applications will become the mainstream, which can be combined with all the typical scenarios of F-RANs. As a suitable implementation strategy of AI, federated learning can be applicable as well. In particular, it can be used for recommending contents to the users based on their interests, which can mitigate the loadings of backhaul in the eMBB scenarios. Compared with the conventional centralized learning paradigms, federated learning can reduce the costs of data offloading, and protect the user privacy. In the URLLC scenarios, real-time surveillance can be implemented by using federated learning, which reduce the end-to-end latency with ultra high reliability for the intelligent transportation applications. In the mMTC scenarios, the conventional IoT can be enhanced as artificial IoT based on federated learning, which can provide great convenience for many novel edge intelligent with low costs.

10.2 Fundamentals of Federated Learning in F-RANs

As a learning-based method, federated learning can approach the optimal results without taking a lot of iterations, which can support real-time processing and management to adapt dynamics of F-RANs. As a distributively cooperative learning paradigm, federated learning faces the following fundamental problems for its implementing in F-RANs:

- **The distributed learning paradigm causes accuracy loss:** The existing high quality learning algorithms, such as deep neural networks, employ complicated nonlinear structure to enhance the capability of feature representation. In federated learning, only linear aggregation is utilized to generate the global averaged model, which will cause accuracy loss. To the best of our knowledge, it is challenging to design feasible schemes to analyze such performance loss, and cannot provide applicable methods to mitigate its impact (Zhao et al.2018).
- **It is difficult to keep a balance between the performance and efficiency of federated learning in F-RANs:** Since the transmission circumstance of F-RANs is not ideal, the communication error is added with the feedback update results, which will cause nonlinear error propagation during the learning procedure. Although the reliability of model exchanging can be guaranteed by employing advanced transmission techniques, it will cause extra communication cost, such as long transmission latency. Therefore, we need to keep a sophisticated tradeoff between the performance and cost of model training procedure.

To improve the accuracy performance of federated learning in a cost-efficient way, the techniques of loss compensation and model compression are necessary, and their implementations in F-RANs can be introduced as follows.

10.2.1 Loss Compensation

The loss function can be used as a metric to evaluate the accuracy performance of learning models, and the accuracy loss of federated learning in F-RANs can be defined as the Euclidean distance of loss functions between the federated learning model $\mathbf{v}_F$ and the theoretically optimal model $\mathbf{v}^*$ (Shamir et al. 2013). Due to the linear federated averaging, the loss function of global model of federated learning can be given as a summation of loss functions of all the local models. In this chapter, the loss function is defined as the empirical risk based on the training datasets. Therefore, $\mathbf{v}^*$ can be defined as the model that can minimize the expected risk of global distribution $P_G(\mathbf{z})$. Recalling the accuracy loss of federated learning can be modeled as follows.

- **Sample selection bias:** In federated learning, the local training dataset S_m can be treated as a sample of a random variable z with distribution $P_m(z)$. To improve the performance of local training models, the empirical risk is employed as an approximation to evaluate the expected risk to reduce the accuracy loss. However, since the data volume of S_m is limited, there always exists a gap between $F(\mathbf{v}_m, S_m)$ and $F(\mathbf{v}_m, P_m(\mathbf{z}))$, due to the central limit theorem, which is named as sample selection bias. It cannot be overlooked in the data-driven schemes, especially when the size training dataset is small.

- **Distribution divergence:** Unlike the existing other applications of AI, such as computer version, the network data of F-RAN is not IID. In particular, during the federated learning procedure, the local model of each user is trained based on its own local dataset, which follows distribution $P_m(\mathbf{z})$. Due to the diversity of user behavior and network circumstance, the distributions of collected data of all the users are quite different, which is not identical with the global distribution $P_G(\mathbf{z})$. Such distribution divergence may lead the learning result to converge to a model with unsatisfying performance.
- **Model convergence:** In Shamir et al. (2013), it shows that gradient-based methods can at least converge to a local optimal result when the centralized learning methods are employed, where the loss function changes continuously. However, in the federated learning paradigm, the global model is generated based on the update results of local models. Therefore, the value of its loss function changes in a discrete way, which impacts the convergence performance of federated learning. In the worst case, the convergence of federated learning even cannot be ensured.

To reduce the performance loss caused by linear federated averaging and distribution diverse data, the loss compensation methods should be used. For example, by using sophisticated design of sample selection and user scheduling, the distribution divergence can be reduced, and so is the sample selection bias. Moreover, a common dataset can be shared among the participators to harmonize the distribution divergence (Zhao et al. 2018). In Wang et al. (2019), the convergence performance of federated learning has been analyzed, which demonstrates that it converges when the IID training data is utilized.

10.2.2 Model Compression

Although the computation capability of edge devices has been significantly improved in the last decade, it is still challenging to deploy the federated learning in large scale, which is restricted by power supply and storage volume. The main reasons can be explained as follows: First, a lot of neurons and interaction links are consisted in deep neural networks, and thus their training procedure will occupy a large volume of memory and consume excessive energy. Second, the cost of message exchanging can be reduced by employing federated learning due to the avoidance of raw data offloading. However, the federated averaging procedure is still with high requirements of transmit power and spectrum bandwidth, which aims to ensure the transmission reliability of feedback high-dimension model parameter vectors.

To overcome the difficulties of local training and federated averaging during the learning procedure, the techniques of model compression can be employed. Motivated by compressing the model structure and parameters, the following two methods can be utilized.

- **Structure compression:** Due to the flexible connection and activation of neurons, the expression ability of deep neural networks can be significantly improved. Meanwhile, the complicated structure causes high cost of model training, especially when the scale of neural networks are large. However, there exist a large amount of redundance in the deep neural networks. For example, some neurons are named as insignificant elements since they contribute little to the improvement of model accuracy. Therefore, the scale of deep neural networks can be compressed by deleting these insignificant neurons, which will not affect its accuracy performance (Han et al. 2016). The neurons that only with a few of connections, as well as the links whose weights are small, can be removed. Then, the deep neural networks are reformed into a sparse form, whose test accuracy is almost the same as the uncompressed redundant versions.
- **Parameter compression:** The model parameters, which are generated via the training procedure, can be modeled as high-dimension vectors with continuous elements. Since these elements are with long digits, they cannot be transmitted through the wireless channels directly. To improve efficiency and guarantee transmission reliability, model parameters can be compressed via quantization (Park et al. 2014). Then, the size of updated model can be shrunk. Moreover, the transmit power consumption can be lowered, and the decoding procedure can be simplified as well. To cope with the error caused by quantization noise, the compression schemes should be designed sophisticatedly, i.e., multiple model parameters can be compressed jointly, which can approach the limits of compression, and the non-uniform quantizing techniques should be selected based on the distribution features of model parameters.

The procedure of model compression with respect to federated learning can be introduced as follows: First, the architecture of local learning models can be simplified by using structure compression. Next, the sparsified model parameters are quantized and transformed into digital versions via compression. Moreover, if all the users transmit their update results simultaneously, the F-AP can receive a mixture of all the feedback messages as long as the synchronization can be guaranteed. Then, the federated averaging model can be obtained straightforwardly via the decompressing procedure, and thus the computational complexity can be reduced significantly.

10.2.3 Performance of Loss Compensation and Model Compression

To evaluate the performance gains of loss compensation and model compression with respect to federated learning in F-RANs, the experiment results of test accuracy are provided based on the MINST dataset. The single tier federated learning scenario is considered in this part, where the deployment of federated learning is implemented between a single F-AP and 20 users. The wireless channels in F-RANs are modeled as flat Rayleigh fading channels. Moreover, the training dataset consists of 12,000 figures, which are sampled randomly from the MINST dataset. Both the IID and non-IID dataset cases are considered here: In the IID dataset case, the local training dataset of each user is selected uniformly, while it is formed by using non-uniform selection in the non-IID dataset case. Ten thousand figures, which are different from the samples in the training dataset, are selected to formulate a test dataset. At each user, the MLP model is employed for generating local learning model. Loss compensation and model compression are employed before update results feedback, with the joint optimization of sampling local training data and user selection.

In Fig. 10.2, the test accuracy is provided. In particular, the convention federated learning method without loss compensation and model compression is used as a comparable scheme. The simulation results show that the update result of global model can converge within 30 iterations. Moreover, compared with the non-optimized scheme, the test accuracy can be improved by 4.67% by employing loss compensation and model compression when the IID datasets are used. In fact, its performance approaches the centralized learning method based on the ideal assumption of transmission circumstance for update results feedback, which is the state-of-art scheme. Since the distribution divergence exists in the non-IID training data case, the test accuracy is worse than that of the IID case. But the accuracy

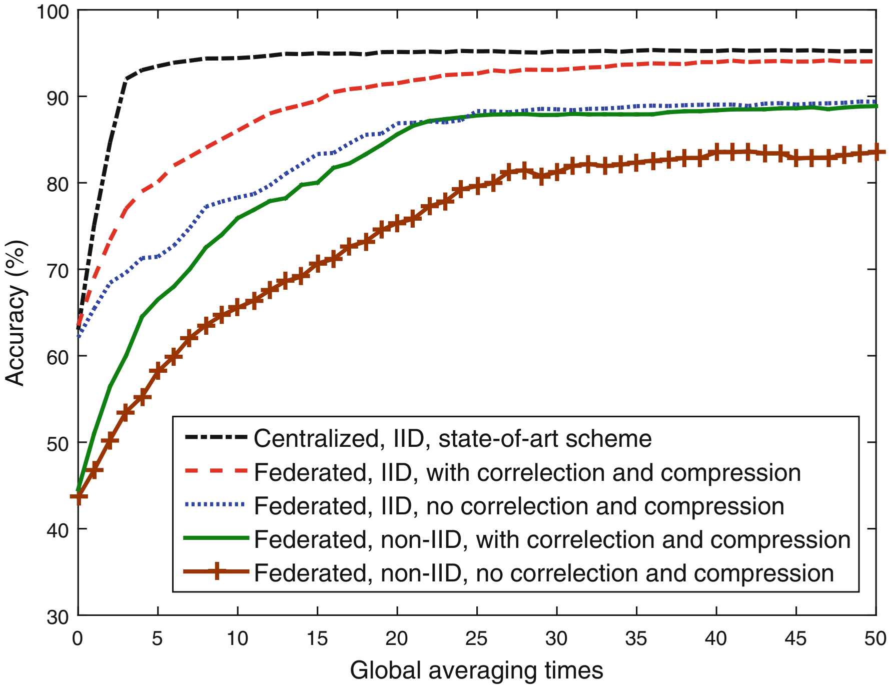


Fig. 10.2 Simulation results of accuracy loss correlation and model compression in F-RANs

performance of federated learning still can be improved by using loss compensation and model compression. In particular, the performance gain is 3.34%.

10.3 Key Enabling Techniques of Future Evolved F-RANs

As a distributively cooperative learning paradigm, federated learning can be deployed to implement flexible and dynamic processing and network management. It requires to employ some key techniques to harmonize the capability of communication, computation, and storage of F-RANs, which can guarantee the adaptivity with respect to the time and space domain. Moreover, to deploy federated learning-enabled network intelligence in F-RANs, some existing techniques also need to be enhanced.

10.3.1 Hierarchical Cloud and Fog Computing

The local model training procedure of federated learning has high requirements of the capability with respect to computation and power supply. Although some users can provide high quality data for model generation, they cannot afford the cost, such as the IoT devices, which may impact the performance and efficiency of federated learning. A feasible solution is to offload the data to other nodes with powerful computation capability, i.e., the cloud and fog computing processors, and adjacent users. As introduced in Zhao et al. (2019), computation offloading can make full use of the cloud and fog computation equipment in F-RANs, and can accumulate the training efficiency of federated learning. Compared with the conventional entire offloading strategy, partial offloading strategy is a more efficient method in federated-learning enabled F-RANs. For example, the computational complexity of model training procedure is mainly determined by model validation, which should be offloaded to nodes, and the parameter update procedure can be executed locally. Moreover, the computational tasks of federated learning can be offloaded to multiple nodes simultaneously, which can ensure the reliability of computation offloading, and improve the processing efficiency of parallel computation. Finally, since the procedure of federated learning is time consuming, the connection stability has to be guaranteed when we consider the selection of offloading nodes.

10.3.2 Advanced Transmission Techniques

To ensure the performance of federated learning, the large-scale learning models, such as deep neural networks that have multiple layers and a lot of neurons, are usually used as the local models, which cause high cost during the model feedback procedure. Therefore, to improve the efficiency of transmitting the update results through wireless channels, some advanced transmission techniques, such as massive MIMO and high-order modulation, are necessary. Moreover, to improve the transmission reliability of model parameters, which can be coded into a long digit form, LDPC code should be employed for channel coding to approach the theoretical limit, as well as the enhanced hybrid retransmission schemes should be designed. Moreover, to adapt with flexible offloading strategies in both the cloud and fog computing layers, the forwarding techniques, such as D2D and relaying transmission, also need to be implemented.

10.3.3 Resource Management and User Scheduling

In F-RANs, the resource management strategy determines the quality and efficiency of model offloading. Moreover, the user scheduling policy is critical to manage the distribution divergence and the convergence rate (Yang et al. 2020). In this part, to show the impact on the convergence of federated learning in F-RANs, the simulation results are provided in Fig. 10.3. In particular, we consider the deployment of federated learning between a single F-AP and 100 users, and 5 users need to be scheduled to participate each iteration of federated averaging. The transmit power is$\rho = 17$ dBm. It shows that the convergence performance can be improved significantly by using proportional fair scheduling, which can choose the users with the best transmission circumstance, and the communication error caused by unideal wireless transmissions can be mitigated.

10.3.4 Intelligent NFV

The key idea of NFV is to support virtualization of network functions by implementing in a software approach, and the virtualized network functions can be decoupled with the hardware. To implement network intelligence in F-RANs, the technique of NFV should be enhanced, and operation systems deployed at the cloud and fog computing layers need to be unified. Then, a common federated learning model can be trained cooperatively among heterogeneous computing nodes straightforwardly with high efficiency. Moreover, the interface between the manager of NFV and deep/machine learning models for network management should be defined, and flexible intelligent network orchestration can be implemented.

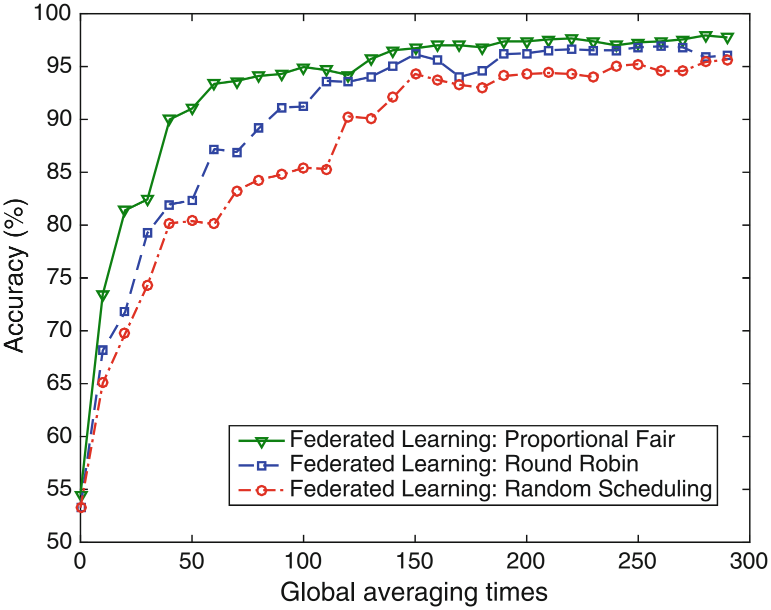


Fig. 10.3 Test accuracy of federated learning with different scheduling strategies in F-RANs

10.4 Open Issues

Although the network edge intelligence can be implemented based on a federated learning-based paradigm, it still faces some critical issues of the communication and computation perspectives, which are discussed as follows.

10.4.1 Massive Multiple Access

It is challenging to deploy federated learning among massive associated user devices, where the quality of generated models cannot be guaranteed. In the massive multiple access scenarios, the network data is collected by a huge amount of users, which cause great diverse of data distribution. It will cause the diffusion of gradient, which impacts the converge performance of federated learning. Moreover, the connection stability of users cannot be ensured when a huge amount of users participate the federated learning procedure, and the performance of federated learning cannot be guaranteed due to the dynamic frequent access and handover. Finally, due to the scarcity, the conflict of radio resource is also critical in the massive multiple access scenarios, especially during the model feedback procedure.

10.4.2 New Theory and Techniques of Deep Learning

Due to the complicated structure, the theoretical performance of deep learning models is not tractable. Therefore, we do not have any insight about the settings of the scales of participators, and the maximum rounds of federated averaging. It is difficult to achieve the best performance in the most efficient way. Moreover, in the existing learning methods, the learning models have to be trained as a whole, which cannot be decoupled. It becomes the bottleneck of fully explore the potential of dispersive computation resource of F-RANs. To improve the training efficiency of federated learning, the novel learning models, which can be separated into individual modules, should be designed. Besides the structure of learning models, new training algorithms are also necessary, since the existing gradient descent-based schemes require the loss functions to follow some specific form to ensure the convergence.

10.4.3 Security and Privacy Issue of Local Model Feedback

Although one of the most important advantages with respect to federated learning is to protect the user privacy by keeping its own data locally, there still exists the risk that its personal information can be inferred, and even reconstructed, only based on the leaked feedback models. In addition, the federated learning procedure can be attacked by spread false models, which will perturb the learning results. Therefore, it is still challenging to guarantee the security and privacy of federated learning in F-RANs. In particular, both the secured transmission techniques and reliable user authentication strategies should be designed.

10.5 Summary

In this chapter, the future trends and open issues of F-RANs have been discussed. To fully explore the potential of F-RANs, network edge intelligence should be implemented. Based on federated learning, which is a model level cooperative learning paradigms, an edge intelligent network architecture of F-RANs has been designed. Then, the key techniques with respect to the communication and computation perspectives are studied. Finally, the existing open issues of federated learning-enabled intelligent F-RANs have been discussed, which can provide some insights for the evolution of F-RANs.

References

Goodfellow I et al (2016) Deep learning. MIT Press, Cambridge

Han S et al (2016) Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding. In: Proceedings of the 33th international conference on machine learning

Park S et al (2014) Fronthaul compression for cloud radio access networks: signal processing advances inspired by network information theory. IEEE Signal Process Mag 31:69–79

Shamir O et al (2013) Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes. In: Proceedings of the 30th international conference on machine learning

Wang S et al (2019) Adaptive federated learning in resource constrained edge computing systems. IEEE J Sel Areas Commun 37:1205–1221

Yang H et al (2020) Scheduling policies for federated learning in wireless networks. IEEE Trans Commun

Zhao Y et al (2018) Federated learning with non-iid data. arXiv:1806.00582

Zhao Z et al (2019) On the design of computation offloading in fog radio access networks. IEEE Trans Veh Technol 68:7136–7149

Acronyms

1G	First generation
2G	Second generation
3G	Third generation
3GPP	Third generation partnership project
4G	Fourth generation
5G	Fifth generation
6G	Sixth generation
AI	Artificial intelligence
AP	Access point
AR	Augmented reality
ARP	Address resolution protocol
BBU	Baseband unit
BPCU	Bits per channel use
BS	Base station
C2X	Car-to-everything
CDF	Cumulative distribution function
CDMA	Code division multiple access
CDN	Content delivery network
CN	Core network
CNNs	Convolutional neural networks
CoMP	Coordinated multiple points
CPRI	Common public radio interface
C-RAN	Cloud radio access network
CRRA	Cooperative radio resource allocation
CRRM	Cooperative radio resource management
CRSP	Collaboration radio signal processing
CSI	Channel state information
D2D	Device to device
DNN	Deep neural network

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0

DQN	Deep Q network
DRL	Deep reinforcement learning
E^3	The economical energy efficiency
ECE	Edge computing environment
EE	Energy efficiency
eMBB	Enhanced mobile broadband
EPC	Evolved packet core
ETSI	European telecommunications standards institute
F-AP	Fog computing based access point
FAP	Femto access point
FDMA	Frequency division multiple access
FL	Federated learning
F-RAN	Fog computing based radio access network
F-UE	Fog UE
GEO	Geosynchronous orbit
GTP	GPRS tunneling protocol
GT	Ground terminal
HetNet	Heterogeneous network
HPN	High power node
I/O	Input and output
ICC	Interference coordination and cancelation
IIC	Industrial Internet consortium
IID	Identical and independent distribution
IoCI	The information of common interest
IoT	Internet of things
IR	Individual rational
ISG	Industry specification group
ITS	Intelligent transportation systems
ITU	International telecommunication union
KPI	Key performance indicator
LDPC	Low density parity check
LEO	Low earth orbit
LoS	Line-of-sight
LPNs	Low power nodes
LSCP	Large-scale collaborative processing
LSE	Least-squares estimate
LTE	Long-term evolution
MAC	Medium access control
MBS	Micro base station
MCC	Mobile cloud computing
MDP	Markov decision process
MEC	Mobile edge computing
MIMO	Multiple-input and multiple-output
ML	Machine learning
MLP	Multi-layer perceptron

mMTC	Massive machine-type communications
MRC	Maximal ratio combining
MSE	Mean square error
MV	Monocular video
NFV	Network function virtualization
NGMN	The next generation mobile networks
NOMA	Non-orthogonal multiple access
NR	New radio
NSA	Non-standalone
OFDMA	Orthogonal frequency division multiple access
OMA	Orthogonal multiple access
PBS	Pico base station
PDF	Probability density function
PGFL	Probability generating functional
PHY	Physical
PPP	Poisson point process
QCQP	Quadratically constrained quadratic programming
QoS	Quality of service
RANaaS	RAN-as-a-Service
RAN	Radio access network
RB	Resource block
RF	Radio frequency
RNNs	Recurrent neural networks
RRAO	Radio resource allocation optimization
RRH	Remote radio head
RS	Relay station
SA	Standalone
SAF-AP	Space or air fog access point
SC	Superposition coding
SCAP	Small cell access point
SE	Spectrum efficiency
SIC	Successive interference cancelation
SINR	Signal-to-interference-plus-noise ratio
SNR	Signal-to-noise-ratio
SON	Self-organizing network
SV	Stereoscopic video
TDIC	Time domain interference cancelation
TDMA	Time division multiple access
TF-AP	Terrestrial for access point
TTI	Time transmission interval
TU	Transferable utility
UAV	Unmanned aerial vehicle
UE	User equipment
uMTC	Ultra-reliable MTC
URLLC	Ultra-reliable-and-low-latency communications

V2C	Vehicle-to-cloud
V2I	Vehicle-to-infrastructure
V2P	Vehicle-to-pedestrians
V2V	Vehicle-to-vehicle
V2X	Vehicular-to-everything
VR	Virtual reality
WLAN	Wireless local area network
WMMSE	Weighted minimum mean square error
WNC	Wireless network cloud
XR	Extended reality

Index

A
Access slicing paradigm
  background, 85–86
  challenges, 91
  hierarchical management architecture, 86–88
  key techniques
    resource management, 88–89
    social-aware, 89–90
  open issues, 91
Artificial intelligence (AI)
  applications, 211
  architecture and infrastructure design, 38
  backhaul and fronthaul links, 38
  intent-based networking, 9
  network
    edge, 16
    slicing, 38
  principles, 36–38
  SF-AP based LEO satellites, 31
  standard promotions and applications, 38
  system architecture, 35, 36

B
Baseband unit (BBU), 2, 5, 6, 16, 21, 23, 69, 71–74, 78, 86, 117

C
Cache resource, 85, 88, 104, 107, 181
Cell fault diagnosis, 196–197
Centralized optimization
  background, 105
  communication model, 106–107
  cost coefficient, 107–108
  formulation, problem, 108–109
  power model, 107
  problem solution and algorithm design
    algorithm framework, 110
    inner loop algorithm, 111–113
    optimization problem reformulation, 109–110
  simulation results
    content caching, 114
    fronthaul and transmitting method selection, 113
Channel state information (CSI)
  imperfect, 76–79
  optimal contract, 82
  partial, 81–82
  perfect, 74–75, 80–81
Client software
  FAPs, 192
  network controller, 193–195
  UEs, 193
Cloud-fog computation offloading
  hierarchical framework, 154–156
  system model, 153–154
Cloud radio access networks (C-RANs)
  BBU pool, 22
  cloud application server, 26
  H-CRAN system, 21
  overview, 5–7
Communication mode selection, 89, 93, 99, 121
Computation complexity analysis, 96, 104, 133, 149

M. Peng et al., *Fog Radio Access Networks (F-RAN)*, Wireless Networks,
https://doi.org/10.1007/978-3-030-50735-0

Computation offloading
- joint optimization, 162–172
- latency performance, 158–161
- performance analysis, 153–162
- probability, 156–158
- simulation results
  - joint optimization algorithm, 173–177
  - latency performance, 173
- strategy, 161–162
- rate-based utility definition, 72–73

Content caching
- effective capacity, 49–52
- hierarchical framework, 133–134
- impact, 114
- joint optimization, 141–148
- pushing and delivering schemes, 134–136

Content pushing
- asymptotic analysis, 141
- delivering scheme, 135
- discussion, 138–140
- NOMA-based, 136–138
- outage probability, 137–138
- system model, 134–135

Contract design
- channel estimation, 75–76
- CSI, 72–79
- perfect CSIs, 74–75

Contract theory, 69, 72, 80, 83

Cooperative game
- background, 114–117
- coalition
  - formation algorithm, 119
  - formulation, 118
- simulation results and analysis
  - algorithm convergence, 120
  - coalition size, 121
- system model, 117

Cooperative NOMA
- comparison results, 67
- outage probability, 68–70
- transmission, 61, 62

Cooperative signal processing
- BBU, 68–69
- framework, 69–72
- NOMA, 61–68
- numerical results, 79–82
- optimal contract design, 72–75
- practical channel estimation, 75–79
- system model, 69–72

D

Database software
- Docker, 189, 190
- MySQL, 187
- PhpMyAdmin, 187, 188
- Python, 188
- TensorFlow, 189

Deep reinforcement learning (DRL)
- algorithm, 122
- caching model, 123–124
- communication
  - model, 122–123
  - mode selection, 121
- complicated resource allocation, 37
- computing model, 123
- energy consumption model, 124
- problem formulation, 124–125
- selection and resource management
  - algorithm design, 126–127
  - computing resource constraint, 125–126
  - precoding design, 125–126
- simulation results and analysis, 127–129

E

Economical energy efficiency (E^3), 105, 106

Edge cache, 16, 17, 22, 23, 52, 54, 55, 57, 114, 116

Effective capacity
- performance analysis
  - cluster C_T, 54–56
  - expected, 57
  - U$_i$ associated with F_T, 52–54
- simulation results, 57–59
- system model, 49, 51–52
- wireless channels, 134

Energy consumption
- acceptable levels, 8
- CPU-cycle frequency, 163
- and latency, 163, 164
- model, 124
- transmit distance, 3

F

Fault detection, 181, 193, 196–197

Federated learning
- conventional, 207–208
- hierarchical paradigm, 208–209
- loss compensation, 210–211, 213–214
- model compression, 211–214
- potential applications, 209

Fifth generation (5G)
- communication systems, 85
- 5GPPP, 91
- F-RANs, 16–17, 23

heterogeneous wireless transmission, 12
IoT paradigm, 89
network architecture, 203
overview, 7–9
possible capabilities, 10
in 3GPP, 22
Fog computing
cloud server, 13
and edge, 15–16
enabling, 179–181
hierarchical cloud, 214
network structure, 13–15
processors, 155
resource-efficient, 34
UAVs in, 33
Fog radio access network (F-RAN)
access slicing (*see* Access slicing paradigm)
application cases
enabled AR/VR, 30
enabled V2X, 28–29
space communications, 31–33
UAV communications, 33–35
client software, 191–195
cooperative signal processing (*see* Cooperative signal processing)
data support perspective, 204–205
evolution
challenges and future works, 38–39
principles of AI, 36–38
system architecture, AI-driven, 35
federated learning-based paradigms, 206–209
5G, 16–17
hardware platforms, 191
learning paradigm perspective, 205–206
network
architecture perspective, 203–204
controller, 181–182
slicing, 24–25
performance evaluation, 197–200
radio interface techniques, 26–27
relative standards, 18
system architecture, 21–23

G
Game theory, 36, 69, 129

H
Heterogeneous networks (HetNets)
C-RANs, 15
overview, 3–5

I
Incentive compatible (IC), 78
Individual rational (IR), 74, 78, 81
Information awareness, 86
Interference coordination, 3, 4, 41, 71
Internet-of-things (IoT), 7–9, 12–15, 24, 33, 89, 205, 214

J
Joint optimization
computation complexity analysis, 149
problem formulation, 162–166
generation of preference lists, 145–146
matching algorithm, 146–148
stable matching theory, 144–145
simulation results, 149–150, 173–177

K
Key enabling techniques
advanced transmission techniques, 215
fog computing, 214
hierarchical cloud, 214
intelligent NFV, 215
resource management, 215
user scheduling, 215

L
Latency performance, 158–161
computation offloading, 174
C-RAN mode, 38
evaluation, 176
Local model feedback, 217

M
Machine learning (ML), 36, 37, 189, 215
complex problems, 36
GEO resource allocation, 32
multi-disciplinary techniques, 36
network management, 215
raw data, 37
Massive multiple access, 216

N
Network architecture, 9, 24, 85, 203, 206, 217
Network management
access slicing, 87
adaptive and convenient management, 87
dynamic processing, 214
F-RANs, 85
local processing, 206
strategies, 203

Network slicing
- AI-driven, 10, 39
- in F-RANs, 24–25
- intent-based networking, 10

Non-convex optimization, 98, 108

Non-orthogonal multiple access (NOMA)
- content pushing, 136
- delivering scheme, 136
- outage probability, 137–138
- performance analysis, 64–68
- principle, 27
- radio resource, 24
- system model
  - cooperative phase, 62–64
  - transmission phase, 62
- transmission, 61

O

Online resource allocation, 33

OpenAirinterface (OAI), 182, 184–185

Optimization theory, 36, 105

Outage probability
- algorithm, 150
- asymptotic analysis, 141
- cooperative NOMA, 68
- F-AP association strategy, 43
- NOMA scheme, 61, 137–138

P

Performance analysis, 42–48
- effective capacity, 49–59
- ergodic capacity
  - numerical results, 48, 49
  - system model, 41–42
- hierarchical cloud-fog computation offloading
  - framework, 154–156
  - probability and latency analysis, 156–161
  - strategy, 161–162
  - system model, 153–154

0-1 Programming, 169

Q

Q-learning, 98, 99, 102, 126

Qt, 186–187

R

Radio access networks (RANs)
- C-RAN, 5–7
- fog computing (*see* Fog computing)
- 5G, 7–9
- F-RAN (*see* Fog radio access network (F-RAN))
- 6G, 9–12
- HetNet overview, 3–5

Radio environment maps (REMs), 194–196

Radio resource
- computing entity, 5
- content
  - pushing, 150
  - transmissions, 51
- data processing, 23
- management functionalities, 2
- multiple access scenarios, 216
- RRAO, 4

Radio resource allocation optimization (RRAO), 4

Request redirection, 199

Resource allocation
- cache management, 141–148
- centralized optimization, 105–114
- cooperative game, 114–121
- DRL (*see* Deep reinforcement learning (DRL))
- key techniques, 86
- UAVs, 33

Resource management
- background, 91–92
- bandwidth allocation subproblem, 167
- CPU frequency optimization subproblem, 166–167
- joint optimization algorithm, 172
- Lyapunov optimization, 96–98
- multiplexed subchannel allocation, 99–101
- offloading decision subproblem, 169–172
- orthogonal subchannel allocation, 98–99
- problem formulation, 95–96
- simulation results
  - parameters, 101, 102
  - performance comparison, 102–104
- system model
  - communication model, 93–94
  - computing model, 94–95
- transmit power optimization subproblem, 168–169

Resource optimization, 89, 205

S

Signal processing
- advanced edge, 17

cooperative processing (*see* Cooperative signal processing)
physical layer, 26
Sixth generation (6G)
AI-based applications, 209
C-RANs, 6
data support perspective, 204
prospect, 9–12
Soft-defined-radio (SDR) software
OpenAirInterface, 184–185
OpenLTE, 182, 183
SrsLTE, 183–184
Stochastic geometry, 41, 154
System architecture
AI-driven F-RANs, 35, 36
C-RANs, 6
F-RAN (*see* Fog radio access network (F-RAN))

T
Test bed design, 200
Third generation partnership project (3GPP), 4, 7–9, 11, 18, 22, 91, 182, 183
Transmission latency, 30, 133, 203, 210

U
Unmanned aerial vehicle (UAV)
F-RAN enabled, 33–35
SAF-AP, 32

V
Vehicular-to-everything (V2X), 28–29

W
Wireless communications, 10, 22, 38, 185