Energy-Efficient Resource Allocation in Cloud and Fog Radio Access Networks

By

Xiangyu He

A thesis submitted to the University of London for the degree of Doctor of Philosophy

School of Electronic Engineering and Computer Science

Queen Mary University of London

Nov, 2019

Abstract

With the development of cloud computing, radio access networks (RAN) is migrating to fully or partially centralised architecture, such as Cloud RAN (C-RAN) or Fog RAN (F-RAN). The novel architectures are able to support new applications with the higher throughput, the higher energy efficiency and the better spectral efficiency performance. However, the more complex energy consumption features brought by these new architectures are challenging. In addition, the usage of Energy Harvesting (EH) technology and the computation offloading in novel architectures requires novel resource allocation designs.

This thesis focuses on the energy efficient resource allocation for Cloud and Fog RAN networks.

Firstly, a joint user association (UA) and power allocation scheme is proposed for the Heterogeneous Cloud Radio Access Networks with hybrid energy sources where Energy Harvesting technology is utilised. The optimisation problem is designed to maximise the utilisation of the renewable energy source. Through solving the proposed optimisation problem, the user association and power allocation policies are derived together to minimise the grid power consumption. Compared to the conventional UAs adopted in RANs, green power harvested by renewable energy source can be better utilised so that the grid power consumption can be greatly reduced with the proposed scheme.

Secondly, a delay-aware energy efficient computation offloading scheme is proposed for the EH enabled F-RANs, where for access points (F-APs) are supported by renewable energy sources. The uneven distribution of the harvested energy brings in dynamics of the offloading design and affects the delay experienced by users. The grid power minimisation problem is formulated. Based on the solutions

derived, an energy efficient offloading decision algorithm is designed. Compared to SINR-based offloading scheme, the total grid power consumption of all F-APs can be reduced significantly with the proposed offloading decision algorithm while meeting the latency constraint.

Thirdly, an energy-efficient computation offloading for mobile applications with shared data is investigated in a multi-user fog computing network. Taking the advantage of shared data property of latency-critical applications such as virtual reality (VR) and augmented reality (AR) into consideration, the energy minimisation problem is formulated. Then the optimal computation offloading and communications resources allocation policy is proposed which is able to minimise the overall energy consumption of mobile users and cloudlet server. Performance analysis indicates that the proposed policy outperforms other offloading schemes in terms of energy efficiency.

The research works conducted in this thesis and the thorough performance analysis have revealed some insights on energy efficient resource allocation design in Cloud and Fog RANs.

Acknowledgements

Firstly, I would like to express my deepest gratitude to my primary supervisor, Prof. Yue Chen. It is Prof. Yue Chen who introduced me to the research field and provided me with the invaluable guidance as well as the persistent encouragement, helping me go through all difficulties I confronted during my study. Not only she has guided me in the research work, she has also given me the much-needed suggestions for daily lives with patient and immense knowledge. Without her persistent support, my researches are not possible to complete.

I would like to express my appreciation to Dr. Hong Xing, Dr. Kok Keong Chai, Dr. Tiankui Zhang, Dr. Maged Elkashlan, Dr. Jesús Requena Carrión and Prof. Arumugam Nallanathan for their valuable suggestions and comments on my research, which are indispensable for my research progress.

I would also like to express my gratitude towards my family for the encouragement which helped me in completion of my research. Without their love and support over the years none of this would have been possible.

Last but not the least, I would like to thank all my colleagues and friends: Dr. Dantong Liu, Dr. Anqi He, Dr. Bingyu Xu, Dr. Yuanwei Liu, Dr. Yun Li, Dr. Liumeng Song, Dr. Jie Deng, Dr. Jingjing Zhao, Dr. Yuan Ma, Dr. Yanru Wang, Dr. Xingjian Zhang, Dr Zixiang Ma, Dr. Yuhang Dai, Bizhu Wang, Xiaoshuai Zhang, Jinze Yang, Zhong Yang, Xiao Liu, Chao Liu, Wanlin Li, Lanting Zha, Xiaolan Liu, Yan Liu, Haoran Qi, Tianwei Hou among others, for the stimulating discussions and for all the fun we have had in the last four years.

Contents

List of Figures

List of Abbreviations

3G	The Third Generation
4G	The Fourth Generation
5G	The Fifth Generation
AP	Access Point
AR	Augmented Reality
AWGN	Additive White Gaussian Noise
BBU	Baseband Unit
BS	Base Stations
BTS	Base Transceiver Station
CAPEX	Capital Expenditure
CDMA	Code Division Multiple Access
CINR	Channel-to-Interference-plus-Noise Ratio
CoMP	Coordinated Multipoint
CPU	Central Processing Unit
C-RANs	Cloud Radio Access Networks
CRE	Cell Range Extension
D2D	Device-to-Device Communications
DL	Downlink

EE	Energy Efficiency
EH	Energy Harvesting
FDMA	Frequency Division Multiple Access
F-AP	Fog-computing-based Access Point
F-RANs	Fog-computing-based Radio Access Networks
GSM	Global System for Mobile Communications
HetNets	Heterogeneous Networks
HPN	High Power Node
H-CRANs	Heterogeneous Cloud Radio Access Networks
ICT	Information and Communication Technology
IoT	Internet of Things
KKT	Karush-Kuhn-Tucker
LP	Linear Programming
LPN	Low Power Node
LTE	Long Term Evolution
MBS	Macrocell Base Station
MCC	Mobile Cloud Computing
MEC	Mobile Edge Computing
MIMO	Multiple-Input Multiple-Output
MINLP	Mixed Integer Non-linear Programming
NFV	Networks Function Virtualisation
NOMA	Non-Orthogonal Multiple Access
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	Operational Expenditure
QoE	Quality of Experience
QoS	Quality of Service

RAN	Radio Access Networks
RB	Resource Blocks
RF	Radio Frequency
RRH	Remote Radio Head
RSRP	Reference Signal Received Power
SDN	Software Defined Network
SINR	Signal-to-Interference-plus-Noise Ratio
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
UA	User Association
UE	User Equipments
UL	Uplink
VR	Virtual Reality
WLAN	Wireless Local Area Networks

Chapter 1

Introduction

1.1 Background

Wireless mobile data traffic volume has been experiencing continuously rapid growth, and it is estimated that there will be a thousand-fold increase over the next decade [OBB+14]. From the environmental aspects, Information and Communication Technology (ICT) represents nearly 2% of global carbon emissions, where cellular networks account for 0.2% [HBB11]. As it has stepped into the 5G era, there is no sign of slowing down in terms of data traffic increase and the huge energy supply requirement as a result. The future networks with even higher data rate will require more power saving techniques to maximise the the energy efficiency of the upcoming systems [AJ16], hence reduce the OPEX of mobile networks operators. In addition to the data traffic increase, the proliferation of computation-intensive applications in new era such as Virtual Reality(VR) and Augmented Reality(AR) brings in the new power consumption challenges which shorten the mobile users' battery life span. Under this circumstance, mobile cloud computing (MCC) and/or mobile edge computing (MEC) are thought to be the energy-efficient solutions in further cellular networks. While the power saving algorithms are not yet fully explored both in the industry and academia.

To meet the exponentially increased traffic demand, future radio access networks are envisioned to be deployed with novel system architecture. Small cells are envisioned to be densely deployed in given areas to leverage the system throughput by maximising the utilisation of existing spectrum, filling up the coverage holes, and bringing the small base stations closer to mobile users [HSS13]. The above-mentioned concept has been defined as Heterogeneous Networks (HetNets), which will be deployed as a multi-tier architecture. Small cell low power nodes (LPNs) are deployed in dense traffic area with high traffic demands to serve high data-rate packet traffic, while powerful high power nodes (HPNs) are responsible for providing ubiquitous coverage [PLJ+14]. Such densified network architecture has been identified as the dominant deployment scenario for the new generation cellular networks [BLM+14]. As proclaimed in [AJ16], small cells deployment reduces the transmit power required to overcome the pathloss since it shortens the distance between mobile users and base stations, which enables better energy efficiency in both downlink and uplink.

Based on the advancement of cloud computing technologies, Centralised processing, Cooperative radio, Cloud-based infrastructure: Radio Access Network (C-RAN) is presented as the candidate for the next generation radio access networks. The C-RAN brings baseband units (BBUs) form multiple BSs to a central pool location to perform baseband processing, while only the radio units are deployed in remote locations [IRH+14]. Such access points are called remote radio heads (RRHs), which are essentially BSs with reduced functionalities. The centralised processing and operation control is enabled by the signalling between BBU pools and RRHs via the fronthaul links [PWLP15a]. In this context, flexible power control and user association techniques are applicable to further enhance the energy saving of the

new systems.

By going step further on incorporating the cloud computing into the heterogeneity of future radio access networks architecture, the Heterogeneous Cloud Radio Access Networks (H-CRANs) was proposed [PLJ+14]. It provides the heterogeneous networks with the possibility of implementing cooperative interference mitigation schemes through its centralised control. BBU pool and scattered RRHs can be regarded as a virtual BS with distributed antennas, enabling smooth handover between small cells with the aid of coordinated control. All control signalling and system broadcast are handled by high power nodes, which alleviates the fronthaul requirements in terms of the capacity and time delay constraints, while the high data packet traffic is mainly served by RRHs. H-CRANs provide an open, simple, controllable, and flexible paradigm for resource allocation [DDD+15]. With the help of some other enabling technologies, large scale resource sharing in H-CRAN is made possible, which can potentially reduce CAPEX and OPEX [MKGM+15].

Except for the above mentioned cloud-computing based scenarios, system architecture based on fog-computing is also proposed as the fog-computing-based radio access network (F-RAN) [PYZW16]. It brings the storage, management and computation to the edge of the network, closer to mobile users to alleviate the potential traffic congestion in the fronthul that causes the long transmission delay [PWLP15b]. Then the heavy burden on the fronthaul networks can be alleviated without having to establish links to the centralised cloud storage and computing. In addition, the transmission delay can be reduced as well [KLL+17].

1.2 Research Motivation

The network densification in the next generation RANs will introduce staggering rise in energy consumption. In order to decrease energy consumption and further save operational cost, energy harvesting is anticipated in next generation systems [PLZW15]. In H-CRAN architecture, small cell transmission points require relatively lower amount of power supply comparing to their macrocell counterparts, which is able to be partly or even fully empowered by renewable energy sources. Unlike conventional grid power supply, the availability of renewable energy suffers high level of uncertainty and fluctuation. Traditional resource management methods are far from enough to address new challenges. Novel schemes are needed so that the renewable energy supply can be better utilised to support the energy efficient operation of future RANs.

The mobile-cloud-computing or mobile-edge-computing enabled by the next generation system architectures bring in applications' computation offloading scenarios, which offloads the computation data from mobile users to the external computing nodes. The computations of mobile applications are offloaded either for saving energy of mobile user or reducing latency of computation execution with the help of external computing nodes. Conventional offloading decision and communications resource allocations cannot be readily applied to the new architectures. As a result, researches on developing novel joint optimisation of offloading decision and resource allocations algorithms are required to achieve the full potentials of new system architecture.

1.3 Research Contributions

In this thesis, the focus is given to the energy-efficient resource allocation for novel RAN architectures. It is aimed to minimise the energy consumption of novel cellular networks architecture in different scenarios, especially the consumption of non-renewable energy that supplied by traditional power grid. In Chapter 3, the non-renewable grid power consumption is aimed to be minimised by proposing an energy efficient user association and power allocation scheme in H-CRAN which maximises the usage of green energy provided by energy harvesting technology. In addition, it is aimed to maximise the grid energy efficiency through proposing an iterative optimisation algorithm in Chapter 3 as well. In Chapter 4, a delay-aware computation offloading scheme for F-RAN is presented, aiming to reduce the non-renewable energy consumption under the hybrid energy supply scenario. In Chapter 5, the energy consumption of the MEC system is minimised through proposed joint computation offloading and communications resources allocation scheme.

The contributions of the thesis are summarised as follows.

- An extensive overview of the state-of-the-art user association, computation offloading, and resource allocation schemes for system architectures of next generation radio access networks is carried out. In addition, open challenges of joint optimisation of the above-mentioned aspects are highlighted, which clarifies the research motivation.

- In the two-tier H-CRAN where macro cells are empowered by grid power and remote radio heads are empowered by renewable energy sources, an energy efficient radio resource optimisation algorithm is proposed to maximise the utilisation of the green power collected by energy harvesters for the RRHs. It leads to the reduction of grid power

consumption for an improved energy efficiency of the whole network. Through applying the joint user association and power allocation algorithms constantly, the instability of power supply from energy harvester can be addressed.

- In the fog-computing-based radio access networks architectures, a novel computation offloading strategy in F-RAN where all access points are equipped with renewable energy sources is proposed. The computation offloading can be implemented to all neighboring access points. The algorithm coordinates the computation offloading according to the availabilities of renewable energy to minimise the grid power consumption. Additionally, the characteristics of fog-computing with multiple choice of offloading paths from the serving F-AP to neighbouring F-APs are reflected in the investigated scenario.

- In the multi-user fog computing system, a scenario where multiple single-antenna mobile users running applications featuring shared data is considered. Mobile users' overall energy consumption is minimised via joint optimisation of computation offloading and communications resource allocation. The optimal solution for the energy minimisation problem of shared-data featured offloading is found, which provides in-depth understanding of the shared-data featured offloading in MEC systems.

- Through mathematical analysis is presented in the design of all proposed algorithms. The effectiveness in energy saving of all investigated scenarios is validated through comprehensive simulations.

1.4 Author's Publication

1. **Xiangyu He**, Hong Xing, Yue Chen, Arumugam Nallanathan, "Energy-Efficient Computations and Communications Resource Allocation for Mobile Applications Featuring Shared Data", submitted to *IEEE Transactions on Communications*.

2. **Xiangyu He**, Anqi He, Yue Chen, Kok Keong Chai and Tiankui Zhang, "Energy Efficient Resource Allocation in Heterogeneous Cloud Radio Access Networks," 2017 *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1-6, Mar. 2017.

3. **Xiangyu He**, Yue Chen and Kok Keong Chai, "Delay-Aware Energy Efficient Computation Offloading for Energy Harvesting Enabled Fog Radio Access Networks," 2018 *IEEE 87th Vehicular Technology Conference (VTC-Spring)*, pp. 1-6, June 2018.

4. **Xiangyu He**, Hong Xing, Yue Chen, Arumugam Nallanathan, "Energy-Efficient Mobile-Edge Computation Offloading for Applications with Shared Data", 2018 *IEEE Global Communications Conference (Globecom)*, pp. 1-6, Dec. 2018.

1.5 Thesis Organisation

Chapter 2 introduces the fundamentals of all related system architecture concepts, such as H-CRAN and F-RAN. The state-of-the-art user association, computation offloading and communication resources allocation algorithms are summarised as well. Open challenges of joint optimisation of the above-mentioned aspects are highlighted. In addition, the main methodology Convex Optimisation is briefly introduced in Chapter 2.

Chapter 3 presents an energy efficient radio resource optimisation algorithm in the two-tier H-CRAN where macro cells are empowered by grid power and remote radio heads are empowered by renewable energy sources. The theoretical analysis and performance evaluation are carried out to prove that proposed algorithm achieve better energy saving performance compared to conventional ones. Then an iterative resource allocation algorithm in H-CRAN is proposed to maximise the energy efficiency of the grid power supply.

Chapter 4 introduces a novel computation offloading strategy in the fog-computing-based radio access networks architectures, where all access points are equipped with renewable energy sources. Theoretical analysis is presented to illustrate how to coordinates the computation offloading according to the availabilities of renewable energy to minimise the grid power consumption.

Chapter 5 investigates the joint optimisation of computation offloading and communications resource allocation in the multi-user fog computing system, where multiple single-antenna mobile users running applications featuring shared data. The detailed analysis on how the shared data property can be utilised to reduce mobile users' energy consumption is presented. Moreover, the performance evaluation is carried out to prove the effectiveness in energy saving compared to all other offloading scenarios.

Chapter 6 summarises the conclusions drawn from current researches and future work.

Chapter 2

Fundamental Concepts and State-of-the-Art

This chapter introduces the fundamentals of next generation system architectures, such as H-CRAN and F-RAN. The state-of-the-art user association, computation offloading and communication resources allocation algorithms are summarised as well. In addition, the main methodology Convex Optimisation is briefly introduced.

2.1 Energy Efficient Radio Access Networks

2.1.1 Power Control in Conventional Cellular Networks

Among all generations of mobile networks, energy efficiency is consistently regraded as one of the most important performance indicators. In GSM system, frequency division multiple access (FDMA) and time division multiple access (TDMA) are applied as the multiplexing schemes [Sch04]. In this case, power control is relatively simple since different mobile users occupy orthogonal channels. The base stations control the power level not too high to avoid energy waste, but sufficient to maintain a good signal to noise ratio.

In the 3G mobile communications system, code division multiple access (CDMA) techniques are applied as the multiplexing scheme [Sch04], where

multiple mobile users share the same frequency band when communicating between user equipments and base stations. In this scenario, the system is vulnerable to "near-far" problem, the open-loop and the closed-loop power control techniques are designed as the solution [YCL09]. Notably, the primary purpose of these power control schemes is to increase the capacity of all users, while saving the energy consumption is just of secondary importance [KMM95].

Before launching the commercial 4G LTE operations, power allocation algorithms aiming to maximise the energy efficiency in OFDMA-based systems have been investigated [KH00], [KGH03], [MHLB08]. Due to more advanced and complex implementations of 4G networks compared to that of the previous generations, novel energy efficient power allocation researches are continuously proposed to further improve the system's energy saving performance, even after the practical deployment of 4G systems has been completed.

In future mobile networks generations, the envisioned RANs architecture evolution will bring in more complexities in their operation, which calls for novel power allocation and power control designs.

2.1.2 Energy Harvesting Technologies

As a consensus, small cell deployment is envisioned to be the candidate RAN architecture of future cellular networks. However, it is not cost-effective to provide all small cell BSs with grid power supply [MLZL15]. Due to the difficulty in supplying the small cells deployment with grid power and the environmental concern on the consumption of non-renewable energy, energy harvesting can be utilised to improve energy efficiency of the cellular networks [HA13]. [BHZ15] reveals that the hybrid solar-wind energy harvester is the

ideal candidate of renewable energy supply.

However, the utilisation of EH in cellular networks brings design challenges because of the fluctuation and uncertainty of the renewable energy supply. The availability of harvested energy differs in varied environmental circumstances, which makes the networks' reliability very hard to guarantee [OTUY15]. In this context, the novel communication protocols, resource allocation and user association schemes are needed to improve the applicability of EH in future cellular networks.

2.2 Next-Generation RANs Architectures

2.2.1 Heterogeneous Networks

Heterogeneous networks is a new mobile networks paradigm as illustrated in Figure 2.1, which is the multi-tiered deployment of small cells access points within the coverage of macrocells. These small cells access points are also described as low-power nodes because of their relatively low level of transmit power. Through densely deploying these low-power nodes, which is also referred as spatial densification, HetNets are expected to be the dominant approach to boost system capacity [HSS13].

The improvement of system capacity is achieved through several aspects. Firstly, for some environments, there are high level of fading and penetration loss which weaken the signal from macrocell base stations. As a consequence, there exist coverage holes within the coverage of the macrocell. The deployment of small cells is able to fill up the coverage holes. Secondly, data traffic can be offloaded from macrocell transmission points, reducing the possibility of traffic congestion. Most importantly, higher degree of resource reuse is en-

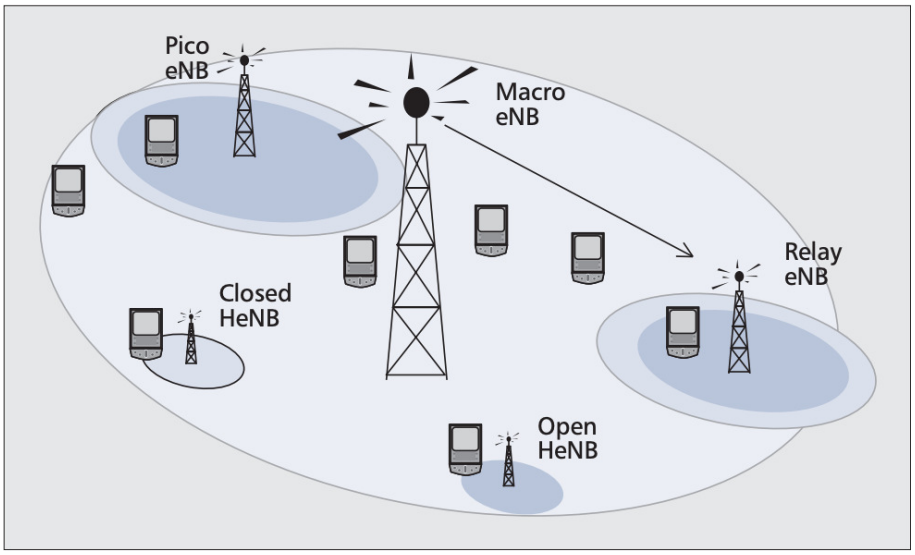


Figure 2.1: Heterogeneous networks (HetNet) with the mix of high power BS and low power BSs [DMW+11].

abled by dense deployment of small cells where each low-power node serves only a small number of mobile users.

2.2.2 Cloud Radio Access Network

C-RAN is proposed as a future radio access networks architecture so as to provide mobile broadband Internet access to wireless customers with low bit-cost, high spectral and energy efficiency. It is a natural evolution toward the distributed Base Transceiver Station (BTS), which is composed of the base-band unit, remote radio heads, backhaul and fronthaul links connecting each components [Chi11]. Thanks to the development of cloud-computing technologies, there implements virtualised BBU Pool in C-RAN where baseband resources of different access point are aggregated under centralised control as depicted in Figure 2.2. In traditional radio access networks, baseband processing capacity is exclusively occupied by each base stations, which impedes the system from operating in an adjustive manner. RRHs are basically

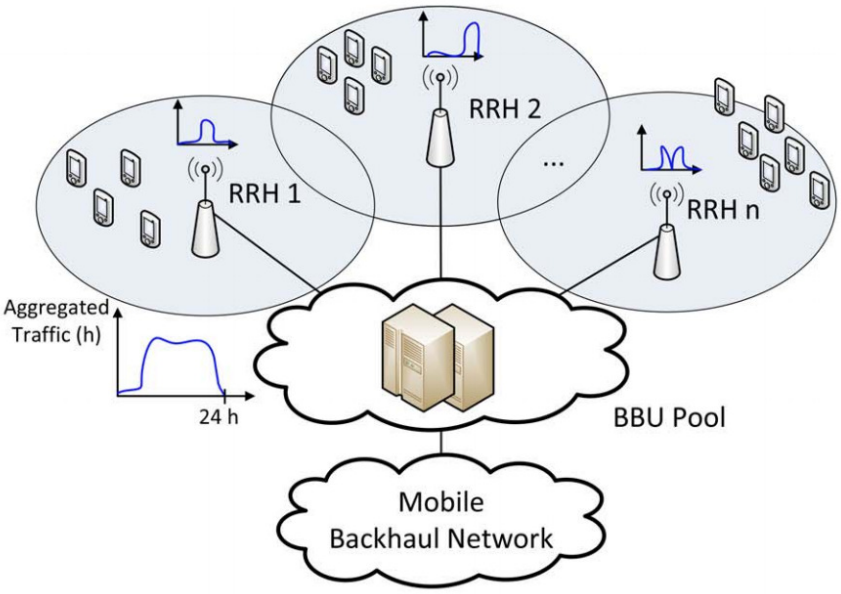


Figure 2.2: Cloud radio access networks with RRHs and centralised BBU pool [CCY+15].

base stations with reduced functionalities, which are only responsible for digital processing, analogue-digital conversion, power amplification and filtering [CCY+15]. Fronthaul capacity is recognised as a critical aspect since the effectiveness of the collaborative operation relies heavily on it.

Many of the conventional operating schemes for resource allocation and interference coordination are envisioned to be enhanced by being adopted in the C-RAN system.

2.2.3 Cloud-assisted Implementation of HetNets

Network densification is regarded as the dominant approach to meet the exponential data traffic increase. However such deployment introduces high level of operational complexity in terms of increased difficulties in management and system control. Additionally, advanced cooperative schemes are needed for the inter-cell interference mitigation in HetNets. As a result,

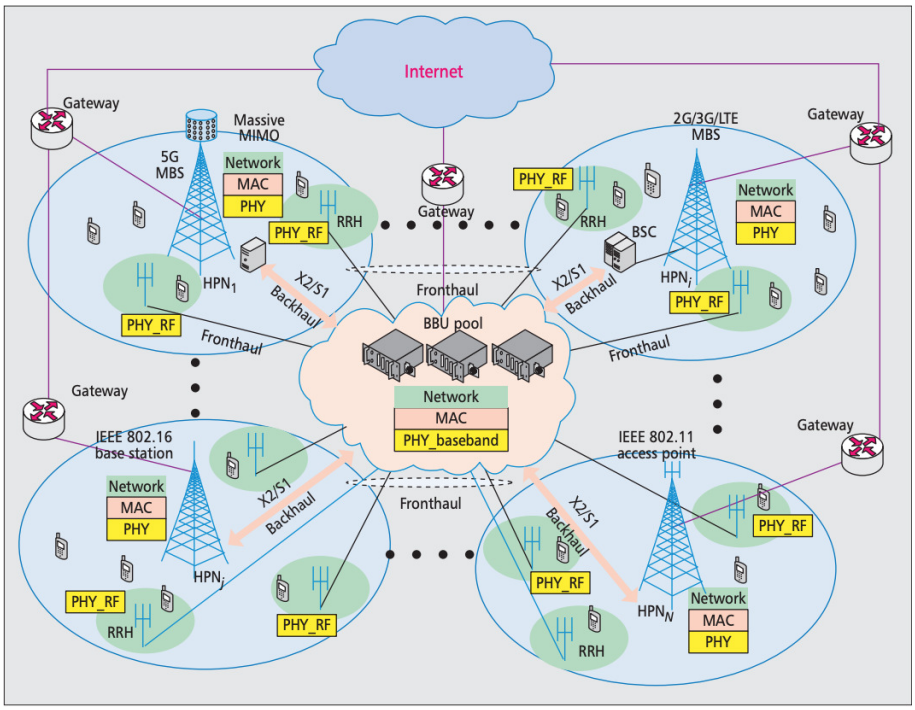


Figure 2.3: System diagram of H-CRAN with different types of APs merging into the united architecture [PLJ+14].

small cell access points are envisioned to be connected to the cloud for centralised management and control [ZCG+15]. This is made possible by incorporating HetNets into C-RAN. It gives rise to a novel architecture known as heterogeneous cloud radio access networks (H-CRAN), which is proposed as a cost-effective potential solution to alleviate inter-tier interference and improve cooperative precessing gains in HetNets through combination with cloud computing [PLJ+14].

The fundamental difference between C-RAN and H-CRAN is the heterogeneity. H-CRAN is a multi-tiered architecture where high-power nodes and low-power nodes coexist. BBU pool in H-CRAN is interfaced with HPN to mitigate inter-tier interference via interference collaboration and beamforming [PXC+15]. The control signalling and system broadcasting are delivered to UEs by HPN, which alleviated the burden of fronthaul link for

signalling exchange. The control signalling and data delivery are decoupled in H-CRANs.

2.2.4 Fog-computing-based Radio Access Networks

It is recognised that the major problems in the centralised architectures is the potential traffic congestion in the fronthul that causes the long transmission delay [PWLP15b]. The fog-computing-based radio access network is introduced by bringing the storage, management and computation to the edge of the network. As indicated in Figure 2.4, in the F-RAN architecture, content servers provide large-scale caching capability, and the controller is used for network control like resource management. In addition, the cloud contains multiple processors of heterogeneous computing capabilities which are connected with each other to achieve computing resource sharing [SPM19]. The heavy burden on the fronthaul networks can be alleviated without having to establish links to the centralised cloud storage and computing. Additionally, the transmission delay can be reduced as well [KLL+17].

The access points equipped with caching, radio signal processing and radio resource management capabilities are referred as the fog-computing-based access points (F-APs). F-APs are evolved from traditional remote radio heads in cloud-assisted platform, support the mobile edge computing.

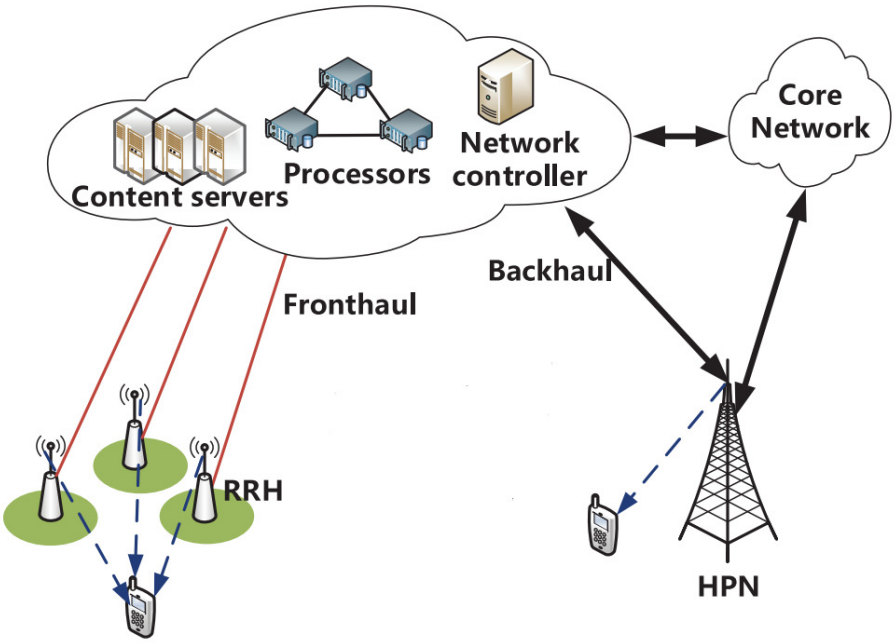


Figure 2.4: The fog-computing based radio access networks [SPM19].

2.3 Resource Allocation in Next-Generation RAN architectures

2.3.1 Radio Resource Management for Next-Generation RANs

Thanks to the cloudification and virtualisation of network functionalities, resource sharing is being extensively applied to realise the H-CRAN's full potential. In [MKGM+15], the resource sharing were investigated in three levels: spectrum sharing, infrastructure sharing and network sharing. Spectrum sharing among operators is regarded as a choice of enlarging the available pool of resources to improve the spectral efficiency, it can be performed through different allocation units, such as RF channels on WLAN and resource blocks of LTE frames [MOP+14]. It is pointed out that the H-CRAN

enables the sharing at the level of symbols through joint processing, while in LTE sharing can only occur in a resource block. Since an H-CRAN can be viewed as a large-scale, highly capable cognitive radio where several distributed radios are connected to a central processing element, it enables the application of cognitive radio. Infrastructure sharing among network operators is discussed in four scenarios, which is envisioned to massively reduce CAPEX and OPEX. With the concept of software defined networks (SDN) and network function virtualisation (NFV), the orchestration of networks can be improved [BdlOS+14]. Network sharing solutions are based on the abstraction of network functionalities of physical level and others.

2.3.1.1 User Association in Cloud-based RAN architectures

With the resource sharing enabled by the system's new characteristics, the H-CRAN is identified as an open, simple, controllable and flexible paradigm for resource allocation [DDD+15]. User association strategies, which is the decision of which user roams to which BS, influence the performance of wireless networks to some extent. The H-CRAN moves the networks toward user-centric architectures where every user can communicate with more than one heterogeneous nodes at the same time. Finding the optimal user association strategy becomes a combinatorial optimisation problem with high complexity.

In [SY14], the optimal joint BS association with power control and beamforming was considered for the downlink HetNets. It followed a pricing-based strategy in which the users are associated with the BS according to the value of a utility at the lowest cost, and showed that the proposed pricing-based distributed user association can significantly improve the conventional max-SINR association. In [ZQL13], the user association in coordinated multipoint

(CoMP) downlink transmission was addressed involving the joint design of transmit beamformers and user data allocation at BSs to minimise the user data transfer from the data center to the BSs in the backhaul. It is notable that the cloud-based architecture will replace the conventional powerful base stations by low-cost low-power RRHs, forming a green and low-cost infrastructure. However, the necessary connections between all the RRHs and the BBU pool require significant power consumption for the transport network. To address this issue, there proposed a novel framework to design a green Cloud-RAN, which is formulated as a joint RRH selection and power minimisation beamforming problem [SZL14]. In both references [ZQL13] [SZL14], the user association problem involves solving a sparse beamforming problem that returns which user should be served by which BSs. It can be noticed that the majority of the research outcomes are for the single-cloud scenario where the whole network is connected to the same cloud. At present, there have already been some research works about the multi-cloud scenario. For example, authors of [DANA15] studied the user-to-cloud-assignment problem by maximizing a network-wide utility subject to practical cloud connectivity constraints, which made an effort toward devising a distributed cloud association strategy via an auction-based iterative approach. In addition, [AAA+19] proposed the framework for the resource allocation in multi-cloud C-RAN industrial Internet of Things scenario, and a low complexity heuristic algorithm to solve the constraint resource allocation problem in linear time.

The energy efficiency-based user association problem in massive MIMO empowered C-RAN was investigated in [ZZY+16]. Three UA algorithms were proposed for different scenarios, which achieved higher energy efficiency than the baseline nearest RRH association scheme. Specifically, the proposed multi-candidate RRHs user association algorithm achieved a good balance

between spectral and energy efficiency, and the performance gain is more significant when the number of users is large. The user association strategy in massive MIMO enabled HetNets was also studied in [LWC+15]. A low complexity distributed UA algorithm was developed for energy efficient fair user association while considering quality of service provision for users, which improves the energy efficiency and user fairness compared to other user association algorithms. A joint downlink and uplink user association and beamforming design to coordinate interference in the C-RAN for energy minimisation was proposed in [LZL15], which is to address the severe interference and also insufficient energy consumption due to the high density of active access points and more stringent uplink requirements in future RANs. This paper was the first attempt to unify the downlink and uplink user association and beamforming design into one general framework. The effectiveness of the proposed algorithms in ensuring the feasibility of both DL and UL transmissions, achieving optimal network energy saving, flexibly adjusting various power consumption trade-offs between active APs and mobile users is verified via extensive simulations. [YA18] took user QoS into consideration when designing user association and BBU-RRH mapping strategy. It minimised the system cost incurred by the energy bill from RRHs and VB rentals. The baseband unit in the system model can be actualised by a virtual machine, making it virtual BBU that can be initiated or shut down as needed to server clusters of RRHs.

2.3.1.2 Energy Efficient Resource Allocation in Next-Generation RANs

The H-CRAN is envisaged to implement a separation between the control plane and the data plane [MKGM+15]. Different to conventional HetNets,

the computation burden of control signalling is carried out by the processors in the central cloud in H-CRANs. According to this new characteristics, [DDD+15] outlined some promising resource allocation strategies in H-CRAN, which are coordinated scheduling, hybrid backhauling, and multicloud association. Scheduling, in the resource allocation context, is the decision on which each user is active at every resource block. With inter-BS coordination in a user-centric architectural model, classical proportional fairness scheduling is no more valid for H-CRANs. In [DDANA15], the authors presented an efficient graph-theoretical-based approach for solving the complicated coordinated scheduling problem in H-CRANs. The heuristic algorithms with low computational complexity were proposed to maximise a network-wide utility under the practical constraint. Simulation results suggested that the proposed algorithms perform near optimal in low shadowing environments. It provided a generic framework that can be used in other resource allocation problems in H-CRANs including power control, user association, and channel assignment. Joint precoding and backhaul compression are studied in both single-cloud [PSSS13] and multi-cloud scenarios [PSSS14]. Additionally, several compression techniques are proposed in an uplink single-cloud framework [ZY14].

By saying H-CRANs user-centric, the guarantee of Quality of Service (QoS) is very important. There is one thing in common for both researches that is the QoS guarantee is the paramount constraint required as a prior condition before any system performance metric to be optimised. In a more recent research, a dynamic resource allocation scheme with three consecutive steps is proposed for H-CRAN in TDD mode, which enhances the bidirectional data rate with low complexity [YWJ+16]. Authors in [PZJ+15] proposed an energy efficient resource allocation scheme for H-CRAN by enhanc-

ing traditional soft fractional frequency reuse. The UEs are grouped into two categorises with high QoS and low QoS requirement, respectively. Then the resource sharing techniques are deigned to meet the data rate requirement of the users, with high priority, while mitigating the inter-tier interference in H-CRAN. Simulation results proved the improvement in energy efficiency when applying enhanced frequency reuse scheme if the resource allocation strategy is designed properly in H-CRAN architecture.

The excessive power usage in heterogeneous architecture is a critical issue to be resolved in future mobile networks. Moreover, future wireless networks are expected to meet the diverse QoS requirements imposed by current and envisioned services apart from achieving higher data rate. [TSA+15] addressed the above-mentioned issues by jointly considering transmit beamforming design and power allocation policies for a heterogeneous real-time and non-real-time traffic to optimise the system's energy efficiency. The proposed algorithm was proved that it can efficiently approach the optimal EE. Additionally, because of the strong intracell and intercell interference due to dense deployment and spectrum reuse, EE and QoS are consequently degraded in future networks if there is no novel solutions towards such issue. The proposed algorithm in [ZDO+16] utilised joint channel selection and power allocation design to improve the QoS performance. In this scenario, the centralised BBU pool is responsible for carrying out interference cancellation and transmission power optimisation. The effectiveness is validated by simulation results achieving a nearly "zero" infeasibility ratio. The infeasibility ratio is defined as the probability of not satisfying QoS requirement. The EE performance is improved by 300% for cellular UEs.

2.3.1.3 Utilisation of Renewable Energy in Next-Generation RANs

Among ideas aiming at improving H-CRAN's performance, utilising renewable energy supply by energy harvesting is a hot research topic [PLZW15]. As a new energy supply option for wireless communication systems, energy harvesting has attracted a lot of attention from industry and academia. Utilising the green power harvested from the renewable energy sources, i.e. solar or wind power, instead of the conventional electricity grid power can effectively reduce OPEX for network operators. However, the inherent intermittent nature of harvested green power challenges the reliable QoS provision in wireless systems [OTUY15]. Hence in practical networks, only the low power base stations (BS) or wireless access points are usually empowered with renewable energy sources, while the high power BSs, i.e. macrocell BSs, are still powered with the steady grid power.

Optimisation of green power base station deployment is investigated in [PE13], covering three aspects that are heterogeneous deployment of base stations, using of renewable base stations and adjustment of transmission powers. The results show that power consumption decrease significantly without sacrificing other performance factor like spectral efficiency and coverage. Moreover, the proposed scheme does not only reduce carbon emission rates, but also decrease the operational cost of the cellular network operators. An user association scheme with the assumption that all BSs are solely empowered with green power harvested from renewable energy sources in developed in [XCE+16], which proves that its proposed algorithm achieves higher throughput than conventional UAs. In [ZXL+15], user association algorithm for HetNets with renewable energy supply is studied, which is proved to be effective in achieving load-balancing and making the best use of renewable energy. The proposed algorithm is proved to converge to the

global optimum of the formulated optimisation problem, which is to maximise the BS utility proportional fairness. The two-dimensional optimisation to lexicographically minimise the on-grid energy consumption in heterogeneous networks with hybrid energy sources is presented in [LCC+15], which involved the optimisation in both the space and time dimensions. An optimal offline algorithm with low complexity was proposed, serving as performance upper bound for evaluating practical online algorithms. Some heuristic online algorithms were further developed. Both offline and online algorithms designed in this paper outperform other algorithms in terms of total and peak on-grid energy consumption reductions. None of these papers has considered the possibility that in H-CRAN small cell serving nodes are solely empowered with green power source while macrocell base stations are still connected to power grid to guarantee the smooth operation. A suitable algorithm for this new consideration is yet to be investigated.

2.3.2 Computation Offloading in F-RANs and MEC Systems

With the advent of the era of IoT, the unprecedented growth of latency-critical applications are nevertheless hardly satisfied by MCC alone. To cater for the low-latency requirements while alleviating the burden over backhaul networks, Mobile Edge Computing, also interchangeably known as *fog computing* has aroused a paradigm shift by extending cloud capabilities to the very edge within the radio access network [MYZ+17]. Both industry and academia have devoted constant effort to fully exploit the potential brought by fog-computing-based architecture. Among pioneering industrialisation on fog computing, Cisco has proposed fog computing as a promising candidate for IoT architecture [BMZA12]. In recent researches, [ZWG+13], [KKLC15],

[YHC16], [LCLH15] focused on one-to-one offloading scheme where there is one mobile user and one corresponding cloudlet, [KNWH13], [CSBC16] presented multiple-user cases where there are multiple edge servers, while [Che15] related to multiple-to-one scenarios where multiple mobile users offload computing to one edge server. In [OSSB15], the authors proposed an AP clustering strategy for distributed fog computing applications based on the minimisation of the transmit power. The proposed approach allows adaptive sizing and resources management of computation clusters, and in the mean time establishes computation clusters for all active requests for better exploitation of available resources, targeting a higher QoE. In [LMZL16] ,the computation task scheduling was researched for MEC systems in terms of shortening the execution delay. The computation tasks are scheduled based on the queueing state of the task buffer, the execution state of the local processing unit, as well as the state of the transmission unit. A power-constrained delay minimisation problem was formulated, and an efficient one-dimensional search algorithm was proposed to find the optimal task scheduling policy that succeeded in achieving a shorter average execution delay. A comprehensive computation offloading solution that uses the multiple radio links available for associated data transfer was provided in [MSS15]. The research presented in [SML$^+$18] examined the computation offloading scenario in mobile wireless sensor networks. An optimal partition is proposed to minimise the total energy consumption in cooperative computing. Going one step further by utilising the optimal results, the authors proposed energy efficient cooperation node selection strategies to achieve fairness and maximal energy saving in a multi-node environment.

2.3.2.1 Joint Design of Computation Offloading and Communication Resources Allocation

The joint optimisation of computation offloading with communications resources (such as power, bandwidth, and rate) proves to improve the performance of fog computing by explicitly taking channel conditions and communications constraints into account. In an early research [WGKN08], the offloading decision making was examined through the estimation of data bandwidth without considering the allocation of communication resources and channel conditions. For communications-aware computation offloading, [XLXN18] minimised the local user's computation latency in a multi-user cooperative scenario, while [WXWC18] minimised the energy consumption of remote fog computing nodes. A game theoretic approach is applied in [GZQL12] for the minimisation of overall energy consumption in MCC system. It is declared to be the first paper that aims to reduce the overall energy consumption of a MCC system under computation offloading context. In [TL17], the authors investigated the joint computation offloading and resource allocation problem exploiting computing resources from both cloud offloading and D2D cases. The proposed scheme achieves high energy saving gains compared with the local computation strategy and cloud offloading strategy. In an MIMO multicell system where multiple mobile users asking for computation offloading to a common cloud server, joint optimisation of radio resources and computational resources is formulated in order to minimise the users' overall energy consumption [SSB15]. The algorithmic framework presented was show that it naturally leads to a distributed and parallel implementation across the radio access points, while only a limited coordination/signalling with the cloud is required. Numerical results show that the proposed schemes outperform disjoint optimisation algorithms.

The authors in [BSD13] propose a method that jointly optimises the transmit power, the number of bits per symbol and the CPU (central processing unit) cycles assigned to each application in order to minimise the power consumption of the mobile users. A one-to-one relation between the power allocated on each channel and the percentage of CPU cycles assigned to the corresponding user was found, and the performance of a scheduler aimed at stabilising the computation queues accumulating on each mobile handset was presented as well. The research scope of [ZWG+13] is divided into two independent sections, one for the scenario of local mobile execution, and the other for the cloud execution where task data are offloaded to edge server for execution. In mobile execution case, the CPU frequency of the mobile user is reconfigured optimally to obtain the least energy consumption. In cloud execution scenario, the transmission rate is optimised in response to the stochastic channel condition to minimise the transmission energy. After that, the optimal application execution policy is designed to choose between mobile execution and cloud execution that consume less energy on the mobile device. A general MEC system with multiple users and one MEC server with limited computation capability is considered in [MZSL17]. The performance metric adopted is the average weighted sum power consumption of the mobile devices and the MEC server. The available radio and computational resources are jointly managed to optimise the MEC system, including the CPU-cycle frequencies for the mobile and server CPUs, the transmit power and bandwidth allocation for computation offloading, as well as the task scheduling decision at the MEC server. It proposes the algorithms that are able to balance the weighted sum power consumption and execution delay performance. However, these line of work have not taken the recently noticed shared data feature into account, thus failing to fully reap the advantage of

fog computing. The resource allocation for a multiuser MEC system based on CDMA/OFDMA accounting for both the cases of infinite and finite cloud computation capacities was researched in [YHCK17]. Algorithms with low complexity that are able to achieve close-to-optimal performance were proposed for all scenarios in this paper, which reduce the weighted sum mobile energy consumption significantly.

2.3.2.2 Computation Offloading for Applications featuring Shared Data

Recently, the intrinsic collaborative properties of the input data for computation offloading was investigated for augmented reality in [ASS17]. Compared to conventional independent offloading across mobile users, the novel resource allocation approach utilising shared data feature presents considerable gains in energy saving performance. In fact, in many mobile applications such as augmented reality and virtual reality, multiple mobile devices share parts of computing input/output in common. The shared-data aware MEC in the emerging 5G applications helps to further improve the system performance (low latency and/or energy). In [PPS18], some important insights on the interplay among the social interactions in the VR mobile social network was revealed, and a significant reduce on the end-to-end latency was achieved through stochastic optimisation technique. [CSYD17] investigated potential spatial data correlation for VR applications to minimise the delay of accomplishing computation. The proposed algorithm can intelligently transfer information on the learned utility across time, and allow adaptation to environmental dynamics due to factors such as changes in the users' data correlation. [WXD17] studied computation offloading in a multi-antenna non-orthogonal multiple access (NOMA)-based system. The users partition

the tasks for local processing and offloading according to the proposed algorithm to minimise the energy consumption of all users. Excpet for the cellular networks, the research on the cooperative energy efficient computation offloading in WLAN networks utilising the similarity of the computation tasks is also presented, where an online task scheduling algorithm is designed to achieve a desirable trade-off between the energy consumption and Internet data traffic by appropriately setting the trade-off coefficient [JYM+14].

2.4 Convex Optimisation

Mathematical modelling is of great importance in designing and optimisation of mobile networks. The optimal algorithms design in terms of energy consumption minimisation relies heavily on obtaining the analytical solutions of the formulated optimisation problem. In this thesis, the main methodology applied in reaching mathematical solutions for theoretical analysis is *Convex Optimisation*.

The theory of Convex Optimisation has been developed for more than a century. As a special category of mathematical optimisation problems, the global optimum of such problems can be obtained very efficiently. There exists well-developed problem-solving techniques for convex optimisation such as ellipsoid methods and sub-gradient methods [Boy]. Then by recognising or formulating a problem as a convex optimisation problem, it can be solved through these techniques very reliably and efficiently. Generally, a typical convex optimisation problem in the following form:

$$\underset{x}{\text{minimise}}\ f_0(x) \tag{2.1}$$

$$\text{subject to} \quad f_i(x) \leq b_i, i = 1, ..., m, \tag{2.2}$$

where the functions $f_0, ..., f_m : \mathbb{R}^n \to \mathbb{R}$ are convex, and satisfy

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y) \tag{2.3}$$

for all $x, y \in \mathbb{R}^n$ and all $\alpha, \beta \in \mathbb{R}$ with $\alpha + \beta = 1, \alpha \geq 0, \beta \geq 0$.

$f_0(x)$ is the objective function of the problem, the vector $x = (x_1, ..., x_n)$ is the optimisation variables of the problem. As for the functions $f_i, \mathbb{R}^n \to \mathbb{R}, i = 1, ..., m$, they are the inequality constraint functions. The objective function and the constraints functions are all required to be convex so that the problem is in the convex form. Then it is able to find the optimal solution x^*, which gives the objective function the smallest value among all vectors that satisfy the constraint.

A widely used techniques in solving resource allocation problems in wireless networks is Lagrange Duality theory. It is applied in all our investigated problems to obtain the optimal solutions, which in turn generates the energy efficient resource allocation design. However, the problems encountered in this thesis may not be readily formulated as standard convex optimisation problems. For example in the User Association problems, the existence of binary variables put the problem into the category of mixed-integer nonlinear programming (MINLP) problems, which are non-convex. In addition, the complexity of target functions or optimisation constraints can also hinder the applicabilities of standard convex optimisation methods. Fortunately, the ways of reformulating original problems can always be found based on some realistic assumptions, which transformed the original problems into traceable ones where standard convex optimisation methods are applicable.

The widely known *water-filling* algorithm in information theory is a typical application of convex optimisation in the area of communication re-

searches. It originates from the following optimisation problem:

$$\text{minimise} \quad -\sum_{i=1}^{n} \log(\alpha_i + x_i)$$

$$\text{subject to} \quad x \succeq 0, \mathbf{1}^\mathsf{T} = 1.$$

The variable x_i represents the transmitter power allocated to the i-th channel, and $\log(\alpha_i + x_i)$ gives the capacity or communication rate of the channel, so the problem is to allocate a total power of one to the channels, in order to maximise the total communication rate. The *water-filling* algorithm is derived through the process of getting the solution of this optimisation problem.

Chapter 3

Resource Allocation Optimisation in H-CRAN with Hybrid Energy Sources

This chapter presents an energy efficient radio resource optimisation algorithm in the two-tier H-CRAN where macrocell HPNs are empowered by grid power and RRHs are empowered by renewable energy sources. The proposed algorithm achieves better energy saving performance compared to conventional ones. Moreover, an iterative resource allocation algorithm in H-CRAN is proposed to maximise the energy efficiency of the grid power supply.

3.1 System Model

The scenario considered is a two-tier H-CRAN deployment as illustrated in Figure 3.1. One tier of it is the macrocells tier whose coverage is served by electricity gird powered macrocell HPNs, and the other one is made up of picocells served by picocell RRHs with energy harvesting power resources (EH-RRH). Grid power minimisation within a single macrocell is considered since the analysis result can be readily extended to other macrocells. Within one macrocell, $m \in \mathcal{M}$ denotes m-th EH-RRH in it. For the ref-

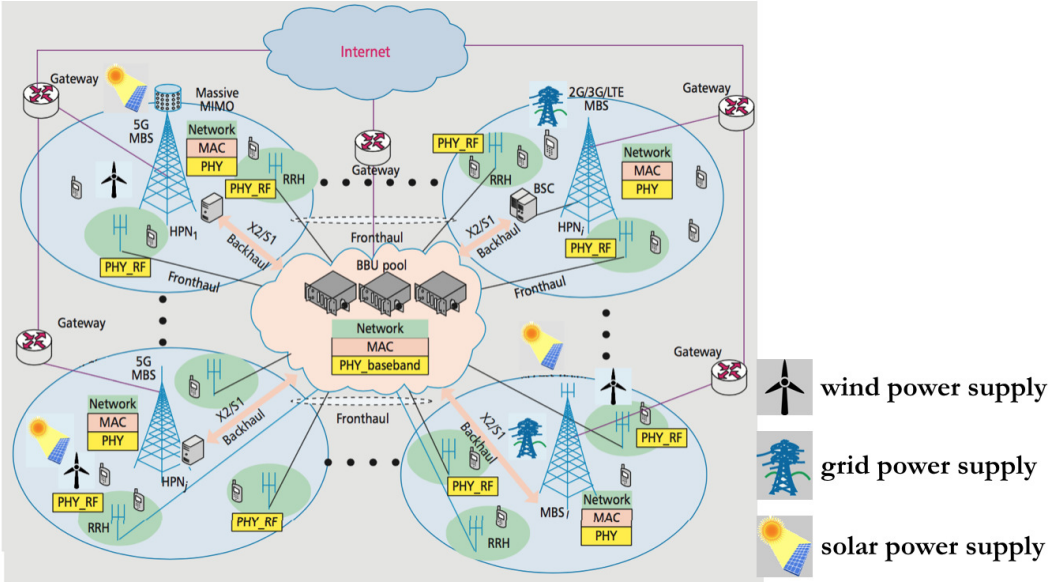


Figure 3.1: The overview of energy harvesting powered H-CRAN.

erence macrocell, there are 12 EH-RRHs uniformly-distributed around the high power node. $u \in \mathcal{U}$ denotes u-th User Equipment connecting to the system, with total number of U. For every UE requesting service, the system will decide whether to associate it with nearest picocell RRH or HPN. Then Resource Blocks (RBs) are allocated to the UE from corresponding transmission points. The system bandwidth of 20MHz containing 100 RBs available.

3.1.1 Energy Harvesting Model

The Energy Harvesters providing picocell RRHs with power supply are considered to be undergoing a stationary stochastic process with the probability density function $f_m(z_m) = 1/(b_m - a_m), \forall z_m \in [a_m, b_m]$ where a_m and b_m are the lowest and highest power harvesting level of m-th RRH. Within one time slot with the duration of 20 ms, the harvesting power is thought to be constant, but may vary in different time slots. In practical scenario, the energy

harvesting rate may remain at the same level for several seconds.

3.1.2 Energy Consumption Model

In real systems deployment case, there is a static circuit power consumption for each RRH denoted as P_{sc}, with the fronthaul power consumption P_{fh} be taken into consideration as well. According to the power consumption model provided by [AGD+11], the amount of available power for data transmit of RRH is given by

$$P^R_{avail} = \kappa^R [p_{EH} - P_{sc} - P_{fh}, 0]^+ \tag{3.1}$$

where $[x]^+ = max\{x, 0\}$, and $\eta^R = 1/\kappa^R$ is the efficiency of power amplifier. This expression tells us that the energy harvesting rate should be larger than that of overall static power consumption of each RRH. Otherwise, it cannot support any data service requests from user equipments.

3.1.3 Downlink transmission model

A important feature of Cloud Radio Access Network is the cooperative operation which enables coordinated allocation of radio resources, avoiding the intra-tier interference among RRHs [PZJ+15]. Moreover, thanks to the relatively low level of transmission power of picocell RRHs, UEs connecting to them can be regarded as suffering no interference from other cells when applying proper frequency reuse schemes. The Carrier-to-Interference-plus-Noise Ratio (CINR) for the the u-th UE to m-th off-grid EH-RRH is expressed as

$$g_{u,m} = h^R_{u,m}(d_{u,m})/B_0 N_0 \tag{3.2}$$

N_0 is the power spectral density of thermal noise. $h^R_{u,m}$ is the path loss as a function of propagation distance between u-th UE and m-th RRH [PZJ+15]

$$h^R_{u,m}(d) = 31.5 + 40\log_{10}(d_{u,m}) \tag{3.3}$$

As for the HPN, due to the much higher level of transmitting power, intra-tier interference among HPNs should be taken into consideration. Accordingly, advanced radio resource reuse scheme is applied for this heterogeneous scenario so that inter-tier interference is negligible. Then the CINR for the u-th UE connected to HPN is given as

$$g_u = h^M_u(d_u)/(B_0 N_0 + P^M_h h^M_{u,h}(d^M_{u,h})) \tag{3.4}$$

P^M_h, $h^M_{u,h}(d^M_{u,h})$ denote the total transmission power of h-th neighbouring HPNs and path loss from h-th neighbouring HPNs to u-th UEs, respectively. $h^M_u(d_u)$ is expressed as

$$h^M_u(d_u) = 15.3 + 37.6\log_{10}(d_u) \tag{3.5}$$

Given all afore-mentioned expressions, the sum data rate of all RRHs in the macrocell is written as:

$$R(\mathbf{a}, \mathbf{p}) = \sum_{u\in\mathcal{U}}\sum_{m\in\mathcal{M}} a_{u,m} B_0 \log_2(1 + p_{u,m} g_{u,m}) \tag{3.6}$$

and the sum data rate of HPN is

$$R^M(\mathbf{a}^M, \mathbf{p}^M) = \sum_{u\in\mathcal{U}} a^M_u B_0 \log_2(1 + p^M_u g_u) \tag{3.7}$$

In above functions, two U×M matrices $\mathbf{a} = [a_{u,m}]_{U\times M}$ and $\mathbf{p} = [p_{u,m}]_{U\times M}$,

two U×1 matrices $\mathbf{a}^M = [a_1^M, \ldots, a_U^M]^T$ and $\mathbf{p}^M = [p_1^M, \ldots, p_U^M]^T$ are introduced, which represent the user association and power allocation policies of RRHs and HPN, respectively. $a_{u,m}$ and a_u^H are binary user association indicator whose value can either be 0 or 1, indicating whether u-th UE is associated to m-th RRH or HPN. $p_{u,m}$ and p_u^M represent the transmission power of m-th RRH and HPN to u-th user equipment.

The energy consumed by HPN to transmit data is expressed by

$$P(\mathbf{a}^M, \mathbf{p}^M) = \sum_{u \in \mathcal{U}} a_u^M p_u^M + P_c^M + P_{bh}^M \tag{3.8}$$

where P_c^M and P_{bh}^M are static circuit power consumption and power consumption for backhaul signalling.

3.2 Energy Efficient Resource Allocation in H-CRAN with Hybrid Energy Sources

3.2.1 Motivation

With the anticipation of mass deployment of small cell networks, it is difficult to provide grid power supply to all of them in a cost-effective way. Energy harvesting provides a suitable solution to this problem. However, new challenges brought by the uncertainty and fluctuation of renewable energy supply need to be addressed. Moreover, because of the disparity of transmit power among transmission points of different tiers, simple user association algorithms such as Reference Signal Receiving Power (RSRP)-based user association cannot achieve best system performance. In order to fully

exploit the utility of harvested energy, this research is conducted to find a good solution on saving grid power.

3.2.2 Problem Formulation

The user association and power allocation algorithms is formulated as an optimisation problem whose target is to minimise the grid power consumed by HPN achieved by offloading the data traffic form HPN to RRH as much as possible. Rate-constrained QoS requirement is introduced, a minimum date rate threshold R_{min} should be achieved as the prerequisite for grid power minimisation.

The objective function together with the constraints are give as below:

$$\min_{\{\mathbf{a}^M, \mathbf{p}^M\}} \quad P(\mathbf{a}^M, \mathbf{p}^M) = \sum_{u \in \mathcal{U}} a_u^M p_u^M + P_c^M + P_{bh}^M \tag{3.9}$$

s.t.

$$\sum_{m \in \mathcal{M}} a_{u,m} + a_u^M = 1, a_{u,m} \in \{0,1\}, a_u^M \in \{0,1\}, \forall u \tag{3.10}$$

$$p_{u,m} \geq 0, p_u^M \geq 0, \forall u, \forall m \tag{3.11}$$

$$\sum_{m \in \mathcal{M}} R_{u,m} + R_u^M \geq R_{min}, \forall u \tag{3.12}$$

$$\sum_{u \in \mathcal{U}} a_{u,m} p_{u,m} \leq P_{avail}^R, \forall m \tag{3.13}$$

where $R_{u,m} = a_{u,m} B_0 \log_2(1+p_{u,m} g_{u,m})$ and $R_u^M = a_u^M B_0 \log_2(1+p_u^M g_u)$. The first constraint indicates that each UE can only be connected to one node, either the selected RRH or HPN. The second one specifies the minimum data-rate requirement for each users, with R_{min} is the data-rate threshold. The third constraint puts the limitation on the overall transmit power of

EH-RRHs, defined by their harvested power.

It is clear that the originally formulated optimisation problem is an non-convex optimisation problem due to the existence of binary variables in the objective function and the user association constraint. More precisely, it is a mixed integer programming problem. Classical convex optimisation methods cannot be applied directly to solve it. Hence, it has to be reformulated in order to obtain the optimal user association policy $\mathbf{a}^*$ and the optimal power allocation policy $\mathbf{p}^*$ for EH-RRHs with corresponding constraint.

In addition to the non-convexity, the difficulties in solving it is also due to the co-existence of optimal resource allocation policy pairs $\mathbf{a}^H$, $\mathbf{p}^H$ and $\mathbf{a}$, $\mathbf{p}$. The optimisation problem will be definitely simpler to solve by reducing it to a form where the all expressions contain only one resource allocation policy pair of the two.

A conclusion can be readily drawn that the minimisation of grid power consumption is correspondent to the maximisation of green power supported throughput. Therefore the algorithm is formulated to a green power maximisation problem which is solely related to $\mathbf{a}$, $\mathbf{p}$. After solving such reformulated problem, through the relationship given by the first constraint of previous optimisation problem, the overall solution of the original problem can be obtained. The green power throughput maximisation problem is propose as below:

$$\max_{\{\mathbf{a},\mathbf{p}\}} \quad R(\mathbf{a},\mathbf{p}) = \sum_{u\in\mathcal{U}}\sum_{m\in\mathcal{M}} R_{u,m} \tag{3.14}$$

s.t.

$$a_{u,m} \in \{0,1\}, p_{u,m} \geq 0, \forall u \in \mathcal{U}, \forall m \in \mathcal{M}, \tag{3.15}$$

$$\sum_{m \in \mathcal{M}} R_{u,m} \geq R_{min}, \forall u \in \mathcal{U}, \tag{3.16}$$

$$\sum_{u \in \mathcal{U}} a_{u,m} p_{u,m} \leq P^{R}_{avail}, \forall m \in \mathcal{M}. \tag{3.17}$$

where $R_{u,m} = a_{u,m} W_0 \log_2(1 + p_{u,m} g_{u,m})$. The first constraint describes the user association constraint that each UE can only be associated with one AP, either the selected RRH or HPN. It is assumed that each UE can connected to only one BS. The second constraint specifies the minimum data-rate requirement for each UE, with R_{min} is the data-rate threshold; being lower than that is considered to be unacceptable. The third one puts the limitation on the overall transmit power of each RRH.

3.2.3 Energy efficient joint user association and power allocation algorithm

To obtain the solutions of the formulated optimisation problem, we should firstly have the Lagrangian of the primal objective function [Boy]:

$$\begin{aligned} L(\mathbf{a}, \mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\nu}) &= \sum_{u \in \mathcal{U}} \sum_{m \in \mathcal{M}} a_{u,m} W_0 \log_2(1 + p_{u,m} g_{u,m}) \\ &+ \sum_{u \in \mathcal{U}} \lambda_u \Big(\sum_{m \in \mathcal{M}} a_{u,m} W_0 \log_2(1 + p_{u,m} g_{u,m}) - R_{min} \Big) \\ &+ \sum_{m \in \mathcal{M}} \nu_m \Big(P^{R}_{avail} - \sum_{u \in \mathcal{U}} a_{u,m} p_{u,m} \Big), \end{aligned} \tag{3.18}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_U) \succeq 0$ is the Lagrange multiplier vector correspond-

ing to the required minimum date rate constraint (3.16). $\boldsymbol{\nu} = (\nu_1, \nu_2, ..., \nu_M) \succeq 0$ is the Lagrange multiplier vector associated with the remote radio head transmission power constraint (3.17). The elements of the vectors are all non-negative.

Then the Lagrangian dual function is given as:

$$
\begin{aligned}
g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \max_{\{\mathbf{a},\mathbf{p}\}} \quad & L(\mathbf{a}, \mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\
\max_{\{\mathbf{a},\mathbf{p}\}} \quad & \{\sum_{u\in\mathcal{U}} \sum_{m\in\mathcal{M}} [(\lambda_u + 1) a_{u,m} W_0 log_2(1 + p_{u,m} g_{u,m}) \\
& - \nu_m a_{u,m} p_{u,m}] - \sum_{u\in\mathcal{U}} \lambda_u R_{min} + \sum_{m\in\mathcal{M}} \nu_m P^R_{avail}\}
\end{aligned}
\tag{3.19}
$$

and then the dual optimisation problem is reformulated as:

$$
\begin{aligned}
\min_{\{\boldsymbol{\lambda},\boldsymbol{\nu}\}} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\
\text{s.t.} \quad & \boldsymbol{\lambda} \succeq 0,\ \boldsymbol{\nu} \succeq 0
\end{aligned}
\tag{3.20}
$$

Obviously, the dual optimisation problem is convex. With fixed $a_{u,m}$ and $p_{u,m}$, the Lagrangian function $L(\mathbf{a}, \mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is linear with λ_u and ν_m, and the dual function $g(\boldsymbol{\beta}, \boldsymbol{\nu})$ is the maximum of these linear functions. Applying dual decomposition method to solve this dual problem, then firstly

it is decomposed into U independent problem:

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \sum_{u \in \mathcal{U}} g_u(\boldsymbol{\lambda}, \boldsymbol{\nu}) - \sum_{u \in \mathcal{U}} \lambda_u R_{min} + \sum_{m \in \mathcal{M}} \nu_m P^R_{avail} \tag{3.21}$$

where

$$\begin{aligned} g_u(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \max_{\{\mathbf{a},\mathbf{p}\}} \quad & \{ \sum_{m \in \mathcal{M}} (\lambda_u + 1) a_{u,m} W_0 log_2(1 + p_{u,m} g_{u,m}) \\ & - \nu_m a_{u,m} p_{u,m} \} \end{aligned} \tag{3.22}$$

Assuming that the m-th RRH is associated to the u-th UE, i.e., $a_{u,m} = 1$, then apparently (3.21) is concave in terms of $p_{u,m}$. Apply the Karush-Kuhn-Tucker (KKT) condition, the optimal power allocation is derived by setting $\partial g_u / \partial p_{u,m} = 0$, which is:

$$p^*_{u,m} = [\omega^*_{u,m} - \frac{1}{g_{u,m}}]^+ \tag{3.23}$$

the optimal water-filling level $\omega^*_{u,m}$ is given as:

$$\omega^*_{u,m} = \frac{(\lambda_u + 1) W_0}{ln2 \nu_m} \tag{3.24}$$

By substituting the optimal power allocation obtained by (3.23) and the water-filling level (3.24) into the decomposed problem (3.21), it becomes

$$g_u(\boldsymbol{\lambda}, \boldsymbol{\nu})$$

$$= \max_{1 \le m \le M} \{(\lambda_u + 1) W_0[\log_2(\omega^*_{u,m} g_{u,m})]^+ \tag{3.25}$$

$$- \nu_m[\omega^*_{u,m} - \frac{1}{g_{u,m}}]^+\}$$

Then the optimal user association indicator is determined by

$$a^*_{u,m} = \begin{cases} 1, & m = arg \max\limits_{1 \le u \le U} D_{u,m}, \\ 0, & otherwise, \end{cases} \tag{3.26}$$

where

$$\begin{aligned} D_{u,m} &= (\lambda_u + 1)[\log_2(\omega^*_{u,m} g_{u,m})]^+ \\ &\quad - \frac{(\lambda_u + 1)}{ln2}[1 - \frac{1}{\omega^*_{u,m} g_{u,m}}]^+ \end{aligned} \tag{3.27}$$

Making use of the sub-gradient method to solve the dual problem, the sub-gradient of the dual function is written as

$$\nabla \lambda_u^{(i+1)} = \sum_{m \in \mathcal{M}} R^{(i)}_{u,m} - R_{min} \tag{3.28}$$

$$\nabla \nu_m^{(i+1)} = P^R_{avail} - \sum_{u \in \mathcal{U}} a^{(i)}_{u,m} p^{(i)}_{u,m} \tag{3.29}$$

where $a_{u,m}^{(i)}$ and $p_{u,m}^{(i)}$ is the user association and power allocation policy derived by the dual variables of the i-th iteration, respectively. $R_{u,m}^{(i)} = a_{u,m}^{(i)} B_0 log_2(1+p_{u,m}^{(i)} g_{u,m})$. $\nabla\lambda_u^{(i+1)}$, $\nabla\nu_m^{(i+1)}$ represent the sub-gradient utilised in the $(i+1)$-th iteration. Accordingly, the update equations for the dual variables in the $(i+1)$-th iteration are expressed as:

$$\lambda_u^{(i+1)} = [\lambda_u^{(i)} - \alpha_\lambda^{(i+1)} \times \nabla\lambda_u^{(i+1)}]^+ \tag{3.30}$$

$$\nu_m^{(i+1)} = [\nu_m^{(i)} - \alpha_\nu^{(i+1)} \times \nabla\nu_m^{(i+1)}]^+ \tag{3.31}$$

with $\alpha_\lambda^{(i+1)}$ and $\alpha_\nu^{(i+1)}$ are the positive step sizes.

3.2.4 Simulation Results and Performance Evaluation

The H-CRAN architecture is comprised of macrocells within each there is one HPN in the center and 12 EH-RRHs uniformly distributed around it. Each resource block occupies a bandwidth of 180kHz and one resource is allocated to one UE at a time. The UEs are randomly located inside the macrocell. The simulation of RB and power allocation is repeated based on the time slot of 1ms, which is the minimum time unit in OFDMA systems. During each time slot, the energy harvesting rates remain constant. According to [AGD+11], the power amplifier efficiency of macrocell HPN is $\eta_{PA}^R = 31.1\%$. It it assumed the static circuit power consumption of HPN to be $P_c^M = 10.3W$ and that of picocell RRH is 0.1W ;as for the fronthaul link and the backhaul link power consumption P_{bh}^M and P_{fh}, they are both assumed to be 0.2W [PZJ+15]. The maximum transmit power of MBS is 20W (43dBm). The lowest and highest power harvesting level are 22dBm and 48dBm respectively. The thermal noise level is -174dBm/Hz. With 20MHz

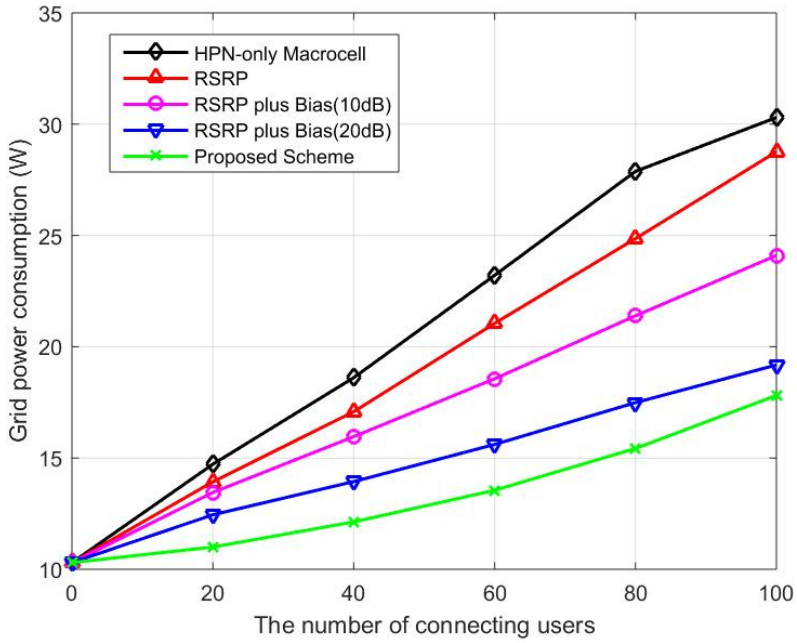


Figure 3.2: Grid power consumption with different number of UEs connecting to the system, $R_{min} = 384kbps$.

as the overall system bandwidth, there are totally 100 Resource Blocks available to be assigned to the UEs.

The single tier HPN-only system is presented as the baseline architecture, where there is only one HPN in the cell center taking on all connection requests. This scenario does not need to consider any specific user association or RB allocation algorithm since the only option for all users is to be associated to the center HPN. The power consumption model is the same as that of the H-CRAN HPN described in the system model section.

Reference signal receiving power (RSRP) is used as the baseline user association algorithm. In RSRP algorithm, a UE is associated to the BS with the strongest received reference signal. If there is not enough power available to satisfy the minimum data rate requirement, the UE would be redirected

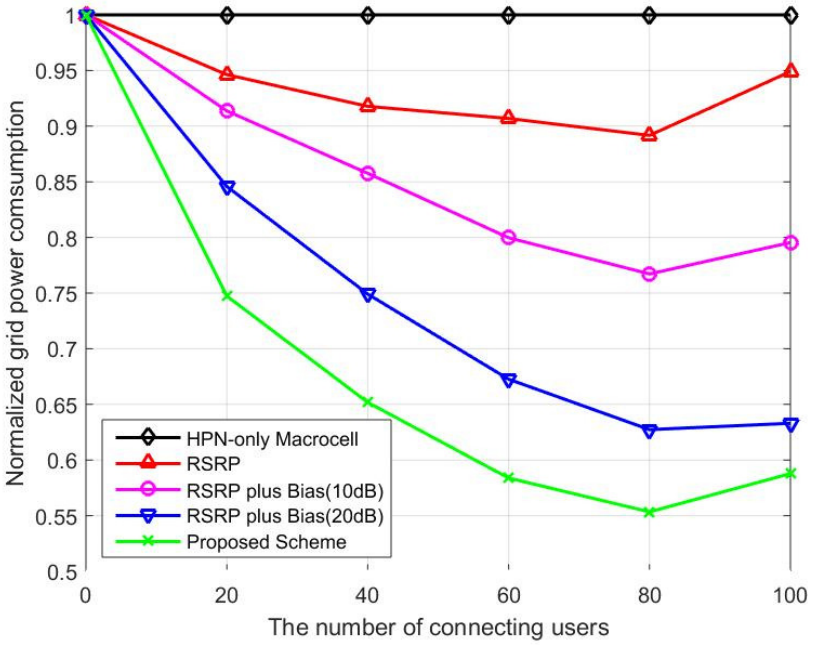


Figure 3.3: Normalised grid power consumption, $R_{min} = 384kbps$.

to the BS with the second highest RSRP, and then so on and so forth. Intuitively, such scheme cannot achieve the best traffic offloading because of the inherent transmit power gap between those of HPN and picocell RRHs. Hence this transmitted signal strength gap needs to be compensated by applying signal bias, as proposed as cell range extension (CRE) in [KBTV10]. The signal bias is the extent to which one response is more probable than another in signal detection.

As shown in Figure 3.2, under the date rate constraint of 384kbps, heterogeneous architecture with RSRP mechanism improves the grid power saving, by offloading mobile data traffic from the macrocell tier to picocell RRHs.

Figure 3.3 shows how much grid power can be saved by applying the proposed algorithm in comparison to the baseline single tier system architecture.

The grid power consumption of the base line system is normalised to one. For RSRP and RSRP-plus-bias schemes, the higher the bias value, the lower the grid power consumption. For example, when 40 users are served, around 8%, 15% and 25% of grid power are saved with RSRP, RSRP-plus-10dB, and RSRP-plus-20dB bias respectively. While around 35% grid power can be saved with the proposed algorithm compared to the baseline algorithm. The grid power saving is down to the improved green power utilisation. It is interesting to observe that the percentage of the grid power saving, in another word, the reduction in grid power consumption is related to the number of users in the system. Under the scenario set in this chapter, the grid power saving performance improves as the number of UEs increases as higher percentages of grid power is saved compared to reference HPN-only architecture, for all algorithms between 0 to 80 UEs. The decreasing speed of the normalised grid power consumption for the proposed optimisation scheme is slowing down as the number of users gets larger, revealing that high density of users diminishes the grid power saving. It is also found that the best grid power saving performance is achieved around 80 users connecting to the system. Note that this peak point for grid power saving is related to the system set up, i.e. number of resource blocks and number of RRHs in each macrocell.

Figure 3.4 presents the green power utilisation and grid power consumption percentage of the proposed optimisation scheme. Percentage results for grid power consumption is obtained from the assumption that 20W is the maximum transmit power level for HPN. The green power utilisation rate increases as the number of connecting users gets higher, and approaches to around 70 percent when there are 100 users being served. It is clear from Figure 3.4 that not all green power gets utilised. This is down to the UE

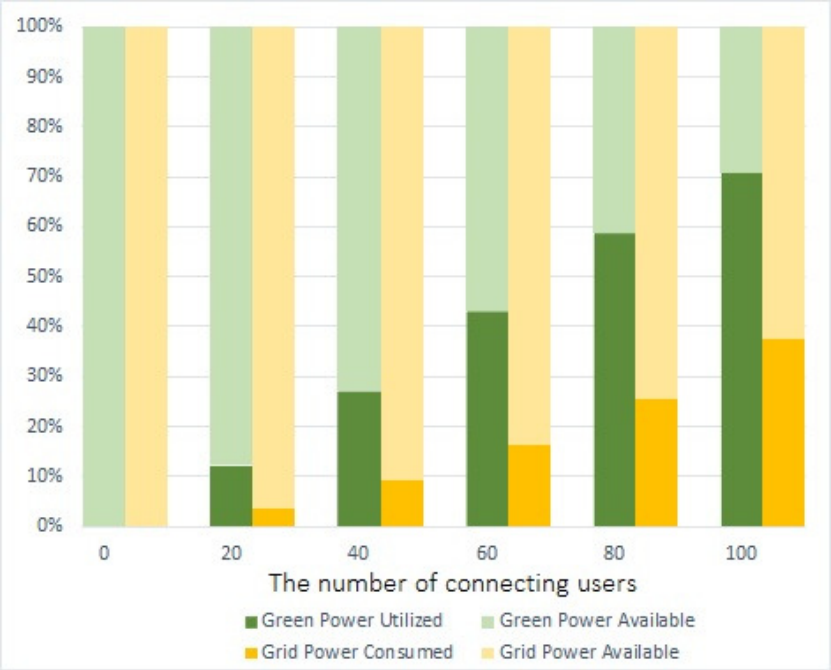


Figure 3.4: Green power and grid power utilisation.

data rate requirments. There might be some residual green power in some RRHs, but it is not enough to support the required date rate. How to utilise this part of residual green power is an interesting direction for future work as an extention of the contribution. More advanced mechanisms are needed to take full advantage of available green power supply. Multipoint transmission where multiple RRHs transmits signal to one single user in a cooperative way if none of the RRHs have enough green power to serve the user by itself, might be an effective mechanism which could further reduce grid power consumption in H-CRAN.

3.3 Iterative Resource Allocation Algorithm for Green Energy Aware H-CRAN

3.3.1 Motivation

In previous sections, we propose an energy efficient resource allocation and user association algorithm based on the formulated optimisation problem that aims to minimise the grid energy consumption of the radio access networks. However, the proposed algorithm cannot guarantee a high transmission rate as a result of taking energy minimisation as the objective. In another word, the proposed algorithm in previous section is suitable for the system whose ultimate concern is the sustainability reflected by non-renewable energy consumption. However, it can still investigate the radio access networks that not only regard the energy consumption as the main performance indicator, but also take the data throughput into account. In this chapter, we present a framework that considers both the energy consumption and the data throughput, which maximises the defined grid energy utility by the proposed iterative resource allocation algorithm.

3.3.2 Problem Formulation

The grid power consumption of m-th AP is given as:

$$P_m^{grid} = [P_{m,0} + \kappa_m \sum_{u \in \mathcal{U}} a_{u,m} p_{u,m} - P_m^g]^+, \quad \forall m \in \mathcal{M} \tag{3.32}$$

where κ_m is the inverse of the power amplifier efficiency of AP m, and $P_{m,0}$ is the static power consumption, P_m^g is the harvested sustainable energy from the green power supply.

The Signal-to-Interference-plus-Noise-Ratio (SINR) between UE u and AP m is

$$\gamma_{u,m} = \frac{p_{u,m} G_{u,m}}{I_{u,m} + \sigma^2} \tag{3.33}$$

where $I_{u,m}$ is the average interference to the downlink transmission from m-th AP to u-th UE [XH12].

Then given the expression of the SINR, the expression of the sum data rate for the whole system based on the Shannon's equation is given by:

$$R(\mathbf{a}, \mathbf{p}) = \sum_{u \in \mathcal{U}} \sum_{m \in \mathcal{M}} a_{u,m} W_0 \log_2(1 + \gamma_{u,m}), \tag{3.34}$$

and the fronthaul capacity requirement of m-th AP is the sum throughput of all the UEs associated with it:

$$Z_m = \sum_{u \in \mathcal{U}} a_{u,m} W_0 \log_2(1 + \gamma_{u,m}), \forall m \in \mathcal{M}. \tag{3.35}$$

The overall amount of grid power consumption is written in the following equation:

$$G(\mathbf{a}, \mathbf{p}) = \sum_{m \in \mathcal{M}} P_m^{grid} = \sum_{m \in \mathcal{M}} [P_{m,0} + \kappa_m \sum_{u \in \mathcal{U}} a_{u,m} p_{u,m} - P_m^g]^+ \tag{3.36}$$

which is the overall static power consumptions and transmitting power subtracting the energy harvesting rates for all picocell remote radio heads and high power node.

Here the grid power utility of the system is introduced, which is defined as the ratio of the overall system throughput over the grid power consumption:

$$\varphi = \quad \frac{R(\mathbf{a}, \mathbf{p})}{G(\mathbf{a}, \mathbf{p})} \tag{3.37}$$

The optimisation is considered to increase the system throughput without adding too much grid power consumption. Then the objective function is given, expressed as the maximisation of the defined grid power utility:

$$\max_{\{\mathbf{a},\mathbf{p}\}} \frac{R(\mathbf{a}, \mathbf{p})}{G(\mathbf{a}, \mathbf{p})} = \frac{\sum_{u \in \mathcal{U}} \sum_{m \in \mathcal{M}} a_{u,m} W_0 \log_2(1 + \gamma_{u,m})}{\sum_{m \in \mathcal{M}} P_m^{grid}} \tag{3.38}$$

s.t.

$$a_{u,m} \in \{0, 1\}, \sum_{m \in \mathcal{M}} a_{u,m} = 1, \forall u \in \mathcal{U}, \tag{3.39}$$

$$Z_m \leq B_m, \forall m \in \mathcal{M}, \tag{3.40}$$

$$\sum_{u \in \mathcal{U}} a_{u,m} p_{u,m} \leq P_m^{max}, \forall m \in \mathcal{M}, \tag{3.41}$$

$$\sum_{m \in \mathcal{M}} a_{u,m} W_0 \log_2(1 + \gamma_{u,m}) \geq \delta_u, \forall u \in \mathcal{U}. \tag{3.42}$$

where constraint (3.39) defines that each user equipment can only be associated with one AP. The constraint (3.40) puts the limitation on the fronthaul capacity of each AP, while (3.41) corresponds to the maximum transmit power of every RRH and HPN. δ_u in (3.42) denotes the minimum data rate of u-th UE, which describes of rate-constrained QoS requirement of the user association.

3.3.3 Iterative Resource Allocation Algorithm Proposal

According to the proof of [PZJ+15], the optimal grid power utility φ^* is achieved if and only if

$$\max_{\{\mathbf{a},\mathbf{p}\}} R(\mathbf{a},\mathbf{p}) - \varphi^* G(\mathbf{a},\mathbf{p}) = R(\mathbf{a}^*,\mathbf{p}^*) - \varphi^* G(\mathbf{a}^*,\mathbf{p}^*) \tag{3.43}$$

Moreover, for all feasible $\mathbf{a}$, $\mathbf{p}$, and φ, $F(\varphi) = R(\mathbf{a},\mathbf{p}) - \varphi G(\mathbf{a},\mathbf{p})$ is a strictly monotonic decreasing function about φ, and $F(\varphi) \geq 0$. Hence, the update algorithm is proposed to find the optimal value of φ:

Algorithm 1 Outer Loop Algorithm

1: Set the maximum number of iterations I_{max}, convergence condition ϵ_φ and the initial value $\varphi^{(1)} = 0$
2: Set the outer loop iteration index $i = 1$ and begin the outer loop iteration
3: **for** $1 \leq i \leq I_{max}$ **do**
4: Solve the user association and power allocation problem with $\varphi^{(i)}$
5: Obtain $a^{(i)}$, $p^{(i)}$, $R(a^{(i)},p^{(i)})$ and $G(a^{(i)},p^{(i)})$
6: **if** $R(a^{(i)},p^{(i)}) - \varphi^{(i)} G(a^{(i)},p^{(i)}) < \epsilon_\varphi$ **then**
7: Set $\{a^*,p^*\} = \{a^{(i)},p^{(i)}\}$ and $\varphi^* = \varphi^{(i)}$
8: break;
9: **else**
10: $i = i+1$, $\varphi^{(i+1)} = \dfrac{R(a^{(i)},p^{(i)})}{G(a^{(i)},p^{(i)})}$
11: **end if**
12: **end for**

Based on the analysis above, the optimisation problem is reformulated into a more tractable form:

$$\max_{\{\mathbf{a},\mathbf{p}\}} \quad R(\mathbf{a},\mathbf{p}) - \varphi^{(i)} G(\mathbf{a},\mathbf{p}) \tag{3.44}$$

$$s.t. (3.39) - (3.42),$$

where $\varphi^{(i)}$ is the obtained utility after i-th update.

Considering all the constraints given above, the Lagrangian expression of the objective function is given as:

$$
\begin{aligned}
L(\mathbf{a}, \mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\nu}) = & \sum_{u \in \mathcal{U}} \sum_{m \in \mathcal{M}} a_{u,m} W_0 \log_2(1 + \gamma_{u,m}) \\
& - \varphi^{(i)} \sum_{m \in \mathcal{M}} \left(P_{m,0} + \kappa_m \sum_{u \in \mathcal{U}} a_{u,m} p_{u,m} - P_m^g \right) \\
& + \sum_{m \in \mathcal{M}} \lambda_m \left[B_m - \sum_{u \in \mathcal{U}} a_{u,m} W_0 \log_2(1 + \gamma_{u,m}) \right] \\
& + \sum_{m \in \mathcal{M}} \beta_m \left(P_m^{max} - \sum_{u \in \mathcal{U}} a_{u,m} p_{u,m} \right), \\
& + \sum_{u \in \mathcal{U}} \nu_u \left[\sum_{m \in \mathcal{M}} a_{u,m} W_0 \log_2(1 + \gamma_{u,m}) - \delta_u \right]
\end{aligned}
\tag{3.45}
$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_M)$ is the Lagrange multiplier vector corresponding to the fronthaul constraints (3.40). $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_M)$ is the Lagrange multiplier vector concerned with the transmit power constraint. $\boldsymbol{\nu} = (\nu_1, \nu_2, ..., \nu_U)$ is the vector corresponding to the minimum data rate constraint for each user equipment.

By going one step further into the Lagrangian dual analysis, the dual

function of the above Lagrangian expression is:

$$
\begin{aligned}
& g(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\nu}) \\
& = \max_{\{\mathbf{a},\mathbf{p}\}} \quad L(\mathbf{a}, \mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\nu}) \\
& = \max_{\{\mathbf{a},\mathbf{p}\}} \quad \Bigg\{ \sum_{u\in\mathcal{U}} \sum_{m\in\mathcal{M}} \Bigg[(1 - \lambda_m + \nu_u) a_{u,m} W_0 log_2(1 + \gamma_{u,m}) \\
& - (\varphi^{(i)} \kappa_m + \beta_m) a_{u,m} p_{u,m} \Bigg] + \sum_{m\in\mathcal{M}} \Bigg[\lambda_m B_m + \beta_m P_m^{max} \\
& - \varphi^{(i)} (P_{m,0} - P_m^g) \Bigg] - \sum_{u\in\mathcal{U}} \nu_u \delta_u \Bigg\}
\end{aligned}
\tag{3.46}
$$

Then by using dual decomposition method, the dual problem is decomposed into K independent sub-problems as

$$
\begin{aligned}
& g(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\nu}) \\
& = \sum_{m\in\mathcal{M}} g_m(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\nu}) \\
& + \sum_{m\in\mathcal{M}} \Bigg[\lambda_m B_m + \beta_m P_m^{max} - \varphi^{(i)} (P_{m,0} - P_m^g) \Bigg] - \sum_{u\in\mathcal{U}} \nu_u \delta_u
\end{aligned}
\tag{3.47}
$$

where

$$
\begin{aligned}
& g_m(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\nu}) \\
& = \max_{\{\mathbf{a},\mathbf{p}\}} \quad \Bigg\{ \sum_{u\in\mathcal{U}} \Bigg[(1 - \lambda_m + \nu_u) a_{u,m} W_0 log_2(1 + \gamma_{u,m}) \\
& - (\varphi^{(i)} \kappa_m + \beta_m) a_{u,m} p_{u,m} \Bigg] \Bigg\}
\end{aligned}
\tag{3.48}
$$

Assuming the u-th UE is associated with m-th AP, it is obvious that (3.47) is concave in terms of $p_{u,m}$. Then applying Karush-Kuhn-Tucker condition,

the expression of optimal power allocation can be obtained:

$$p^*_{u,m} = [\omega^*_{u,m} - \frac{I_{u,m} + \sigma^2}{g_{u,m}}]^+ \tag{3.49}$$

and the optimal water-filling level $\omega^*_{u,m}$ is derived as:

$$\omega^*_{u,m} = \frac{(1 - \lambda_m + \nu_u)W_0}{ln2(\varphi^{(i)}\kappa_m + \beta_m)} \tag{3.50}$$

By substituting the derived optimal power allocation into the decomposed problem (3.47), it gives the following expression:

$$\begin{aligned} & g_m(\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\nu}) \\ & = \max_{1 \le u \le U} \quad \left\{ (1 - \lambda_m + \nu_u) W_0 \left[log_2(\omega^*_{u,m} \frac{g_{u,m}}{I_{u,m} + \sigma^2}) \right]^+ \right. \\ & \left. - (\varphi^{(i)}\kappa_m + \beta_m) \left[\omega^*_{u,m} - \frac{I_{u,m} + \sigma^2}{g_{u,m}} \right]^+ \right\} \end{aligned} \tag{3.51}$$

Then with (3.49) and (3.50), the optimal user association indicator is derived as:

$$a^*_{u,m} = \begin{cases} 1, & u = arg \max\limits_{1 \le u \le U} D_{u,m}, \\ 0, & otherwise, \end{cases} \tag{3.52}$$

where

$$\begin{aligned} D_{u,m} = & (1 - \lambda_m + \nu_u) W_0 \left[log_2(\omega^*_{u,m} \frac{g_{u,m}}{I_{u,m} + \sigma^2}) \right]^+ \\ & - \frac{(1 - \lambda_m + \nu_u)}{ln2} \left[1 - \frac{I_{u,m} + \sigma^2}{g_{u,m}\omega^*_{u,m}} \right]^+ \end{aligned} \tag{3.53}$$

The sub-gradient based method is used to update the dual variables λ_m, β_m and ν_u to generate the power allocation and user association policies. The update equations for the dual variables in the $(l+1)$-th iteration are given by

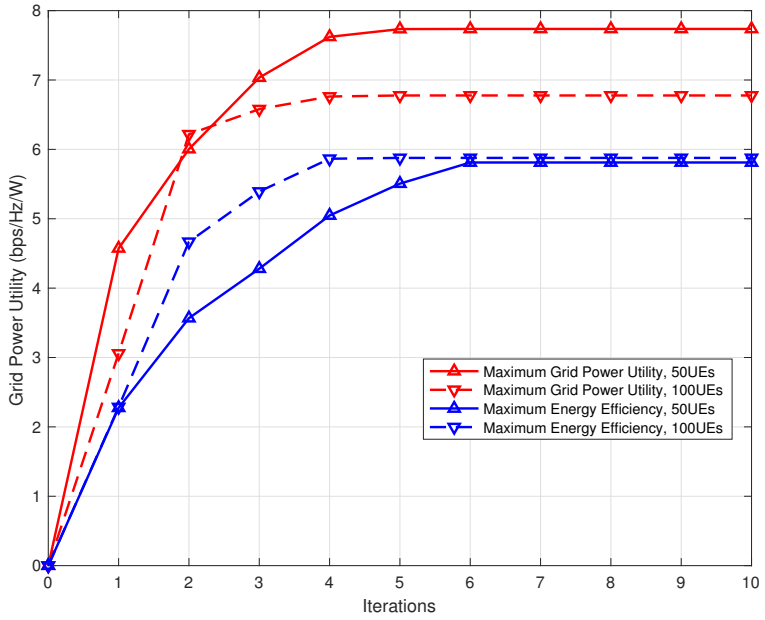


Figure 3.5: Convergence performance of the grid power utility.

$$\lambda_m^{(l+1)} = [\lambda_m^{(l)} - \alpha_\lambda^{(l+1)}(B_m - \sum_{u\in\mathcal{U}} R_{u,m}^{(l)})]^+, \quad \forall m \in \mathcal{M}, \tag{3.54}$$

$$\beta_m^{(l+1)} = [\beta_m^{(l)} - \alpha_\beta^{(l+1)}(P_m^{max} - \sum_{u\in\mathcal{U}} a_{u,m}p_{u,m})]^+, \forall m \in \mathcal{M}, \tag{3.55}$$

$$\nu_u^{(l+1)} = [\nu_u^{(l)} - \alpha_\nu^{(l+1)}(\sum_{m\in\mathcal{M}} R_{u,m}^{(l)} - \delta_u)]^+, \quad \forall u \in \mathcal{U}, \tag{3.56}$$

where $R_{u,m}^{(l)} = a_{u,m}^{(l)} W_0 log_2(1 + \frac{p_{u,m}^{(l)} g_{u,m}}{I_{u,m} + \sigma^2})$, and $\alpha_\lambda^{(l+1)}$, $\alpha_\beta^{(l+1)}$, $\alpha_\nu^{(l+1)}$ are the positive step sizes.

3.3.4 Simulation Results and Conclusions

In this section, the numerical results for the system model and algorithm presented in previous sections are presented. There is one HPN in the cell center and 12 EH-RRHs uniformly distributed around it. Each resource block occupies a bandwidth of 180kHz, and one resource block is allocated to one UE at a time. The UEs are randomly distributed inside the macrocell. The simulation of RB and power allocation is repeated based on the time slot of 1ms, which is the minimum time unit in OFDMA systems. During each time slot, the energy harvesting rates remain constant. According to [AGD+11], the power amplifier efficiencies of macrocell HPN and picocell RRHs are $\eta = 31.1\%$ and $\eta_R = 6.7\%$ respectively. It is assumed that the static circuit power consumption of HPN to be $P_c^M = 10.3W$ and that of picocell RRH is 0.1W. As for the fronthaul link and the backhaul link power consumption P_{bh}^M and P_{fh}, they are both assumed to be 0.2W [PZJ+15]. The maximum transmit power of HPN is 20W (43dBm). Energy harvesting rate P_k^g is assumed to be undergoing a stochastic process [XCE+16]. The lowest and highest power harvesting level are 22dBm and 48dBm respectively. The thermal noise level is -174dBm/Hz. With 20MHz as the overall system bandwidth, there are 100 resource blocks available to be assigned to the UEs.

Figure 3.5 demonstrates the convergence performance of the proposed algorithm. The convergence of the baseline algorithm – Maximum Energy Efficiency which is similar with that of [PZJ+15] is presented as well for the comparison. The curves representing the convergence of 50 and 100 UEs in the system for the two algorithms are presented respectively. In the first few iterations, the algorithm improves the grid power utility approaching the convergence level. As shown in the Figure 3.5, with different scenarios and

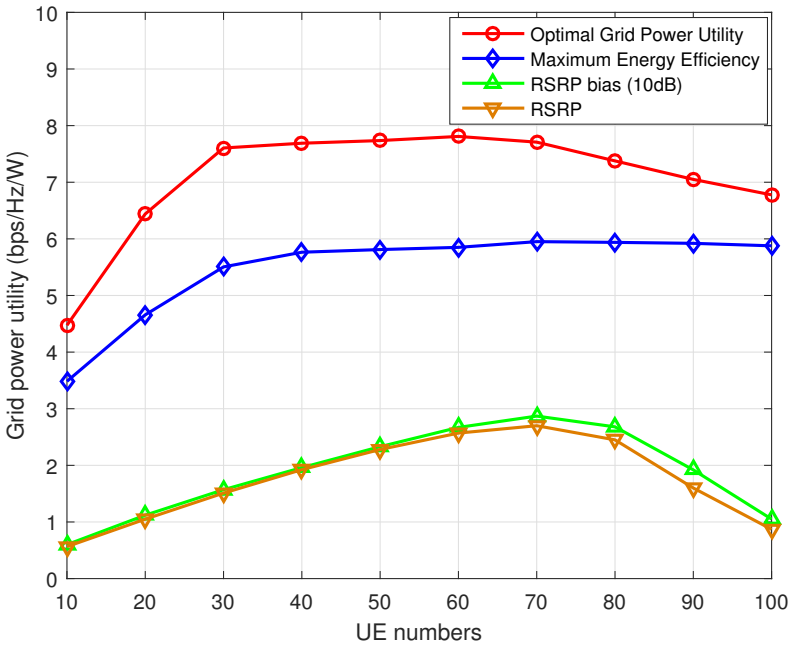


Figure 3.6: Grid power utility versus different numbers of UEs.

UE numbers, they have different tempos towards the convergence value, but all converge after five updates of the outer loop iteration. The change of optimality target from energy efficiency to grid power utility do not sacrifice the convergence speed of such iterative algorithm. For the proposed algorithm, system with 50 UEs outperforms the system with 100 UEs in terms of grid power utility. As for the maximum energy efficiency algorithm, although system with 50 UEs has slightly better utility performance, they almost have the same value. Then it turns to the detailed grid power utility performance versus different number of UEs.

The grid power utilities versus UE numbers for different scenarios are depicted in the Figure 3.6. The curve as the highest level represents the obtained utility of proposed algorithm. The value goes an apparent rise as

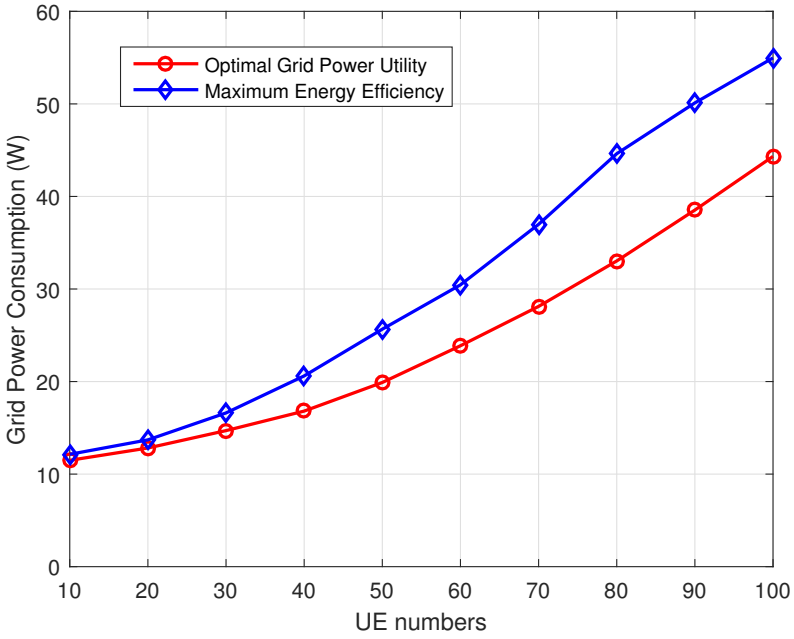


Figure 3.7: Grid power consumption for different scenarios.

the number of UEs increases from 10 to 30. The reason behind is that before the number of connected users reaches 30, there is much harvested renewable remained unused; then as more and more users connected to the system, that part of renewable which is not utilised would be allocated to the newly connected UEs. Under such condition, data rate requirement can be satisfied without consuming much more grid power. Then the value remains at a relatively stable level for UE numbers from 30 to 70, which means that the utility optimality algorithm reassigns the radio resource according to the new system characteristics to satisfy the service requirement without degrading the utility. After that, the utility decreases because much higher grid energy consumption is needed to overcome the increased interference level.

Another three scenarios are added as the benchmarks. The blue line is

the maximum energy efficiency algorithm. After the go-up for the UE numbers from 10 to 30, the curve reaches the stable level, and the position of this curve is completely under that of the proposed algorithm. The lines of RSRP and biased RSRP occupy the lowest positions in the plot. In addition, both two curves show a straight up-and-down with a peak and without a stability level. That is because as a simple algorithm without considering other system parameters except for the receiving power of the UEs, RSRP does not collect system information for centralised control. Under such scenario, the system just allocates the remained green energy to the newly added UEs until it is exhausted. That in turns provides the proof of the superiority of the advanced centralised control. In summary, the proposed algorithm gives the best performance both in terms of the grid power utility and the efficiency of resource allocation.

Figure 3.7 compares the grid power consumption for two different scenarios. The curve representing proposed algorithm is always under that of the baseline energy efficiency maximisation algorithm, meaning that proposed algorithm always consumes less grid power for whatever the number of user equipments. The amount of grid power saved by applying our algorithm may not be very large for UE numbers from 10 to 30, but as the number of user equipments gets larger, the grid power saving becomes more obvious. There are almost 20 to 30 percent of grid power saving compared to baseline scenario for UE numbers larger than 30, which is quite considerable when it comes to the absolute value for hundreds of macrocells adding together.

The highest energy harvesting level for renewable energy source is also an important factor influencing the grid power utility performance. This

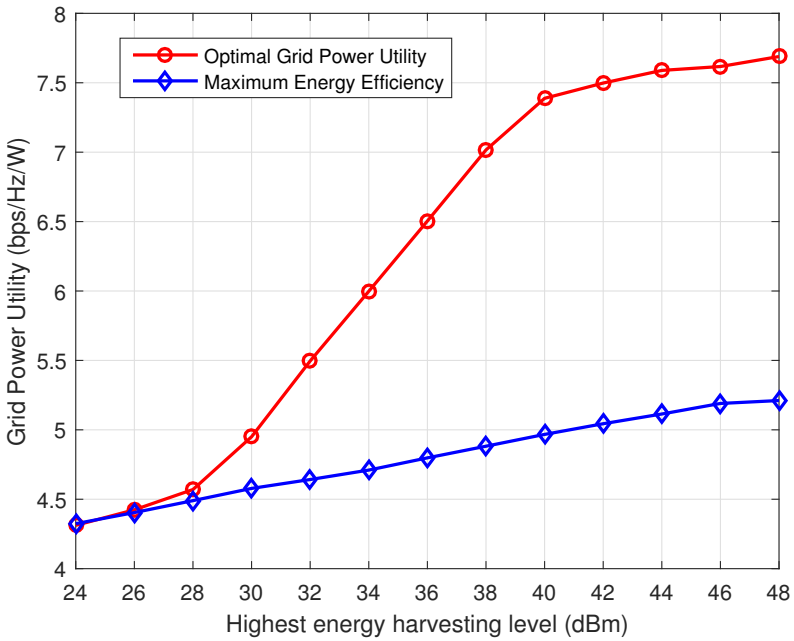


Figure 3.8: Grid power utility comparisons under different upper bounds of the energy harvesting rate.

parameter itself is decided by the condition of ambient environment. Figure 3.8 compares the utility as the maximum energy harvesting rate ranges from 24dBm to 46dBm. At the first 3 points, the grid power utility increases slowly, because most of the harvested renewable energy can just compensate the static power consumption of the APs. This restricts the growth of the green power supported service rate. Then there goes a rapid increase for range from 28dBm to 40dBm. As most of the static power consumption has already been satisfied by the energy harvester, excess green energy is able to be allocated accordingly to the served UEs, which makes the throughput rise up without consuming extra grid power. After 40dBm, the increase slows down. That is because the energy harvesting rate excesses the maximum allowed transmitting power for most of the APs most of the time, only a portion of the renewable will be used for signal transmission. Compared with

the grid power utility maximisation algorithm, energy efficiency optimisation algorithm undergoes a linear increase. The reason is that EE optimisation obtains the optimal resource allocation without considering the green power supply, so the grid power utility just increases as the portion of green energy usage goes higher.

3.4 Summary

In this chapter, the heterogeneous cloud radio access network with hybrid energy sources is investigated. Firstly, the joint user association and power allocation scheme for hybrid energy source powered H-CRAN is proposed. The optimisation problem is formulated as on-grid energy minimisation problem. Applying Lagrange dual decomposition method, the user association and power allocation policy is obtained, generating the numerical results of grid power consumption by macrocell high power node. Numerical results reveal that the proposed algorithm is able to introduce considerably large amount of on-grid energy saving comparing to baseline algorithms. The power saving performance is improved by the optimal algorithm comparing to traditional RSRP schemes. Afterwards, an iterative optimisation algorithm is proposed to maximise the grid power utility. Numerical results demonstrate that the grid power utility of the proposed optimisation scheme outperforms all other baseline schemes to a large extent. In terms of the grid power saving performance, 20 to 30 percent of grid power consumption can be reduced compared to baseline algorithm.

Chapter 4

Delay-aware Energy Efficient Computation Offloading for Energy Harvesting Enabled Fog Radio Access Networks

This chapter introduces a novel computation offloading strategy in the fog-computing-based radio access networks architectures, where all access points are equipped with renewable energy sources. Theoretical analysis is presented to illustrate how to coordinates the computation offloading according to the availabilities of renewable energy to minimise the grid power consumption.

4.1 Motivation

Although there are existing literatures on resource allocation strategy of F-RANs' offloading, there is no existing work on how to coordinate the computation offloading according to the availabilities of renewable energy to minimise the grid power consumption. In addition, many researches on F-RAN and mobile edge computing do not fully reflect the characteristics of fog-computing by examining the computation offloading strategies with multiple choice of offloading paths. In this chapter, a energy efficient computation

offloading design for energy harvesting enabled F-RAN is presented, which is to fill in the above-mentioned research gaps.

4.2 System Model

This chapter focuses on a multi-user scenarios where the set $\mathcal{U}$ of U mobile user equipments (UEs) are served by F-APs. The F-APs are denoted as set $\mathcal{N} = \{1, ..., N\}$. For specific F-AP $n \in \mathcal{N}$, $\mathcal{U}_n$ represents the set of UEs associated with it, while $\mathcal{U}_n \cap \mathcal{U}_{n'} = \emptyset, \forall n, n' \in \mathcal{N}, n \neq n'$ and $\bigcup_{n \in \mathcal{N}} \mathcal{U}_n = \mathcal{U}$. $\mathcal{N}_u$ is defined as the serving small cell of u. The service request for each UE is represented as $(D_u^{in}, D_u^{out}, L_u)$, where D_u^{in} is the number of input data bits for computation, D_u^{out} is the output bits of the computation results returned, and L_u denotes how many instructions are needed to complete the computation. In theory, the computational load L_u can be offloaded to the cloud center instead of going to the neighbouring F-APs. However, this contradicts the intention of alleviating the traffic burden and eliminating transmission delay of the fronthaul network. Moreover, the high level of static power consumption in the cloud center hinders the applicability of renewable power supply.

4.2.1 Transmission Model

For all F-APs $n \in \mathcal{N}$, $a_{nk}^u = 1$ if they decides to offload the computation task of $u \in \mathcal{U}_n$ to the neighbouring F-AP $k \in \mathcal{N}/\mathcal{N}_u$, and $a_{nk}^u = 0$ otherwise. Since there would be excessive signalling overhead if the offloading is implemented among multiple F-APs, it is assumed that if $a_{nk}^u = 1$, then $\sum_{k \in \mathcal{N}/\mathcal{N}_u} \sum_{u \in \mathcal{U}_n} a_{nk}^u = 1, \forall n \in \mathcal{N}$ since the computation for specific u in its serving F-AP $n \in \mathcal{N}_u$ can only be offloaded to one of the neighbouring

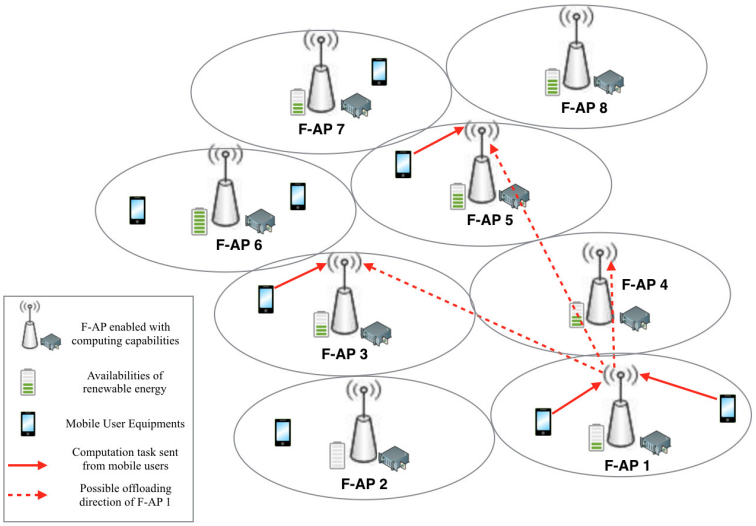


Figure 4.1: System diagram for F-RAN offloading.

F-APs.

Given the channel between two access points $\mathbf{h}_{nk}$,the channel-to-interference-plus-noise ratio (CINR) is expressed as:

$$\omega_{nk} = \frac{|\mathbf{h}_{nk}|^2}{N_0}, \tag{4.1}$$

where N_0 is the noise power spectral density. By assuming that effective resource reuse scheme is applied, the interference from other transmission can be ignored [PZJ+15]. Then the achievable data transmission rate for the n-th F-AP to transmit D_u to neighbouring F-AP k can be given as

$$R^u_{nk} = B_0 log_2(1 + p^u_{nk}\omega_{nk}), \tag{4.2}$$

where B_0 is the bandwidth occupied for data transmission, p^u_{nk} is the transmit power level. Hence, the latency introduced by the data transmit from serving

F-AP to the neighbouring access point can be expressed as

$$T_{nk}^{u} = \frac{D_{u}^{in}}{B_0 log_2(1 + p_{nk}^{u}\omega_{nk})}. \tag{4.3}$$

Similarly, given the transmit power for returning the computation results as $\widehat{p}_{kn}^{u}$ and the CINR $\widehat{\omega}_{kn}$, the transmit delay for returning computation result is given as:

$$\widehat{T}_{kn}^{u} = \frac{D_{u}^{out}}{B_0 log_2(1 + \widehat{p}_{kn}^{u}\widehat{\omega}_{kn})}, \tag{4.4}$$

and the total transmission latency be written as

$$T_{u}^{T} = \sum_{n\in\mathcal{N}_u} \sum_{k\in\mathcal{N}/\mathcal{N}_u} a_{nk}^{u} \Bigg(T_{nk}^{u} + \widehat{T}_{kn}^{u} \Bigg). \tag{4.5}$$

Based on the transmit power and the latency, the energy consumption for the wireless transmission of forward offloading is represented by $E_{nk}^{u} = \frac{1}{\eta_0} p_{nk}^{u} T_{nk}^{u}$, and for returning computation results $\widehat{E}_{kn}^{u} = \frac{1}{\eta_0} \widehat{p}_{kn}^{u} \widehat{T}_{kn}^{u}$ where η_0 is the power amplifier efficiency of F-APs. Then the total transmit energy consumed by n-th F-AP is given as:

$$E_{n}^{T} = \sum_{k\in\mathcal{N}/\mathcal{N}_u} \Bigg(\sum_{u\in\mathcal{U}_n} a_{nk}^{u} E_{nk}^{u} + \sum_{u\in\mathcal{U}/\mathcal{U}_n} a_{kn}^{u} \widehat{E}_{kn}^{u} \Bigg), \tag{4.6}$$

and the total transmit power as:

$$P_{n}^{T} = \sum_{k\in\mathcal{N}/\mathcal{N}_u} \Bigg(\sum_{u\in\mathcal{U}_n} a_{nk}^{u} p_{nk}^{u} + \sum_{u\in\mathcal{U}/\mathcal{U}_n} a_{kn}^{u} \widehat{p}_{kn}^{u} \Bigg). \tag{4.7}$$

4.2.2 Computation Model

It is assumed that the overall computation capability, which is characterised as the CPU cycles per second, is fixed for each access point denoted by

F_n^{max}. Therefore, by scheduling f_{un} as the computation capability for the u-th UE's service requests, the constraint should be satisfied as the sum capability allocated should not excess F_n^{max}. Denoting m_u as the number of CPU cycles required for each instruction, the latency of completing the computing task for u-th UE is

$$\begin{aligned} T_u^C &= \sum_{n\in\mathcal{N}_u}\left[(1-\sum_{k\in\mathcal{N}/\mathcal{N}_u} a_{nk}^u)\frac{m_u L_u}{f_{un}} + \sum_{k\in\mathcal{N}/\mathcal{N}_u} a_{nk}^u \frac{m_u L_u}{f_{uk}}\right] \\ &= \sum_{n\in\mathcal{N}_u}\left[\sum_{k\in\mathcal{N}/\mathcal{N}_u} a_{nk}^u(\frac{m_u L_u}{f_{uk}} - \frac{m_u L_u}{f_{un}}) + \frac{m_u L_u}{f_{un}}\right] \end{aligned} \tag{4.8}$$

According to [WZL12], the energy consumption per CPU cycle is proportional to the square of the clock frequency of the chip. Then denoting $\varepsilon_n = \kappa_0(F_n)^2$ as the energy consumption per CPU cycle of the n-th F-AP, with

$$F_n = \sum_{k\in\mathcal{N}/\mathcal{N}_u}\left[\sum_{u\in\mathcal{U}_n}(1-a_{nk}^u)f_{un} + \sum_{u\in\mathcal{U}/\mathcal{U}_n} a_{kn}^u f_{un}\right] \tag{4.9}$$

Hence, given the total computation load of n-th fog-computing access point as:

$$C_n = \sum_{k\in\mathcal{N}/\mathcal{N}_u}\left[\sum_{u\in\mathcal{U}_n}(1-a_{nk}^u)L_u + \sum_{u\in\mathcal{U}/\mathcal{U}_n} a_{kn}^u L_u\right] \tag{4.10}$$

the total energy consumption for completing the computation in n-th access point can be written as:

$$E_n^C = \varepsilon_n C_n = \kappa_0(F_n)^2 C_n, \forall n \in \mathcal{N} \tag{4.11}$$

where $\kappa_0 = 10^{-8}$ according to realistic measurement [Che15].

4.2.3 Energy Harvesting Model

It is assumed that the energy harvested by green energy source cannot be stored in the access points because of the disadvantage of doing so [GLN+12]. Energy collected from ambient environment should be utilised in time, otherwise it will dissipated after next time slot. This means that the green energy should be fully utilised to serve the computation and offloading transmission in a short time slot. The available renewable energy for F-AP n is denoted as E_n^g, which is i.i.d between 0 and E_H^{max} uniformly for each access point. The energy consumed from the power grid is:

$$G_n = [E_n^T + E_n^C - E_n^g]^+, \forall n \in \mathcal{N} \tag{4.12}$$

where $[x]^+ = max\{x, 0\}$.

4.3 Problem Formulation

4.3.1 Original Problem

The grid energy consumption minimisation problem is formulated in this section. The delay requirement of mobile UEs should not excess the maximum allowed delay T_{max}. The total latency including the transmission delay and computation execution delay is

$$t_u = T_u^T + T_u^C \tag{4.13}$$

The maximum overall latency for finishing the computation is imposed as T_{max}. Given the computational capabilities for each F-AP F_n^{max}, and the

maximum transmission power level P_n^{max}, the network grid energy minimisation problem is formulated as:

$$\min_{\{\mathbf{a},\mathbf{p},\widehat{\mathbf{p}},\mathbf{f}\}} \quad \sum_{n\in\mathcal{N}} G_n \tag{4.14}$$

s.t.

$$P_n^T \leq P_n^{max}, \forall n \in \mathcal{N} \tag{4.15}$$

$$t_u \leq T_{max}, \forall u \in \mathcal{U} \tag{4.16}$$

$$F_n \leq F_n^{max}, \forall n \in \mathcal{N} \tag{4.17}$$

where $\mathbf{a} \triangleq (a_{nk}^u)_{\forall n\in\mathcal{N}_u,\forall k\in\mathcal{N}/\mathcal{N}_u,\forall u\in\mathcal{U}}$, is defined as the computation offloading indicator, $a_{nk}^u = 1$ if the computation task of $u \in \mathcal{U}_n$ is offloaded to the neighbouring F-AP $k \in \mathcal{N}/\mathcal{N}_u$, and $a_{nk}^u = 0$, otherwise; $\mathbf{p} \triangleq (p_{nk}^u)_{\forall n\in\mathcal{N}_u,\forall k\in\mathcal{N}/\mathcal{N}_u,\forall u\in\mathcal{U}}$ is the forward offloading power allocation policies, $\widehat{\mathbf{p}} \triangleq (\widehat{p}_{kn}^u)_{\forall n\in\mathcal{N}_u,\forall k\in\mathcal{N}/\mathcal{N}_u,\forall u\in\mathcal{U}}$ is the backward computation result returning power allocation policies; $\mathbf{f} \triangleq (f_{un})_{\forall u,n}$ is the computation capability allocation policy.

Because of the non-convexity of constraint (4.16) and the objective function itself, and the binary offloading indicator $\mathbf{a} \triangleq (a_{nk}^u)_{\forall n\in\mathcal{N}_u,\forall k\in\mathcal{N}/\mathcal{N}_u,\forall u\in\mathcal{U}}$, the formulated optimisation problem falls into the category of mixed integer non-linear programming (MINLP) which is NP-hard [CSBC16]. The problem is reformulated into the convex form in the following subsection.

4.3.2 Convex Reformulation

This section addresses the non-convexity of the optimisation problem. Firstly, the non-convex term in the objective function is investigated. In the objective function, the term representing the total transmit energy, E_n^T, is non-convex in terms of p_{nk}^u and $\widehat{p}_{kn}^u$. However, from the rough calculation according to

eq. (11), the energy consumed for transmitting input and output bits of computation is less than one tenth of the corresponding computation. As a result, the non-convex term of E_n^T can be eliminated from the objective function with negligible impact on the solution obtained.

Secondly, the constraint (4.16) is transformed into its convex equivalence. By separating the forward offloading and backward computation result returning into two independent power allocation policies, the signalling and decision making process may get complicated and therefore degrade the system performance in practical applications. Hence, by assuming that $p_{nk}^u = \widehat{p}_{kn}^u, \forall u \in \mathcal{U}, \forall n \in \mathcal{N}_u, \forall k \in \mathcal{N}/\mathcal{N}_u$, the forward and backward transmit power can be determined together with equal level. Assuming that the channel condition between two access points remains unchanged within the service time for specific u, i.e. $\omega_{nk} = \widehat{\omega}_{kn}, \forall n \in \mathcal{N}_u, \forall k \in \mathcal{N}/\mathcal{N}_u$, the constraint (4.16) is rewritten as

$$\sum_{n\in\mathcal{N}_u}\sum_{k\in\mathcal{N}/\mathcal{N}_u} a_{nk}^u \Bigg[\frac{D_u^{in}+D_n^{out}}{B_0 log_2(1+p_{nk}^u\omega_{nk})} + (\frac{m_u L_u}{f_{uk}} - \frac{m_u L_u}{f_{un}})\Bigg] + \sum_{n\in\mathcal{N}_u}\frac{m_u L_u}{f_{un}} \leq T_{max}, \forall u \in \mathcal{U} \tag{4.18}$$

The latency constraint (4.18) is still non-convex, further transformation is required. Supposing that for specific u, if there exists the offloading from its serving F-AP $\mathcal{N}_u$ to one of the neighbouring access points $k \in \mathcal{N}/\mathcal{N}_n$, i.e. $\sum_{n\in\mathcal{N}_u}\sum_{k\in\mathcal{N}/\mathcal{N}_u} a_{nk}^u = 1$, the expression (4.18) is reduced to

$$\frac{(D_u^{in}+D_u^{out})}{B_0 log_2(1+p_{nk}^u\omega_{nk})} + \frac{m_u L_u}{f_{uk}} \leq T_{max} \tag{4.19}$$

otherwise, i.e. $\sum_{n\in\mathcal{N}_u}\sum_{k\in\mathcal{N}/\mathcal{N}_u} a^u_{nk} = 0$, the expression would be reduced to

$$\frac{m_u L_u}{f_{un}} \leq T_{max} \tag{4.20}$$

Under the feasibility condition $T_{max} f_{uk} > m_u L_u$ [OSSB15], the reduced delay constraint (4.19) can be rewritten as

$$\frac{(D^{in}_u + D^{out}_u) f_{uk}}{T_{max} f_{uk} - m_u L_u} - B_0 log_2(1 + p^u_{nk}\omega_{nk}) \leq 0 \tag{4.21}$$

which is convex in term of both f_{uk} and p^u_{nk}. Then the simplified latency constraint (4.16) can be replaced with the three following constraints equivalently

$$\begin{aligned}\sum_{n\in\mathcal{N}_u}\sum_{k\in\mathcal{N}/\mathcal{N}_u} a^u_{nk}\Bigg[&\frac{(D^{in}_u + D^{out}_u) f_{uk}}{T_{max} f_{uk} - m_u L_u} \\ &- B_0 log_2(1 + p^u_{nk}\omega_{nk})\Bigg] \leq 0\end{aligned} \tag{4.22}$$

$$(1 - \sum_{n\in\mathcal{N}_u}\sum_{k\in\mathcal{N}/\mathcal{N}_u} a^u_{nk})(m_u L_u - f_{un} T_{max}) \leq 0 \tag{4.23}$$

$$m_u L_u - T_{max} f_{uk} < 0, \forall u \in \mathcal{U}, k \in \mathcal{N}/\mathcal{N}_u \tag{4.24}$$

4.4 Offloading Decision Algorithm

Although the objective function and non-convex constraints have been transformed into convex form, the reformulated problem is still a mix-integer programming with binary variables. If the exhaustive search method is applied,

the computational complexity would increase exponentially as the network size grows, which hinders the applicability on the scenario of large scale. Here the reformulated optimisation is denoted as the function of offloading decision $\mathbf{a} \triangleq (a_{nk}^u)_{\forall n \in \mathcal{N}_u, \forall k \in \mathcal{N}/\mathcal{N}_u, \forall u \in \mathcal{U}}$ as

$$\begin{aligned} \mathcal{P}(\mathbf{a}) : \min_{\{\mathbf{p},\mathbf{f}\}} \quad & \sum_{n \in \mathcal{N}} G_n' = \sum_{n \in \mathcal{N}} [E_n^C - E_n^g]^+ \\ \text{s.t.} \quad & (4.15) \ \& \ (4.17), \ (4.22)\text{-}(4.24) \end{aligned} \tag{4.25}$$

The following offloading decision algorithm is proposed to determine the value of each $a_{nk}^u \in \mathbf{a}$. The algorithm goes iteratively to find the set of F-AP with unused green energy and find the specific F-AP with largest grid energy consumption. Then the offloading decision index is updated each iteration by trying to find the possible offloading which can reduce the total network grid power consumption.

At the beginning of the decision making process, the energy consumption for each F-AP under no-offloading scenario would be estimated based on the collected parameters, e.g. latency constraints and required CUP cycles. If all access points have sufficient renewable energy or all access points have no spare renewable energy, the algorithm stops since there is no meaning making offloading. Usually it will not be the case, so from the access point which consumes the highest level of non-renewable energy, the computation task is assumed to be offloaded from its serving UEs set in turn. By finding the offloading which makes the non-renewable energy consumption the lowest, the iteration updates the offloading indicator accordingly. The algorithms iterates until it is found that further iterations will not obtain lower non-renewable energy consumption.

Algorithm 2 Offloading Decision Algorithm

Calculate the grid energy consumption of each F-AP under the scenario of no offloading, find the access point set with surplus green energy $\mathcal{M}^{[0]} = \{n|G_n^{[0]} = 0\}$
if $\mathcal{M}^{[0]} = \emptyset$ or $\mathcal{M}^{[0]} = \mathcal{N}$ **then**
  go to End
end if
Initialisation $r = 0$, $\mathbf{a}^{[r]} = \mathbf{0}$
repeat
  Set $r = r + 1$, $i = 1$ and $j = 1$
  $\tilde{n}= arg\max_{n\in\mathcal{N}/\mathcal{M}} G_n^{[r-1]}$
  for $i \leq |\mathcal{U}_{\tilde{n}}|$ **do**
    for $j \leq |\mathcal{M}^{[r-1]}|$ **do**
      Update $\mathbf{a}^{[r]}$ by setting $a_{\tilde{n}\tilde{m}}^{\tilde{u}} = 1$, where $\tilde{u} = u_i \in \mathcal{U}_{\tilde{n}}$, $\tilde{m} = m_j \in \mathcal{M}^{[r-1]}$
      Solve the problem $\mathcal{P}(\mathbf{a}^{[r]})$
      if it is feasible **and** $\sum_{n\in\mathcal{N}} G_n^{[r]} < \sum_{n\in\mathcal{N}} G_n^{[r-1]}$ **then**
        Update $G_n^{[r]}$ for each $n \in \mathcal{N}$, $\mathcal{M}^{[r]} = \{n|G_n^{[r]} = 0\}$
        break
      else
        $\mathbf{a}^{[r]} = \mathbf{a}^{[r-1]}$, $G_n^{[r]} = G_n^{[r-1]}, \forall n \in \mathcal{N}$
      end if
    end for
  end for
until $\mathcal{M}^{[r]} = \emptyset$ or $\mathcal{M}^{[r]} = \mathcal{N}$ or $\sum_{n\in\mathcal{N}} G_n^{[r]} = \sum_{n\in\mathcal{N}} G_n^{[r-1]}$
End

4.5 Simulation Results and Performance Analysis

This section elaborates the choosing of the system parameters and presents the numerical results produced.

4.5.1 System Parameters

Referencing the given channel gain between two access points [CSBC16]:

$$\mathbf{h}_{nk} = 10^{-\frac{L(d_{nk})}{20}} \sqrt{\phi_{nk}\theta_{nk}}\mathbf{g}_{nk} \tag{4.26}$$

where $L(d_{nk}) = 15.3 + 37.6log_{10}d_{nk}$ is the pathloss between serving F-AP n and computing F-AP k as the function of distance $d_{nk} \in \mathbb{R}$ in meters. In addition, ϕ_{nk} represents the log-normal shadowing with the standard deviation as 8 dB, θ_{nk} is defined as the antenna gain with value 15dBi, and g_{nk} is used to denote the small scale fading coefficient.

It is assumed that the F-APs with coverage radius of 200m and mobile users are geographically distributed in the $600 \times 600m^2$ area, so that there are $N = 9$ F-APs in total and the number of mobile users varies from 10 to 50. Referencing the face recognition applications[Che15], the computation tasks are with input and output data size D_u^{in} and D_u^{out} as 420KB, and the operations required is 500 Mega operations. The CPU cycles required for each operation is $m_u = 2$. The computational capabilities for each F-AP is $F_{max} = 10$GHz. Maximum allowed transmit power of access points $P_n^{max} = 43dBm$.

4.5.2 Numerical Results

Figure 4.2 represents the average grid energy consumption for completing each UE's computation. It worth mentioning that as the allocated computational capacity for each F-AP increases, energy consumption of each CPU cycle gets higher. The "SINR-based Offloading" scheme is presented as a benchmark. The F-APs which lack of sufficient renewable energy will try to offload the computation to the access point with the best channel condition.

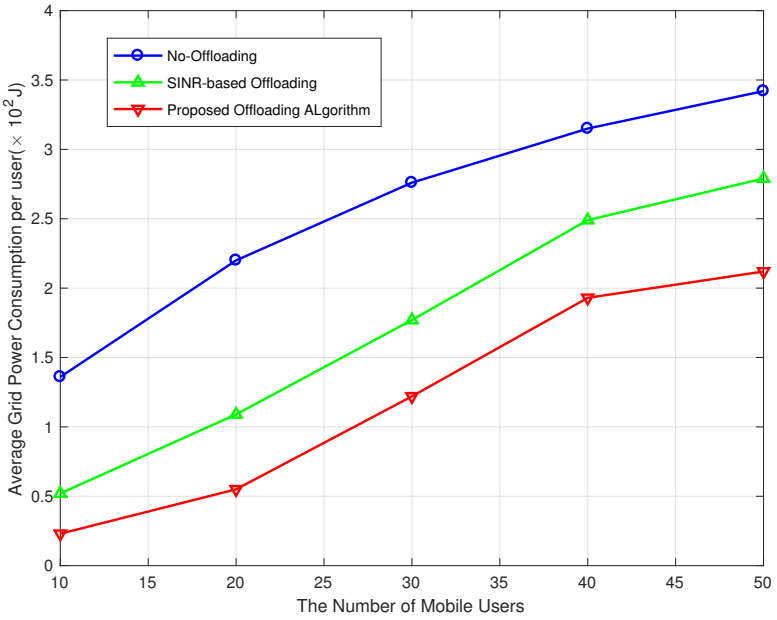


Figure 4.2: Average non-renewable energy consumption for each served UE in dependence on different user numbers, $T_{max} = 0.2s$

Compared to no-offloading scenario, this channel condition based offloading scheme does reduce the grid energy consumption. The proposed offloading scheme gives even lower grid energy consumption. That is because good channel condition is not directly related to the most important influencing factor, namely the availability of harvested energy. For specific offloading F-AP, the transmit of data can take more time if it offloads computation tasks to one of its neighbouring access points that are not with the best channel condition. However, the access point can be with higher availability of renewable energy, so that even the total energy consumption increases since the time window for completing computation is shorter, the less grid energy would consumed.

Figure 4.3 depicts the percentage of computation tasks completed by access points within the imposed latency constraints. This ratio is obtained

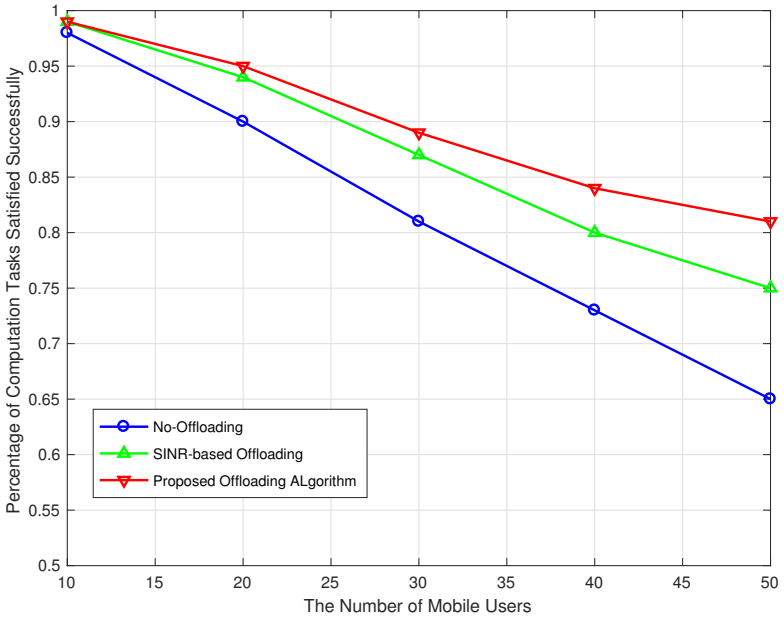


Figure 4.3: The percentage of computation tasks completed within imposed latency constraint in dependence on different user numbers, $T_{max} = 0.2s$

by removing the user from most task-intensive cell upon failing to reach a solution of the optimisation. The completing percentage are presented in dependence on the number of distributed mobile users from 10 to 50. As long as some offloading scheme is implemented, the satisfaction ratio would be improved compared to no offloading scheme. Still, the proposed offloading scheme gives better performance, namely higher percentage of satisfied computation, compared to channel condition based offloading. That is because of the richer choice of offloading paths for the proposed schemes.

Figure 4.4 represents how the latency for finishing computation tasks varies for different scenarios. The proposed offloading scheme takes more time to finish the computation than non-offloading scenario, because offloading takes time for data transmission between access points. For this performance indicator, "SINR-based offloading" gives better results than proposed

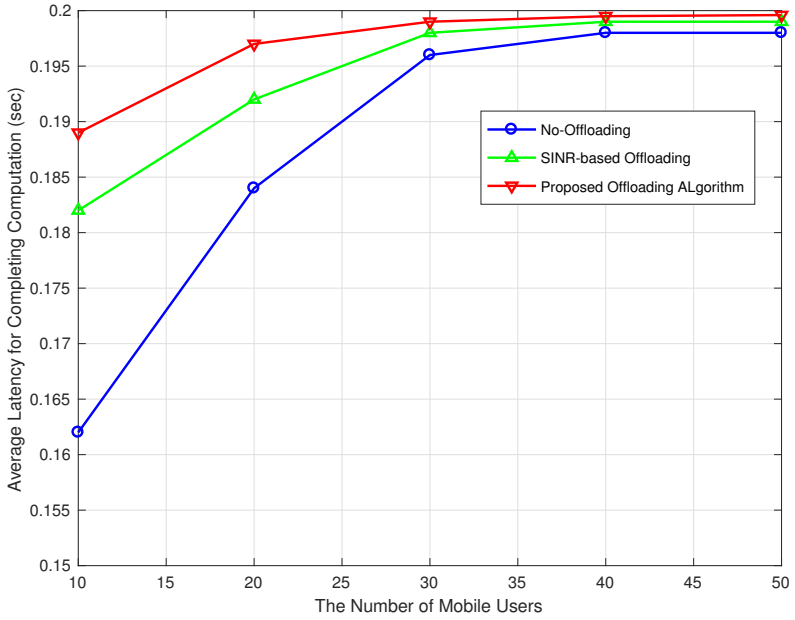


Figure 4.4: The average delay for completing computation in dependence on mobile user numbers, $T_{max} = 0.2s$

scheme. The reason behind is that channel condition based offloading takes less time for data exchange. However as the trade-off, this come at the cost of dropping more users' computation request. This also applies for the no offloading scenario, and this scenario can gives quicker average completion time but with the lowest percentage of computation task completion as show in Figure 4.3. It is possible for some APs with sufficient renewable energy to finish some mobile users' computation quicker than the imposed constraint, especially in low user numbers. As for the higher densities, average latencies of all scenarios tend to approach the imposed delay constraint.

4.6 Summary

In this chapter, a novel offloading design in energy harvesting enable F-RAN is proposed to achieve energy efficient computation offloading. A green energy aware offloading decision algorithm is proposed to determine the offloading strategy, then the power and computation capabilities allocations are obtained through solving the optimisation problem. The proposed offloading design has successfully minimised the grid power consumption with the help of renewable energy sources. Numerical results show that the offloading scheme can save the average grid power consumption for completing the computing task of each UE. Moreover, the percentage of computation request served by fog-computing also increases.

Chapter 5

Mobile-Edge Computation Offloading for Applications Featuring Shared Data

This chapter investigates the joint optimisation of computation offloading and communications resource allocation in the multi-user MEC system, where multiple single-antenna mobile users running applications featuring shared data. The detailed analysis on how the shared data property can be utilised to reduce mobile users' energy consumption is presented. Moreover, the performance evaluation is carried out to prove the effectiveness in energy saving.

5.1 System Model

A mobile-edge system is considered consisting of U mobile users running AR/VR applications, denoted as $\mathcal{U} = \{1, ..., U\}$, and one access point (AP) equipped with computing facilities working as a cloudlet server. All of the mobile users and the AP are assumed to be equipped with single antenna. The input data size for user u is denoted by D_u^I, $\forall u \in \mathcal{U}$, in which one fraction data size of D_S^I bits are the shared data that is the same across all U mobile users. The rest $D_u^I - D_S^I$ bits are held individually that are

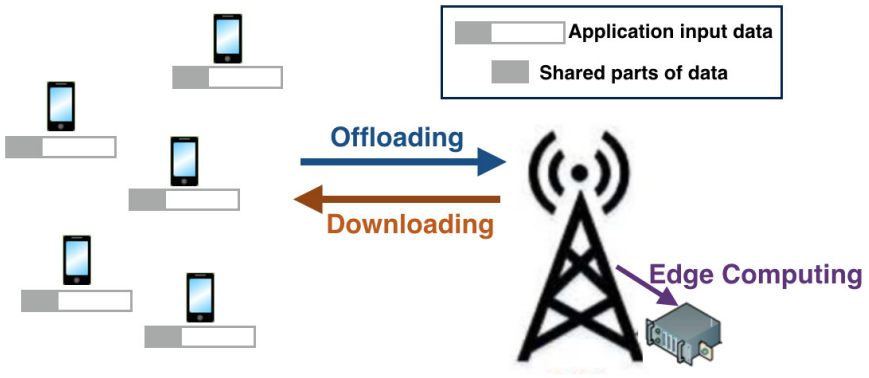


Figure 5.1: The system diagram of the investigated multi-user MEC system.

exclusive among different users. The computation data go through a partial offloading scheme, where both the shared data and the exclusively individual data are transmitted from each user partially. The shared data transmitted from each user is denoted by $D^I_{u,S}, \forall u \in \mathcal{U}$, such that $\sum_{u=1}^{U} D^I_{u,S} = D^I_S$. The amount of input data that is exclusively transmitted by u is thus given by $\bar{D}^I_u = D^I_u - D^I_S - D^L_u, \forall u \in \mathcal{U}$, where the D^L_u bits are remained for local execution by user u. For every bit of input data, there requires λ_0 CPU cycles to finish to computation. The proportion of the shared data in the input data bits is denoted as $\epsilon_u = D^I_S/D^I_u, \forall u \in \mathcal{U}$.

It can be seen from Figure 5.2 that there are two consecutive sub-phases for both input data offloading and the results downloading phases: the shared data transmission and the individual data transmission. The data are transmitted between mobile users and the cloudlet server via frequency division multiple access (FDMA). The uploading and downloading of all mobile users are conducted simultaneously occupying different frequency bands. The overall bandwidth are averagely allocated to all mobile users. The transmission duration for offloading the shared input data is denoted by $t^{ul}_{u,S}, \forall u \in \mathcal{U}$; the

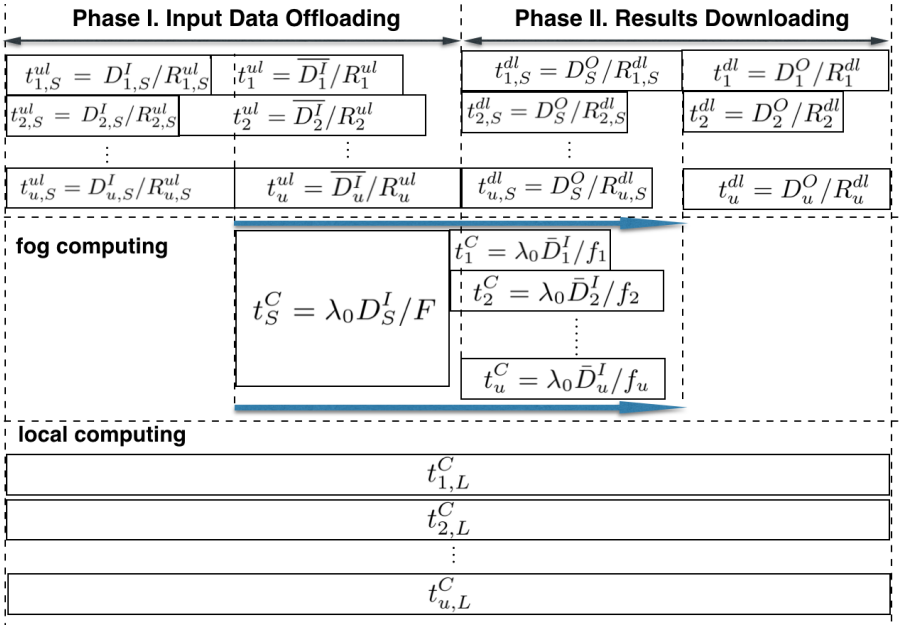


Figure 5.2: The data transmission protocol of the investigated multi-user MEC system.

offloading duration for the individual data is denoted as t^{ul}_u, $\forall u \in \mathcal{U}$; and the durations for downloading the shared and the individual output data are $t^{dl}_{u,S}$, t^{dl}_u, $\forall u \in \mathcal{U}$ respectively. The remote computation times are also illustrated in Figure 5.2, where t^C_S and t^C_u, $\forall u \in \mathcal{U}$ denote the computation of the shared input data and that of the offloaded individual data, respectively. Correspondingly, F and f_u, $\forall u \in \mathcal{U}$, denote the computational frequency (in *cycles/s*) allocated to process the shared and the individual tasks by the cloudlet server respectively. In addition, the local computation time is denoted by $t^C_{u,L}$, $\forall u \in \mathcal{U}$.

5.1.1 Task Data Transmission

As observed from the timing diagram Figure 5.2, there are two consecutive offloading transmission sub-phases: the shared data and the individual data offloading [ASS17]. Each mobile user occupying different frequency bands via

FDMA, and offloads their computation tasks to the cloudlet server simultaneously. The forward link channel coefficient of user u is given by $h_u, \forall u \in \mathcal{U}$, which is assumed to remain unchanged during the offloading transmission. With the transmission power given by $p^{ul}_{u,S}$, the achievable individual data rate for offloading the shared data is expressed as:

$$R^{ul}_{u,S} = W^{ul}_u log_2(1 + \frac{p^{ul}_{u,S}|h_u|^2}{N_0}), \forall u \in \mathcal{U}, \tag{5.1}$$

where $W^{ul}_u = \frac{W^{ul}}{U}$, W^{ul} denoting the overall bandwidth available for the input data offloading, U is the number of active mobile users. N_0 is the additive white Gaussian noise (AWGN) power. Accordingly, $t^{ul}_{u,S} = D^I_{u,S}/R^{ul}_{u,S}$, and the energy consumed by the u-th user in the shared data offloading sub-phase is given as

$$E^{ul}_{u,S} = t^{ul}_{u,S} p^{ul}_{u,S} = \frac{t^{ul}_{u,S}}{|h_u|^2} f(\frac{D^I_{u,S}}{t^{ul}_{u,S}}), \forall u \in \mathcal{U}, \tag{5.2}$$

where $f(x)$ is defined as $f(x) = N_0(2^{\frac{x}{W^{ul}_u}} - 1)$ [XLXN19].

Recalling the same derivation process, denoting the transmission power applied for individual data offloading as p^{ul}_u, the energy consumption for the u-th user in the individual data offloading sub-phase is expressed as:

$$E^{ul}_u = t^{ul}_u p^{ul}_u = \frac{t^{ul}_u}{|h_u|^2} f(\frac{D^I_u - D^I_S - D^L_u}{t^{ul}_u}), \forall u \in \mathcal{U}. \tag{5.3}$$

Similar to the offloading transmission, the downloading transmission is separated into two sub-phases: the shared output data and the individual results downloading. The shared output data are multicasted to the mobile users via the whole available frequency band W^{dl}. The backward link channel coefficient for user u is given by $g_u, \forall u \in \mathcal{U}$, which calculates the signal fading when transmitted from the cloudlet server to the mobile user. It is assumed

that $g_u, \forall u \in \mathcal{U}$ remain constant during the results downloading. Given the transmission time of the shared data computation results as $t^{dl}_{u,S}, \forall u \in \mathcal{U}$, the energy expended to return the shared data computation results to mobile user u is

$$E^{dl}_{u,S} = \frac{t^{dl}_{u,S}}{|g_u|^2}\Gamma(\frac{a_0 D^I_S}{t^{dl}_{u,S}}), \forall u \in \mathcal{U}, \tag{5.4}$$

where $\Gamma(x) = N_0(2^{\frac{x}{W^{dl}}} - 1)$ [XLXN19], W^{dl} is the overall available frequency bands available. a_0 is the coefficient representing the number of output bits for executing one bit of input data.

Similarly, the amount of energy expended for transmitting individual output results back to user u is:

$$E^{dl}_u = \frac{t^{dl}_u}{|g_u|^2} f(\frac{a_0(D^I_u - D^I_S - D^L_u)}{t^{dl}_u}), \forall u \in \mathcal{U}. \tag{5.5}$$

In the mobile users side, the energy consumption for receiving and decoding downloaded data from the cloudlet server is proportional to the length of downloading time. According to the model given by [ASS17], the energy consumed at the u-th mobile user for the whole downloading phase is written as:

$$E^{down}_u = (t^{dl}_{u,S} + t^{dl}_u)\rho^{dl}_u, \forall u \in \mathcal{U}, \tag{5.6}$$

where ρ^{dl}_u (in *Joules/second*) captures the energy expenditure per second.

5.1.2 Computation Model

The computation in the cloudlet server is divided into two different subphases. Firstly, after the D^I_S bits of shared data are received by the cloudlet server, it is processed by the allocated computing capability. In this subphase, the part of computing capability allocated by the cloudlet server is

denoted as f_s. Given that the CPU cycles needed to finish the computation of shared input-bits is $\lambda_0 D_S^I$, the computation latency induced by processing shared data is:

$$t_S^C = \lambda_0 D_S^I / f_s, \tag{5.7}$$

According to the energy model in [Che15], the energy expended per CPU cycle is proportional to the square of the clock frequency of the chip. Given the allocated computing frequency f_s, the energy consumption is given as:

$$E_S^C = \kappa_0 {f_s}^2 \lambda_0 D_S^I, \tag{5.8}$$

where κ_0 is the energy consumption capacitance coefficient.

After the completion of the shared data computation, the computing capability is released to handle the computation of exclusive data offloaded by each user individually. In this sub-phase, the computing capability is divided into different fractions f_u for tasks from different users u. The latency induced is then given as:

$$t_u^C = \lambda_0 \bar{D}_u^I / f_u, \forall u \in \mathcal{U}, \tag{5.9}$$

and the energy consumption for finishing the computation of u-th mobile user's individual data bits is

$$E_u^C = \kappa_0 {f_u}^2 \lambda_0 \bar{D}_u^I, \forall u \in \mathcal{U}. \tag{5.10}$$

Given the local computation time as $t_{u,L}^C, \forall u \in \mathcal{U}$, and the local computing bits D_u^L, then the local clock frequency of u-th mobile user is given as $f_u^L = \lambda_0 D_u^L / t_{u,L}^C$. Recalling the energy consumption model in [Che15], the energy expended per CPU cycle is proportional to the square of the clock frequency

$f_u^{L^2}$. Since the CPU cycles needed to complete the local computation is $\lambda_0 D_u^L$, the energy consumption for executing local computing in terms of computational time is expressed as:

$$E_u^L = \kappa_0 \frac{(\lambda_0 D_u^L)^3}{t_{u,L}^{C\ 2}}, \forall u \in \mathcal{U}, \tag{5.11}$$

5.1.3 Total Latency

For the purpose of introducing the expression of total latency, a point needs to be understood firstly that some phases of computing or transmitting can only start after the end of some other specific progresses. For example, the uploaded shared data in the cloudlet server cannot start being processed and computed until the transmission of it completes. Given that all mobile users upload the shared input data cooperatively taking the time length as $t_{u,S}^{ul}$, the induced latency until the end of shared input data computation is $t_S^{ul} + t_S^C$, where $t_S^{ul} = \max\{t_{u,S}^{ul}\}$. The individual offloaded data can only set out to be computed both after the completion of its transmission and the end of shared data computation which means the release of computing capability. Hence, denoting the latency from the beginning of the data offloading until the end of individual input data computation as $\tau_{1,u}$, it is expressed as

$$\tau_{1,u} = \max\{t_{u,S}^{ul} + t_u^{ul}, \max\{t_{u,S}^{ul} + t_S^C\}\} + t_u^C \tag{5.12}$$

Moreover, by looking at the output data downloading, the shared output data transmission begins after the end of all offloading transmission since in our design the multicasting needs to occupy the whole bandwidth. More importantly, the transmission is only possible by the time that the computation results are available. Denoting the latency from the beginning of the data

offloading until the end of shared data output multicasting as τ_2, it is given as:

$$\tau_2 = \max\left\{ \max_{u\in\mathcal{U}}\{t^{ul}_{u,S} + t^{ul}_{u}\}, \max_{u\in\mathcal{U}}\{t^{ul}_{u,S}\} + t^{C}_{S} \right\} + \max_{u\in\mathcal{U}}\{t^{dl}_{u,S}\} \tag{5.13}$$

Note that there is no subtitle u for the denotation of τ_2 since it is irrelevant with specific mobile user u.

Finally, it comes to the individual output data transmission. As the last phase of the data transmission, it should wait for the idle state of the transmission channels, which means that the shared output data multicasting needs to come to the end before the transmission of the individual output data. Similar to the shared output data transmission, another aspect should be met except for the idle channels, which is the availability of the computing results. As a result, the latency from the beginning of input data offloading to the end of individual output data downloading for each mobile user u is:

$$\tau_u = \max\{\tau_{1,u}, \tau_2\} + t^{dl}_{u} \tag{5.14}$$

Combining the above facts, the expanded total latency expression is finally given as follows:

$$\begin{aligned}\tau_u = \max\Bigg\{ & \max\{t^{ul}_{u,S} + t^{ul}_{u}, \max_{u\in\mathcal{U}}\{t^{ul}_{u,S}\} + t^{C}_{S}\} + t^{C}_{u}, \\ & \max\left\{ \max_{u\in\mathcal{U}}\{t^{ul}_{u,S}\} + t^{C}_{S}, \max_{u\in\mathcal{U}}\{t^{ul}_{u,S} + t^{ul}_{u}\} \right\} + \max_{u\in\mathcal{U}}\{t^{dl}_{u,S}\} \Bigg\} + t^{dl}_{u}, \forall u \in \mathcal{U}.\end{aligned} \tag{5.15}$$

5.2 Energy Efficient Computation Offloading and Communications Resources Allocation for Applications Featuring Shared Data

5.2.1 Motivation

The joint optimisation of computation offloading with communications resources (such as power, bandwidth, and rate) proves to improve the performance of fog computing by explicitly taking channel conditions and communications constraints into account. The intrinsic collaborative properties of the input data for computation offloading was investigated for augmented reality in [ASS17]. However the authors of [ASS17] fail to find the optimal solutions of the proposed energy minimisation problem. In order to provide the in-depth understanding of the shared-data featured offloading in MEC systems and fully reap the advantage of fog computing, this research is conducted.

5.2.2 Problem Formulation

The overall energy consumption at the mobile users consists of three parts: data offloading over the uplink (c.f. (5.2) and (5.3)), results retrieving (c.f. (5.6)), and local computing (c.f. (5.11)), which is thus given by

$$\begin{aligned} E_m = &\sum_{u\in\mathcal{U}} \kappa_0 \frac{(\lambda_0 D_u^L)^3}{{t_{u,L}^C}^2} + \sum_{u\in\mathcal{U}} \frac{t_{u,S}^{ul}}{|h_u|^2} f(\frac{D_{u,S}^I}{t_{u,S}^{ul}}) \\ &+ \sum_{u\in\mathcal{U}} \frac{t_u^{ul}}{|h_u|^2} f(\frac{D_u^I - D_S^I - D_u^L}{t_u^{ul}}) + \sum_{u\in\mathcal{U}} (t_{u,S}^{dl} + t_u^{dl})\rho_u^{dl}. \end{aligned} \tag{5.16}$$

The objective is to minimise the mobile users' energy consumption given

by E_m, subject to the computing latency constraints, the maximum local computing frequencies, and the total energy consumption on the individual data at the BS. Specifically, the optimisation problem is formulated as below:

$$
\begin{aligned}
(\mathrm{P1}): \min_{\{t^{ul}_{u,S}, t^{ul}_{u}, t^{C}_{u,L}, t^{dl}_{u}, D^{L}_{u}, D^{I}_{u,S}\}} \quad & E_m \\
\texttt{s.t.} \quad & \\
& \tau_u \leq T_{max}, \forall u \in \mathcal{U}, && \text{(5.17a)} \\
& \sum_{u \in \mathcal{U}} \frac{t^{dl}_{u}}{|g_u|^2} f\left(\frac{a_0(D^{I}_{u} - D^{I}_{S} - D^{L}_{u})}{t^{dl}_{u}}\right) \leq E_{max}, && \text{(5.17b)} \\
& 0 \leq t^{C}_{u,L} \leq T_{max}, \forall u \in \mathcal{U}, && \text{(5.17c)} \\
& \lambda_0 D^{L}_{u} \leq t^{C}_{u,L} f_{u,max}, && \text{(5.17d)} \\
& 0 \leq D^{L}_{u} \leq D^{I}_{u} - D^{I}_{S}, \forall u \in \mathcal{U}, && \text{(5.17e)} \\
& \sum_{u \in \mathcal{U}} D^{I}_{u,S} = D^{I}_{S}, D^{I}_{u,S} \geq 0, && \text{(5.17f)} \\
& t^{ul}_{u,S} \geq 0, t^{ul}_{u} \geq 0, t^{C}_{u,L} \geq 0, t^{dl}_{u} \geq 0, \forall u \in \mathcal{U}. && \text{(5.17g)}
\end{aligned}
$$

Constraint (5.17a) and (5.17c) gives the latency constraints that the time taken for accomplishing computing tasks cannot excess the maximum allowed length, both for offloading and local computing. (5.17b) tells that the available energy for downlink transmission of remote computing node should be lower than a maximum level. (5.17d) restricts the number of allowable local computing bits imposed by local computing capabilities. In addition, (5.17f) puts that adding all the shared data bits offloaded by all mobile users respectively, the value should be equal to the exact amount of shared bits existing in the same user group.

Although the latency expression (5.15) looks complex in its from, (5.17a)

is still a convex constraint. For the ease of exposition, it is assumed herein that the cloudlet executes the shared and the individual computing within the duration of the individual data offloading and the shared results downloading, respectively, i.e., $t_S^C \ll t_u^{ul}$, and $t_u^C \ll t_{u,S}^{dl}$, $\forall u \in \mathcal{U}$[1]. As a result, (5.17a) can be simplified as below:

$$\max\{t_{u,S}^{ul} + t_u^{ul}\} + t_S^{dl} + t_u^{dl} \leq T_{max}, \forall u \in \mathcal{U}. \tag{5.18}$$

by introducing the auxiliary variable t^{dl}, which satisfies $t_u^{dl} \leq t^{dl}, \forall u \in \mathcal{U}$, (5.18) reduces to

$$t_{u,S}^{ul} + t_u^{ul} \leq T_{max} - t_S^{dl} - t^{dl}, \forall u \in \mathcal{U}. \tag{5.19}$$

Notice that E_u^C's (c.f. (5.10)) is monotonically decreases with respect to the local computing time $t_{u,L}^C$ for each mobile user. To obtain the minimal energy consumption, it is obvious that $t_{u,L}^C = T_{max}, \forall u \in \mathcal{U}$. Then the optimisation problem to be solved is reformulated as:

$$(\text{P1}') : \min_{\{t_{u,S}^{ul}, t_u^{ul}, t_u^{dl}, t^{dl}, D_u^L, D_{u,S}^I\}} E_m \tag{5.20a}$$

$$\texttt{s.t.}$$

$$(5.17\text{b}) - (5.17\text{g}), (5.19). \tag{5.20b}$$

$$t_u^{dl} \leq t^{dl}, \forall u \in \mathcal{U}. \tag{5.20c}$$

[1]It is assumed herein that the computation capacities at the cloudlet is relatively much higher than those at the mobile users, and thus the computing time taken is much shorter than the data transmission time.

5.2.3 Joint offloading and communication resource allocation

Introducing dual variables $\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \nu$, the Lagrangian of problem $(P1')$ is presented as:

$$\begin{aligned}
&L(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \nu, t_{u,S}^{ul}, t_u^{ul}, t_u^{dl}, t^{dl}, D_u^L, D_{u,S}^I) = \\
&\sum_{u\in\mathcal{U}} \frac{t_{u,S}^{ul}}{|h_u|^2} f(\frac{D_{u,S}^I}{t_{u,S}^{ul}}) + \sum_{u\in\mathcal{U}} \frac{t_u^{ul}}{|h_u|^2} f(\frac{D_u^I - D_S^I - D_u^L}{t_u^{ul}}) \\
&+ \sum_{u\in\mathcal{U}} \kappa_0 \frac{(\lambda_0 D_u^L)^3}{{t_{u,L}^C}^2} + \sum_{u\in\mathcal{U}} (t_{u,S}^{dl} + t_u^{dl})\rho_u^{dl} + \sum_{u\in\mathcal{U}} \beta_u (t_{u,S}^{ul} \\
&+ t_u^{ul} - T_{max} + t_S^{dl} + t^{dl}) + \sum_{u\in\mathcal{U}} \omega_u (\lambda_0 D_u^L \\
&- t_{u,L}^C f_{u,max}) + \sum_{u\in\mathcal{U}} \sigma_u (t_u^{dl} - t^{dl}) \\
&+ \nu [\sum_{u\in\mathcal{U}} \frac{t_u^{dl}}{|g_u|^2} f(\frac{a_0(D_u^I - D_S^I - D_u^L)}{t_u^{dl}}) - E_{max}],
\end{aligned} \tag{5.21}$$

where $\boldsymbol{\beta} = \{\beta_1, ..., \beta_U\}$ are dual variables associated with the latency constraint (5.19), $\boldsymbol{\omega} = \{\omega_1, ..., \omega_U\}$ are associated with local computing bits constraint (5.17d)), $\boldsymbol{\sigma} = \{\sigma_1, ..., \sigma_U\}$ are connected with the constraint for auxiliary variable t^{dl}, and ν catches the downlink transmission energy constraint (5.17b). Hence, the Lagrangian dual function expressed as:

$$\begin{aligned}
&g(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \nu) \\
&= \min_{\{t_{u,S}^{ul}, t_u^{ul}, t_u^{dl}, t^{dl}, D_u^L, D_{u,S}^I\}} L(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \nu, t_{u,S}^{ul}, t_u^{ul}, t_u^{dl}, t^{dl}, \\
&D_u^L, D_{u,S}^I),
\end{aligned} \tag{5.22}$$

s.t. (5.17e)-(5.17g).

Consequently, the corresponding dual problem is formulated as:

$$\max_{\{\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\sigma},\nu\}} g(\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\sigma},\nu) \tag{5.23}$$

s.t.

$$\boldsymbol{\beta} \succeq 0, \boldsymbol{\omega} \succeq 0, \boldsymbol{\sigma} \succeq 0, \nu \geq 0.$$

Proposition 1. *Given a determined set of dual variables $\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\sigma},\nu$, the optimal solution to the Lagrangian dual problem* (5.23) *can be determined as follows.*

The optimal primal variables $t_{u,S}^{ul}$, t_u^{ul}, and t_u^{dl}, are given by

$$\hat{t}_{u,S}^{ul} = \frac{\hat{D}_{u,S}^I}{\frac{W_u^{ul}}{ln2}[W_0(\frac{1}{e}(\frac{\beta_u|h_u|^2}{N_0}-1))+1]}, \forall u \in \mathcal{U}. \tag{5.24}$$

$$\hat{t}_u^{ul} = \frac{D_u^I - D_S^I - \hat{D}_u^L}{\frac{W_u^{ul}}{ln2}[W_0(\frac{1}{e}(\frac{\beta_u|h_u|^2}{N_0}-1))+1]}, \forall u \in \mathcal{U}. \tag{5.25}$$

$$\hat{t}_u^{dl} = \frac{a_0(D_u^I - D_S^I - \hat{D}_u^L)}{\frac{W_u^{dl}}{a_0 ln2}[W_0(\frac{1}{e}(\frac{(\rho_u^{dl}+\sigma_u)|g_u|^2}{\nu N_0}-1))+1]}, \forall u \in \mathcal{U}. \tag{5.26}$$

where $W_0(x)$ is the principle branch of the Lambert W function defined as the solution for $W_0(x)e^{W_0(x)} = x$ [WXWC18], e is the base of the natural logarithm; the optimal auxiliary variable t^{dl} is given by:

$$\hat{t}^{dl} = \begin{cases} 0, & \sum_{u\in\mathcal{U}} \beta_u - \sum_{u\in\mathcal{U}} \sigma_u > 0, \\ T_{max} - t_S^{dl}, & otherwise; \end{cases} \tag{5.27}$$

and the optimal local computing data size is given by

$$\hat{D}_u^L = \min\left\{T_{max}\sqrt{\left[\frac{N_0 ln2}{3\kappa_0{\lambda_0}^3}\left(\frac{2^{\frac{\hat{r}_u^{ul}}{W_u^{ul}}}}{W_u^{ul}|h_u|^2}+\frac{\nu a_0\cdot 2^{\frac{a_0\hat{r}_u^{dl}}{W_u^{dl}}}}{W_u^{dl}|g_u|^2}\right)-\frac{\omega_u}{3\kappa_0\lambda_0^2}\right]^+},\right.$$
$$\left. D_u^I - D_S^I\right\}, \forall u\in\mathcal{U},$$

where $\hat{r}_u^{ul} = \frac{W_u^{ul}}{ln2}[W_0(\frac{1}{e}(\frac{\beta_u|h_u|^2}{N_0}-1))+1]$ *and* $\hat{r}_u^{dl} = \frac{W_u^{dl}}{a_0 ln2}[W_0(\frac{1}{e}(\frac{(\rho_u^{dl}+\sigma_u)|g_u|^2}{\nu N_0}-1))+1]$, $\forall u\in\mathcal{U}$.

In fact, on one hand, $\hat{r}_u^{ul}$'s and $\hat{r}_u^{dl}$'s can be interpreted as the optimum transmission rate for the shared/individual data offloading and the individual data downloading, respectively, given the dual variables. On the other hand, for each user u, the optimal transmission rate for the shared data is seen to be identical to that of the individual data over the uplink, given that the uplink channel gains remain unchanged during the whole offloading phase.

Next, to obtain the optimal offloading bits of the shared data for each user, i.e., $\hat{D}_{u,S}^I$, the following lemma is needed.

Lemma 1. *The optimal offloaded shared data for user u is expressed as,*

$$\hat{D}_{u,S}^I = \begin{cases} D_S^I, & \hat{u} = arg\min\limits_{1\le u\le U}\Delta_u, \\ 0, & otherwise, \end{cases} \tag{5.28}$$

where $\Delta_u = \frac{f(\hat{r}_{u,S}^{ul})}{\hat{r}_{u,S}^{ul}|h_u|^2}+\frac{\beta_u}{\hat{r}_{u,S}^{ul}}, \forall u\in\mathcal{U}$.

Proof. To obtain how the shared input data offloading $\hat{D}_{u,S}^I$ are distributed among users, it needs to examine the partial Lagrangian regarding $D_{u,S}^I$ and $t_{u,S}^{ul}$. Replacing the shared data offloading time $t_{u,S}^{ul}$ with $\frac{D_{u,S}^I}{\hat{r}_u^{ul}}$, the partial

Lagrangian is expressed as

$$\min_{\{D^I_{u,S}\}} \overline{L} = \sum_{u\in\mathcal{U}} [\frac{t^{ul}_{u,S}}{|h_u|^2} f(\frac{D^I_{u,S}}{t^{ul}_{u,S}}) + \beta_u t^{ul}_{u,S}]$$

$$= \sum_{u\in\mathcal{U}} [\frac{D^I_{u,S}}{\hat{r}^{ul}_{u,S}|h_u|^2} f(\hat{r}^{ul}_{u,S}) + \beta_u \frac{D^I_{u,S}}{\hat{r}^{ul}_{u,S}}] \tag{5.29}$$

$$= \sum_{u\in\mathcal{U}} \Delta_u \cdot D^I_{u,S}$$

s.t.

$$\sum_{u\in\mathcal{U}} D^I_{u,S} = D^I_S, D^I_{u,S} \geq 0, \forall u \in \mathcal{U}, \tag{5.30}$$

where it is defined that $\Delta_u = \frac{f(\hat{r}^{ul}_{u,S})}{\hat{r}^{ul}_{u,S}|h_u|^2} + \frac{\beta_u}{\hat{r}^{ul}_{u,S}}$ as a constant given the dual variable β_u's. As a result, the optimal solution to the above linear programming is easily obtained as shown in (5.46). □

Notable, it is easily observed from Lemma 2 that the shared data is optimally offloaded by one specific user instead of multiple ones.

Based on Proposition 1, the dual problem can thus be iteratively solved according to ellipsoid method (with constraints), the detail of which can be referred to [Boy]. The algorithm for solving (P1′) is summarised in Table 5.1.

5.2.4 Simulation Results and Performance Analysis

This section elaborates the set-up of the system parameters and presents the generated simulation results. Except for the local computing only scheme where users execute all the data bits locally, there are three other offloading schemes presented as baseline algorithms: 1) *Offloading without considering*

Table 5.1: Algorithm I for solving (P1′)

Require: $(\boldsymbol{\beta^{(0)}}, \boldsymbol{\omega^{(0)}}, \boldsymbol{\sigma^{(0)}}, \nu^{(0)})$

1: **repeat**

2: Solve (5.22) given $(\boldsymbol{\beta^{(i)}}, \boldsymbol{\omega^{(i)}}, \boldsymbol{\sigma^{(i)}}, \nu^{(i)})$ according to Proposition 1 and obtain $\{\hat{t}^{ul}_{u,S}, \hat{t}^{ul}_{u}, \hat{t}^{dl}_{u}, \hat{t}_{dl}, \hat{D}^{L}_{u}, \hat{D}^{I}_{u,S}\}$;

3: update the subgradient of $\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \nu$ respectively, i.e., $t^{ul}_{u,S} + t^{ul}_{u} - T_{max} + \max\limits_{u\in\mathcal{U}}\{t^{dl}_{u,S}\} + t^{dl}$, $\lambda_0 D^{L}_{u} - t^{C}_{u,L} f_{u,max}$, $t^{dl}_{u} - t^{dl}$, $\sum_{u\in\mathcal{U}} \frac{t^{dl}_{u}}{|g_u|^2} f(\frac{a_0(D^{I}_{u} - D^{I}_{S} - D^{L}_{u})}{t^{dl}_{u}}) - E_{max}$ in accordance with the ellipsoid method [Boy];

4: **until** the predefined accuracy threshold is satisfied.

Ensure: The optimal dual variables to the dual problem (5.23) $(\boldsymbol{\beta^*}, \boldsymbol{\omega^*}, \boldsymbol{\sigma^*}, \nu^*)$

5: Solve (5.22) again with $(\boldsymbol{\beta^*}, \boldsymbol{\omega^*}, \boldsymbol{\sigma^*}, \nu^*)$

Ensure: $\{t^{ul*}_{u,S}, t^{ul*}_{u}, t^{dl*}_{u}, t^{dl*}, D^{L*}_{u}, D^{*}_{u,S}\}$

the shared data: the collaborative properties are ignored, every user makes the offloading decision without coordination among other users; 2) *Full offloading only*: the shared data is taken into consideration, but the whole chunks of input data of every user are forced to be offloaded to the edge computing node, excluding the local computing capability from participating in the computation tasks; 3) *Offloading with equal time length*: taking the correlated data into consideration, the data offloading and downloading are performed for each user with equal time length, with optimal solutions obtained through CVX.

In the simulation, the bandwidth avaialble is assumed to be $W^{ul} = W^{dl}$=10MHz, the maximum downlink transmit power $P_{max} = 1W$, and the input data size $D^{I}_{u} = 10kbits$ for all users. The spectral density of the AWGN power is -169 dBm/Hz. The mobile energy expenditure per second in the downlink is ρ^{dl}_{u}=0.625 J/s [ASS17], the maximum local computing capability $f_{u,max} = 1G$Hz. In addition, $\lambda_0 = 1 \times 10^3$ cycle/bit, $a_0 = 1$, $\kappa_0 = 10^{-26}$. The pathloss model is $PL = 128.1 + 37.6log_{10}(d_u)$, where d_u represents the

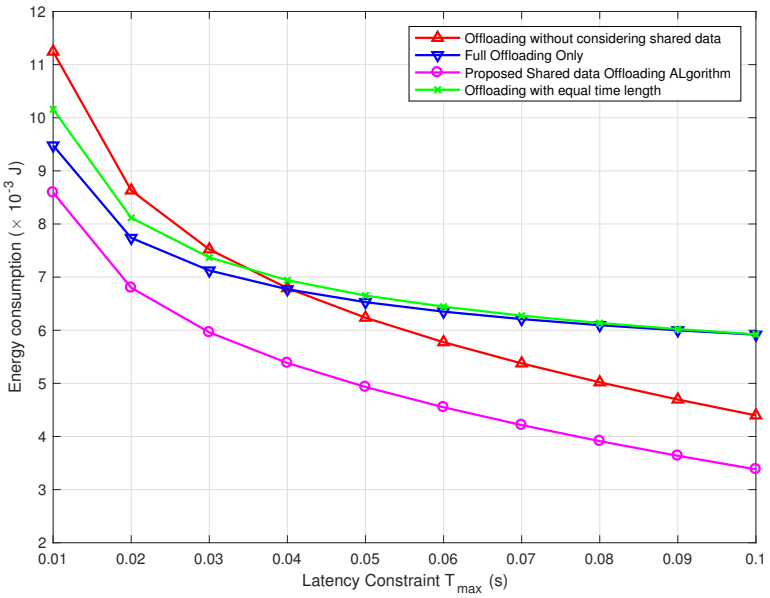


Figure 5.3: Energy consumption versus different latency constraints.

distance between user u and edge computing node in kilometers.

Figure 5.3 depicts how the energy consumption changes with different latency constraints. The energy consumption are becoming lower as the latency requirement gets longer for all listed offloading algorithms. Only the proposed offloading scheme can give the lowest energy consuming performance. The best energy saving improvement can only be achieved through the joint participation of local computing and shared data coordination. Moreover, even though the equal time length offloading has lower complexity than the proposed algorithm, it cannot compete with the proposed one in terms of energy saving. Recalling the conclusion that the best way to achieve the energy saving is to let these correlated bits transmitted by one specific user, the reason is that forcing offloading time duration to be equal makes the shared data to be transmitted by all users simultaneously.

The energy consumed for computing one data bit increases exponentially

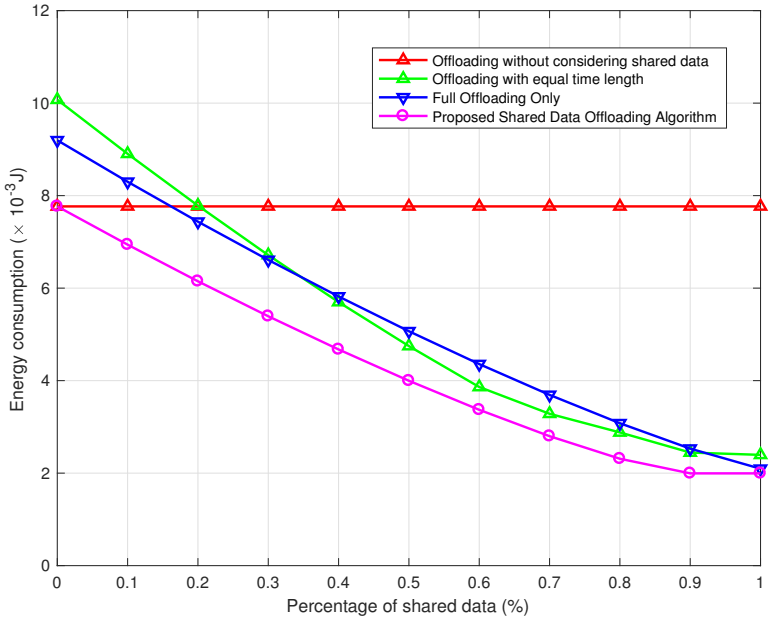


Figure 5.4: Energy consumption versus different percentage of shared data.

as the latency constraint diminishes. Hence for the local computing only scheme, when latency constraint comes to 0.01 second the energy taken to finish the computation tasks, which is 1000 mJoules, can reach up to nearly 100 times more than those of all the offloading algorithms. Then it drops exponentially to 10 mJoules when the latency constraint goes to 0.1 second. As a result, the curve representing local computing only is not added in Figure 5.3, otherwise the comparison of the offloading schemes will not be clear.

In Figure 5.4, the energy consumption changes with the percentage of shared data is demonstrated. Evidently, as long as the shared data is taken into consideration when making offloading decisions, the lower overall energy consumption is achieved when the proportion of shared data gets higher. More energy will be saved when the percentage of shared data gets higher for proposed offloading scheme compared to the scheme without consider-

ing the existence of shared data. This trend applies to the full offloading only algorithm as well, because it also cares about the existence of shared data when making offloading decisions. The energy consumptions for full offloading only do not always go under that of offloading without considering shared data. That is because when given specific latency constraint, the importance of local computing capabilities diminishes in saving mobile users' energy consumption as the share of common data increases. Since most of the data will be offloaded to the edge node, few input bits would remain local for computing. Then the energy consumption of the full offloading only scenario represents that it get closer to that of the proposed algorithm when the percentage of shared data increases. Similar trend applies to the equal time length offloading as well.

5.3 Joint Energy-Efficient Computation Offloading, Communications and Computational Resources Design

5.3.1 Motivation

In the previous section, the energy efficient communications resource allocation design is presented to minimise the mobile user's overall energy consumption. In this section, I am going to investigate the impact of computational resources in edge computing node on the energy efficient computation offloading design and the allocation of communications resources.

5.3.2 Problem Formulation

The overall energy consumption of the system consists of that from mobile users side and that of the cloudlet server side. The amount of energy expensed

at the mobile users side includes three parts: data offloading over the uplink (c.f. (5.2) and (5.3)), results retrieving (c.f. (5.6)), and local computing (c.f. (5.11)), which is thus given by

$$\begin{aligned} E_m = &\sum_{u\in\mathcal{U}} \kappa \frac{(\lambda_0 D_u^L)^3}{{t_{u,L}^C}^2} + \sum_{u\in\mathcal{U}} \frac{t_{u,S}^{ul}}{|h_u|^2} f(\frac{D_{u,S}^I}{t_{u,S}^{ul}}) \\ &+ \sum_{u\in\mathcal{U}} \frac{t_u^{ul}}{|h_u|^2} f(\frac{D_u^I - D_S^I - D_u^L}{t_u^{ul}}) + \sum_{u\in\mathcal{U}} (t_{u,S}^{dl} + t_u^{dl})\rho_u^{dl}. \end{aligned} \tag{5.31}$$

As for the energy depleted in the cloudlet server and AP side, taking computation and transmission energy consumption into consideration gives the expression as:

$$\begin{aligned} E_e = &\kappa_0 \sum_{u\in\mathcal{U}} [{f_u}^2 \lambda_0 (D_u^I - D_S^I - D_u^L) + \kappa_0 {f_s}^2 \lambda_0 D_S^I \\ &+ \sum_{u\in\mathcal{U}} \frac{t_{u,S}^{dl}}{|g_u|^2} \Gamma(\frac{a_0 D_S^I}{t_{u,S}^{dl}}) + \sum_{u\in\mathcal{U}} \frac{t_u^{dl}}{|g_u|^2} f(\frac{a_0 (D_u^I - D_S^I - D_u^L)}{t_u^{dl}})]. \end{aligned} \tag{5.32}$$

Upon considering the energy consumption of the whole system, different weights may be put on the energy consumption in the mobile users side and that in the cloudlet server side. By introducing the weights Ω_1 and Ω_2 which satisfies $\Omega_1 + \Omega_2 = 1$, the weighted sum energy consumption is:

$$E_{total} = \Omega_1 E_m + \Omega_2 E_e. \tag{5.33}$$

In this section, it aims to minimise the weighted sum of the energy consumption for completing the VR/AR computation tasks which include task offloading, local/edge computing and edge computing results downloading by optimizing the transmission and computation time, computational capacity of both mobile users and cloudlet server sides, and the offloaded input-bits.

Specially, we are interested in the following problem:

$$\begin{aligned}
(\mathrm{P1}): \underset{\{t^{ul}_{u,S}, t^{ul}_{u}, t^{C}_{u,L}, t^{dl}_{u,S}, t^{dl}_{u}, D^{I}_{u,S}, D^{L}_{u}, f_s, f_u\}}{\texttt{Minimise}} \quad & \Omega_1 E_m(t^{ul}_{u,S}, t^{ul}_{u}, t^{C}_{u,L}, t^{dl}_{u}, t^{dl}_{u,S}, D^{L}_{u}, D^{I}_{u,S}) \\
& + \Omega_2 E_e(t^{dl}_{u,S}, t^{dl}_{u}, D^{L}_{u}, f_s, f_u) \\
& \texttt{Subject to} \\
& \tau_u(t^{ul}_{u,S}, t^{ul}_{u}, t^{C}_{u,L}, t^{dl}_{u}, t^{dl}_{u,S}, D^{L}_{u}, D^{I}_{u,S}, f_s, f_u) \leq T_{\max}, \forall u \in \mathcal{U}, \quad (5.34a) \\
& 0 \leq t^{C}_{u,L} \leq T_{\max}, \forall u \in \mathcal{U}, \quad (5.34b) \\
& \lambda_0 D^{L}_{u} / t^{C}_{u,L} \leq f_{u,\max}, \forall u \in \mathcal{U}, \quad (5.34c) \\
& 0 \leq D^{L}_{u} \leq D^{I}_{u} - D^{I}_{S}, \forall u \in \mathcal{U}, \quad (5.34d) \\
& \sum_{u \in \mathcal{U}} D^{I}_{u,S} = D^{I}_{S}, D^{I}_{u,S} \geq 0, \forall u \in \mathcal{U}, \quad (5.34e) \\
& f_s \leq F, \quad (5.34f) \\
& \sum_{u \in \mathcal{U}} f_u \leq F, f_u \geq 0, \forall u \in \mathcal{U}, \quad (5.34g) \\
& t^{ul}_{u,S} \geq 0, t^{ul}_{u} \geq 0, t^{C}_{u,L} \geq 0, t^{dl}_{u} \geq 0, t^{dl}_{u,S} \geq 0, \forall u \in \mathcal{U}. \quad (5.34h)
\end{aligned}$$

In the formulated problem above, the objective function is the weighted sum of the system's energy consumption, where E_m (c.f. (5.31)) and E_e (c.f. (5.32)) are energy consumptions of mobile users side and edge server side in terms of optimisation variables. The latency constraint (5.34a) guarantees that the overall latency induced by transmission and computation will not exceed specific unified deadline $T_{\max}$. The local computing latency constraint (5.34b) states that the local computation conducted in the mobile users side should be finished within latency deadline $T_{\max}$, just as those task bits undergo offloading, remote computing and downloading returning to the mobile user end. In (5.34c), the application tasks bits remained for local computing is constrained by the maximum local computing capabilities $f_{u,\max}, \forall u \in \mathcal{U}$. The maximum amount of local computation bits are the

overall tasks bits subtracted by the shared data bits according to our design, which is reflected in constraint (5.34d). The shared input-bits (5.34e) reflects that the sum of the cooperatively offloaded shared input-bits $D^I_{u,S}$ from all mobile users add up to its total length D^I_S. The CPU computation frequency f_s allocated for shared data computation in the cloudlet server cannot exceed the maximum available CPU capabilities F (c.f. (5.34f)). Moreover, the sum of the edge server's allocated capabilities f_u's for each mobile user's individual input data is constrained by F as well (c.f. (5.34g)).

5.3.3 Joint Energy-Efficient Computation Offloading, Communications and Computational Resources Algorithm

Note that problem (P1) is not a convex problem due to the fractional form of two primal variables D^L_u's and f_u's in the expression of computational latency $t^C_u = \frac{\lambda_0(D^I_u - D^I_S - D^L_u)}{f_u}$. It is found that by fixing one of these two variable at one time, the optimisation problem goes to be convex. In the following subsections, the original problem (P1) is divided into two sub-problems (P1.1) and (P1.2) by fixing the value of D^L_u's and f_u's in (P1) respectively. The corresponding optimal solutions for the following two sub-problems are presented as well. Finally, the proposed alternating optimisation algorithm for solving (P1) is introduced based on iteratively solving sub-problems (P1.1) and (P1.2).

5.3.3.1 Optimal Solutions for (P1.1)

The optimisation problem (P1.1) is formulated by fixing the edge computing allocation f_u as:

$$
\begin{aligned}
(\text{P1.1}): \underset{\{t^{ul}_{u,S}, t^{ul}_{u}, t^{C}_{u,L}, t^{dl}_{u,S}, t^{dl}_{u}, D^{I}_{u,S}, D^{L}_{u}, f_s\}}{\texttt{Minimise}} \quad & \Omega_1 E_m(t^{ul}_{u,S}, t^{ul}_{u}, t^{C}_{u,L}, t^{dl}_{u}, t^{dl}_{u,S}, D^{L}_{u}, D^{I}_{u,S}) \\
& + \Omega_2 E_e(t^{dl}_{u,S}, t^{dl}_{u}, D^{L}_{u}, f_s, \overline{f}_u) \\
& \texttt{Subject to} \\
& \tau_u \leq T_{\max}, \forall u \in \mathcal{U}, && (5.35a) \\
& 0 \leq t^{C}_{u,L} \leq T_{\max}, \forall u \in \mathcal{U}, && (5.35b) \\
& \lambda_0 D^{L}_{u} / t^{C}_{u,L} \leq f_{u,\max}, && (5.35c) \\
& 0 \leq D^{L}_{u} \leq D^{I}_{u} - D^{I}_{S}, \forall u \in \mathcal{U}, && (5.35d) \\
& \sum_{u \in \mathcal{U}} D^{I}_{u,S} = D^{I}_{S}, D^{I}_{u,S} \geq 0, && (5.35e) \\
& f_s \leq F, && (5.35f) \\
& t^{ul}_{u,S} \geq 0, t^{ul}_{u} \geq 0, t^{C}_{u,L} \geq 0, t^{dl}_{u} \geq 0, t^{dl}_{u,S} \geq 0, \forall u \in \mathcal{U}. && (5.35g)
\end{aligned}
$$

It can be seen from (5.11) that the minimal energy consumption is obtained apparently by $t^{C}_{u,L} = T_{\max}$. In order to make the latency expression (5.34a) with $\max\{x, 0\}$ continuous and derivable with respect to primal variables, it needs to introduce necessary auxiliary variables $\{\mathbf{a}_1, a_2, a_3, \mathbf{a}_4, t^{ul}_{S}, t^{dl}_{S}\}$.

Then the optimisation problem is reformulated equivalently as:

$$(\text{P1.1}'): \underset{\{t^{ul}_{u,S}, t^{ul}_u, t^{dl}_u, t^{dl}_{u,S}, D^L_u, D^I_{u,S}, f_s, a_{1,u}, a_2, a_3, a_{4,u}, t^{ul}_S, t^{dl}_S\}}{\texttt{Minimise}} \quad \Omega_1 E_m(t^{ul}_{u,S}, t^{ul}_u, t^{dl}_u, t^{dl}_{u,S}, D^L_u, D^I_{u,S}) + \Omega_2 E_e(t^{dl}_{u,S}, t^{dl}_u, D^L_u, f_s, \overline{f}_u)$$

Subject to

(5.35c) − (5.35g),

$$t^{ul}_{u,S} \leq t^{ul}_S, \forall u \in \mathcal{U}, \tag{5.36a}$$

$$t^{dl}_{u,S} \leq t^{dl}_S, \forall u \in \mathcal{U}, \tag{5.36b}$$

$$t^{ul}_{u,S} + t^{ul}_u \leq a_{1,u}, \forall u \in \mathcal{U}, \tag{5.36c}$$

$$t^{ul}_S + \frac{\lambda_0 D^I_S}{f_s} \leq a_{1,u}, \forall u \in \mathcal{U}, \tag{5.36d}$$

$$a_{1,u} + \frac{\lambda_0 (D^I_u - D^I_S - D^L_u)}{\overline{f}_u} \leq a_{4,u}, \forall u \in \mathcal{U}, \tag{5.36e}$$

$$t^{ul}_{u,S} + t^{ul}_u \leq a_2, \tag{5.36f}$$

$$a_2 \leq a_3, \tag{5.36g}$$

$$t^{ul}_S + \frac{\lambda_0 D^I_S}{f_s} \leq a_3, \tag{5.36h}$$

$$a_3 + t^{dl}_S \leq a_{4,u}, \forall u \in \mathcal{U}, \tag{5.36i}$$

$$a_{4,u} + t^{dl}_u \leq T_{max}, \forall u \in \mathcal{U}. \tag{5.36j}$$

The Lagrangian is presented as:

$$
\begin{aligned}
&L(\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\phi},\boldsymbol{\varpi},\nu,\psi,\boldsymbol{\sigma},\boldsymbol{\theta},t_{u,S}^{ul},t_u^{ul},t_{u,S}^{dl},t_u^{dl},D_u^L,D_{u,S}^I,f_s,\mathbf{a}_1,a_2,a_3,\mathbf{a}_4,t_S^{ul},t_S^{dl}) = \\
&\Omega_1\Bigg[\sum_{u\in\mathcal{U}}\frac{t_{u,S}^{ul}}{|h_u|^2}f(\frac{D_{u,S}^I}{t_{u,S}^{ul}})+\sum_{u\in\mathcal{U}}\frac{t_u^{ul}}{|h_u|^2}f(\frac{D_u^I-D_S^I-D_u^L}{t_u^{ul}})+\sum_{u\in\mathcal{U}}\kappa\frac{(\lambda_0 D_u^L)^3}{{T_{max}}^2} \\
&+\sum_{u\in\mathcal{U}}(t_{u,S}^{dl}+t_u^{dl})\rho_u^{dl}\Bigg]+\Omega_2\Bigg[\kappa_0\sum_{u\in\mathcal{U}}[\overline{f}_u]^2\lambda_0(D_u^I-D_S^I-D_u^L)+\kappa_0{f_s}^2\lambda_0 D_S^I \\
&+\sum_{u\in\mathcal{U}}\frac{t_{u,S}^{dl}}{|g_u|^2}f(\frac{a_0 D_S^I}{t_{u,S}^{dl}})+\sum_{u\in\mathcal{U}}\frac{t_u^{dl}}{|g_u|^2}f(\frac{a_0(D_u^I-D_S^I-D_u^L)}{t_u^{dl}})]\Bigg]+\sum_{u\in\mathcal{U}}\beta_u(t_{u,S}^{ul}-t_S^{ul}) \\
&+\sum_{u\in\mathcal{U}}\omega_u(t_{u,S}^{dl}-t_S^{dl})+\sum_{u\in\mathcal{U}}\gamma_u(t_{u,S}^{ul}+t_u^{ul}-a_{1,u})+\sum_{u\in\mathcal{U}}\delta_u(t_S^{ul}+\frac{\lambda_0 D_S^I}{f_s}-a_{1,u}) \\
&+\sum_{u\in\mathcal{U}}\phi_u(a_{1,u}+\frac{\lambda_0(D_u^I-D_S^I-D_u^L)}{\overline{f}_u}-a_{4,u})+\sum_{u\in\mathcal{U}}\varpi_u(t_{u,S}^{ul}+t_u^{ul}-a_2)+\nu(a_2-a_3) \\
&+\psi(t_S^{ul}+\frac{\lambda_0 D_S^I}{f_s}-a_3)+\sum_{u\in\mathcal{U}}\sigma_u(a_3+t_S^{dl}-a_{4,u})+\sum_{u\in\mathcal{U}}\theta_u(a_{4,u}+t_u^{dl}-T_{max}),
\end{aligned}
\tag{5.37}
$$

where $\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\phi},\boldsymbol{\varpi},\nu,\psi,\boldsymbol{\sigma},\boldsymbol{\theta}$ are Lagrangian multipliers associated with constraints (5.36a)-(5.36j), respectively. Hence, the dual function is expressed as:

$$
\begin{aligned}
&g(\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\phi},\boldsymbol{\varpi},\nu,\psi,\boldsymbol{\sigma},\boldsymbol{\theta}) = \\
&\underset{\{t_{u,S}^{ul},t_u^{ul},t_{u,S}^{dl},t_u^{dl},D_u^L,D_{u,S}^I,f_s,\mathbf{a}_1,a_2,a_3,\mathbf{a}_4,t_S^{ul},t_S^{dl}\}}{\texttt{Minimise}} L(\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\phi},\boldsymbol{\varpi},\nu,\psi,\boldsymbol{\sigma},\boldsymbol{\theta}, \\
&t_{u,S}^{ul},t_u^{ul},t_{u,S}^{dl},t_u^{dl},D_u^L,D_{u,S}^I,f_s,\mathbf{a}_1,a_2,a_3,\mathbf{a}_4,t_S^{ul},t_S^{dl})
\end{aligned}
\tag{5.38}
$$

s.t.

$$
\begin{aligned}
&(5.35\text{d})-(5.35\text{g}), \\
&D_u^L \le T_{\max} f_{u,\max}/\lambda_0.
\end{aligned}
$$

$$\hat{D}_u^L = \min\left\{T_{\max}\sqrt{\left[\frac{N_0 ln2}{3\kappa_0{\lambda_0}^3}(\frac{2^{\frac{\hat{r}_u^{ul}}{W_u^{ul}}}}{W_u^{ul}|h_u|^2}+\frac{\Omega_2 a_0\cdot 2^{\frac{a_0\hat{r}_u^{dl}}{W_u^{dl}}}}{\Omega_1 W_u^{dl}|g_u|^2})+\frac{\phi_u}{3\overline{f}_u\kappa_0\lambda_0^2\Omega_1}+\frac{\Omega_2\kappa_0\lambda_0\overline{f}_u^{\,2}}{\Omega_1}\right]^+},\right.$$
$$\left.\min\left\{\frac{T_{\max}f_{u,\max}}{\lambda_0}, D_u^I - D_S^I\right\}\right\}, \forall u\in\mathcal{U}, \tag{5.39}$$

Consequently, the corresponding dual problem is formulated as:

$$\max_{\{\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\phi},\boldsymbol{\varpi},\nu,\psi,\boldsymbol{\sigma},\boldsymbol{\theta}\}} g(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\phi},\boldsymbol{\varpi},\nu,\psi,\boldsymbol{\sigma},\boldsymbol{\theta}) \tag{5.40}$$

s.t.

$$\boldsymbol{\beta}\succeq 0,\boldsymbol{\omega}\succeq 0,\boldsymbol{\gamma}\succeq 0,\boldsymbol{\delta}\succeq 0,\boldsymbol{\phi}\succeq 0,$$
$$\boldsymbol{\varpi}\succeq 0,\nu>0,\psi>0,\boldsymbol{\sigma}\succeq 0,\boldsymbol{\theta}\succeq 0$$

Proposition 2. *Given a determined set of dual variables $\boldsymbol{\beta},\boldsymbol{\omega},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\phi},\boldsymbol{\varpi},\nu,\psi,\boldsymbol{\sigma},\boldsymbol{\theta}$, the optimal solution to the problem* (5.38) *can be determined as follows:*

The optimal primal variables $t_{u,S}^{ul}$, t_u^{ul}, $t_{u,S}^{dl}$ and t_u^{dl}, are given by

$$\hat{t}_{u,S}^{ul} = \frac{\hat{D}_{u,S}^I}{\frac{W}{ln2}\left[W_0(\frac{1}{e}(\frac{(\beta_u+\gamma_u+\varpi_u)|h_u|^2}{\Omega_1 N_0}-1)+1)\right]}, \forall u\in\mathcal{U}, \tag{5.41}$$

$$\hat{t}_u^{ul} = \frac{D_u^I - D_S^I - \hat{D}_u^L}{\frac{W}{ln2}\left[W_0(\frac{1}{e}(\frac{(\gamma_u+\varpi_u)|h_u|^2}{\Omega_1 N_0}-1)+1)\right]}, \forall u\in\mathcal{U}, \tag{5.42}$$

$$\hat{t}_{u,S}^{dl} = \frac{a_0 D_S^I}{\frac{W_{all}}{ln2}\left[W_0(\frac{1}{e}(\frac{(\Omega_1\rho_u^{dl}+\sigma_u)|g_u|^2}{\Omega_2 N_0}-1)+1)\right]}, \forall u\in\mathcal{U}, \tag{5.43}$$

$$\hat{t}_u^{dl} = \frac{a_0(D_u^I - D_S^I - \hat{D}_u^L)}{\frac{W}{ln2}\left[W_0(\frac{1}{e}(\frac{(\Omega_1\rho_u^{dl} + \theta_u)|g_u|^2}{\Omega_2 N_0} - 1) + 1)\right]}, \forall u \in \mathcal{U}, \tag{5.44}$$

where $W_0(x)$ is the principle branch of the Lambert W function defined as the solution for $W_0(x)e^{W_0(x)} = x$ [WXWC18], e is the base of the natural logarithm; given dual variables, the denominators of the above-mentioned expression in this proposition, which is denoted by $\hat{r}_{u,S}^{ul}$'s($\hat{r}_u^{ul}$'s) and $\hat{r}_{u,S}^{dl}$'s($\hat{r}_u^{dl}$'s), stand for the optimum transmission rate for the shared(individual) data offloading and the shared(individual) data downloading, respectively; notably, the optimal values of $\hat{r}_{u,S}^{ul}$'s should be obtained given $\hat{D}_{u,S}^I, \forall u \in \mathcal{U}$, which is provided by Lemma 2; the optimal solutions of the local computing bits are provided in (5.39)*; the optimal auxiliary variables are given as:*

$$\hat{a}_{1,u} = \begin{cases} T_{\max}, & -\gamma_u - \delta_u + \phi_u < 0, \\ 0, & otherwise, \end{cases}$$

$$\hat{a}_2 = \begin{cases} T_{\max}, & -\sum_{u\in\mathcal{U}} \varpi_u + \nu < 0, \\ 0, & otherwise, \end{cases}$$

$$\hat{a}_3 = \begin{cases} T_{\max}, & -\nu - \psi + \sum_{u\in\mathcal{U}} \sigma_u < 0, \\ 0, & otherwise, \end{cases}$$

$$\hat{a}_{4,u} = \begin{cases} T_{\max}, & -\phi_u - \sigma_u + \theta_u < 0, \\ 0, & otherwise, \end{cases}$$

$$\hat{t}_S^{ul} = \begin{cases} T_{\max}, & -\sum_{u\in\mathcal{U}} \beta_u + \sum_{u\in\mathcal{U}} \delta_u < 0, \\ 0, & otherwise, \end{cases}$$

$$\hat{t}_S^{dl} = \begin{cases} T_{\max}, & -\sum_{u\in\mathcal{U}} \omega_u + \sum_{u\in\mathcal{U}} \sigma_u < 0, \\ 0, & otherwise, \end{cases}$$

and the optimal local computing data size is given by (5.39)*. The remote computing rate for shared data,* $\hat{f}_s$ *would be calculated by*

$$\hat{f}_s = \min\left(\sqrt[3]{\frac{\sum_{u\in\mathcal{U}} \delta_u + \psi}{2\kappa_0}}, F\right). \tag{5.45}$$

The following lemma gives the optimal offloaded input-bits of the shared data for each user, i.e., $\hat{D}_{u,S}^I$.

Lemma 2. *The optimal offloaded shared data for user* u *is expressed as,*

$$\hat{D}_{u,S}^I = \begin{cases} D_S^I, & \hat{u} = arg\min_{1\le u\le U} \Delta_u, \\ 0, & otherwise, \end{cases} \tag{5.46}$$

where $\Delta_u = \frac{\Omega_1 f(\hat{r}_{u,S}^{ul})}{\hat{r}_{u,S}^{ul}|h_u|^2} + \frac{\beta_u+\gamma_u+\varpi_u}{\hat{r}_{u,S}^{ul}}, \forall u \in \mathcal{U}$.

It is concluded from the above Lemma that the shared data is optimally offloaded by one specific user instead of multiple ones simultaneously. Remarkably, the user with maximum $\hat{r}_{u,S}^{ul}$ is not necessarily the one that is responsible for the shared input-bits offloading. It can be seen that the expression of $\Delta_u, \forall u \in \mathcal{U}$ is also associated with some Lagrangian variables. For example, among these Lagrangian variables γ_u's and ϖ_u's are related to the transmission rate of individual input-bits and relevant auxiliary variables, which may make their value large and therefore the user with maximum $\hat{r}_{u,S}^{ul}$ not the one with minimum Δ_u.

Based on Proposition 1, the dual problem can thus be iteratively solved according to ellipsoid method (with constraints), the detail of which can be

referred to [Boy]. The algorithm for solving (P1.1) is summarised in Table 5.2.

Table 5.2: Algorithm I for solving (P1.1′)

Require: $(\boldsymbol{\beta^{(0)}}, \boldsymbol{\omega^{(0)}}, \boldsymbol{\gamma^{(0)}}, \boldsymbol{\delta^{(0)}}, \boldsymbol{\phi^{(0)}}, \boldsymbol{\varpi^{(0)}}, \nu^{(0)}, \psi^{(0)}, \boldsymbol{\sigma^{(0)}}, \boldsymbol{\theta^{(0)}})$

1: **repeat**

2: Solve (P1.1′) given $(\boldsymbol{\beta^{(i)}}, \boldsymbol{\omega^{(i)}}, \boldsymbol{\gamma^{(i)}}, \boldsymbol{\delta^{(i)}}, \boldsymbol{\phi^{(i)}}, \boldsymbol{\varpi^{(i)}}, \nu^{(i)}, \psi^{(i)}, \boldsymbol{\sigma^{(i)}}, \boldsymbol{\theta^{(i)}})$ according to Proposition 2 and obtain $\{\hat{t}^{ul}_{u,S}, \hat{t}^{ul}_{u}, \hat{t}^{dl}_{u,S}, \hat{t}^{dl}_{u}, \hat{D}^{L}_{u}, \hat{D}^{I}_{u,S}, \hat{f}_s, \hat{a}_{1,u}, \hat{a}_2, \hat{a}_3, \hat{a}_{4,u}, \hat{t}^{ul}_{S}, \hat{t}^{dl}_{S}\}$;

3: update the subgradient of $\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\phi}, \boldsymbol{\varpi}, \nu, \psi, \boldsymbol{\sigma}, \boldsymbol{\theta}$ respectively, i.e., $t^{ul}_{u,S} - t^{ul}_{S}$, $t^{dl}_{u,S} - t^{dl}_{S}$, $t^{ul}_{u,S} + t^{ul}_{u} - a_{1,u}$, $t^{ul}_{S} + \frac{\lambda_0 D^I_S}{f_s} - a_{1,u}$, $a_{1,u} + \frac{\lambda_0 (D^I_u - D^I_S - D^L_u)}{\overline{f}_u} - a_{4,u}$, $t^{ul}_{u,S} + t^{ul}_{u} - a_2$, $a_2 - a_3$, $t^{ul}_{S} + \frac{\lambda_0 D^I_S}{f_s} - a_3$, $a_3 + t^{dl}_{S} - a_{4,u}$, $a_{4,u} + t^{dl}_{u} - T_{max}$, in accordance with the ellipsoid method [Boy];

4: **until** the predefined accuracy threshold is satisfied.

Ensure: The optimal dual variables to the dual problem (5.40) $(\boldsymbol{\beta^*}, \boldsymbol{\omega^*}, \boldsymbol{\gamma^*}, \boldsymbol{\delta^*}, \boldsymbol{\phi^*}, \boldsymbol{\varpi^*}, \nu^*, \psi^*, \boldsymbol{\sigma^*}, \boldsymbol{\theta^*})$

5: Solve (P1.1′) again with $(\boldsymbol{\beta^*}, \boldsymbol{\omega^*}, \boldsymbol{\gamma^*}, \boldsymbol{\delta^*}, \boldsymbol{\phi^*}, \boldsymbol{\varpi^*}, \nu^*, \psi^*, \boldsymbol{\sigma^*}, \boldsymbol{\theta^*})$, update the primal variables

Ensure: $\{t^{ul*}_{u,S}, t^{ul*}_{u}, t^{dl*}_{u,S}, t^{dl*}_{u}, D^{L*}_{u}, D^{I*}_{u,S}, f^*_s, a^*_{1,u}, a^*_2, a^*_3, a^*_{4,u}, t^{ul*}_{S}, t^{dl*}_{S}\}$

5.3.3.2 Optimal Solutions for (P1.2)

Problem (P1.1) is given by fixing the primal variable f_u in the original problem (P1). In this section, through fixing the primal variable D^L_u, the optimi-

sation problem (P1.2) is formulated as:

$$\begin{aligned}
(\text{P1.2}): \underset{\{t^{ul}_{u,S}, t^{ul}_{u}, t^{C}_{u,L}, t^{dl}_{u,S}, t^{dl}_{u}, D^{I}_{u,S}, f_s, f_u\}}{\texttt{Minimise}} \quad & \Omega_1 E_m(t^{ul}_{u,S}, t^{ul}_{u}, t^{dl}_{u}, t^{dl}_{u,S}, \overline{D}^{L}_{u}, D^{I}_{u,S}) \\
& + \Omega_2 E_e(t^{dl}_{u,S}, t^{dl}_{u}, D^{L}_{u}, f_s, f_u) \\
& \texttt{Subject to} \\
& \tau_u \leq T_{\max}, \forall u \in \mathcal{U}, && \text{(5.47a)} \\
& \sum_{u \in \mathcal{U}} f_u \leq F, f_u \geq 0, \forall u \in \mathcal{U}, && \text{(5.47b)} \\
& 0 \leq t^{C}_{u,L} \leq T_{\max}, \forall u \in \mathcal{U}, && \text{(5.47c)} \\
& \sum_{u \in \mathcal{U}} D^{I}_{u,S} = D^{I}_{S}, D^{I}_{u,S} \geq 0, && \text{(5.47d)} \\
& f_s \leq F, && \text{(5.47e)} \\
& t^{ul}_{u,S} \geq 0, t^{ul}_{u} \geq 0, t^{C}_{u,L} \geq 0, t^{dl}_{u,S} \geq 0, t^{dl}_{u} \geq 0, \forall u \in \mathcal{U}. && \text{(5.47f)}
\end{aligned}$$

Given corresponding dual variables, the solutions of the primal variables $t^{ul}_{u,S}$, t^{ul}_{u}, $t^{dl}_{u,S}$, t^{dl}_{u}, f_s, and $D^{I}_{u,S}$ for the problem (P1.2) are similar to those in equation (5.41), (5.42), (5.43), (5.44), (5.45), (5.46) for the problem (P1.1), respectively. The solution for $t^{C}_{u,L}$ is T_{max}. The optimal solutions to the primal variables f_u is given by the solutions of the following formula:

$$2\Omega_2\kappa_0\lambda_0(D^{I}_{u} - D^{I}_{S} - \overline{D}^{L}_{u}){f_u}^3 + \varphi {f_u}^2 - \lambda_0\phi_u(D^{I}_{u} - D^{I}_{S} - \overline{D}^{L}_{u}) = 0 \quad (5.48)$$

Lemma 3. *According to Vieta's formulas that relate the coefficients of a polynomial to sums and products of its roots, there exists a unique non-negative real number that would be the value of* $\hat{f}_u$.

Proof. Assuming that x_1, x_2 and x_3 are the roots of the polynomial (5.48). Based

on Vieta's formulas, the relationships between the coefficients and the sums and products of its roots are given as:

$$x_1 + x_2 + x_3 = -\varphi/2\Omega_2\kappa_0\lambda_0(D_u^I - D_S^I - \overline{D}_u^L) \leq 0,$$
$$x_1x_2 + x_1x_3 + x_2x_3 = 0,$$
$$x_1x_2x_3 = \phi_u/2\Omega_2\kappa_0 \geq 0.$$

According to their relationships with zero, if there are three real number roots, except for the case that x_1=x_2=x_3=0, there can only be one positive real number and two negative real numbers that are the roots of (5.48). If there is one real number root and two complex conjugates roots, the real number must be greater than zero. To sum up, there always exists $\hat{f}_u \geq 0$, $\forall u \in \mathcal{U}$ from the roots of (5.48). □

5.3.3.3 Alternating Optimisation for Solving P1

By fixing specific primal variable in problem (P1), it gives (P1.1) and (P1.2) which are both convex and ready for obtaining optimal solutions, respectively. It can be proved that through solving (P1.1) and (P1.2) in an iterative manner, the local optimum of the original problem can be attained [BH03]. In this section, the alternating optimisation algorithm is proposed for obtaining the solution to achieve a local optimum of the formulated problem, which is introduced and explained as follows

The algorithm starts in initializing the allocated edge computing capabilities as $\hat{f}_u^{(0)} = F/U$, which means that the cloudlet computing capabilities are assumed to be equally allocated to process the offloaded input-bits of every mobile user. Under this circumstance, the first set of optimal solutions $\{\hat{t}_{u,S}^{ul(i)}, \hat{t}_u^{ul(i)}, \hat{t}_{u,S}^{dl(i)}, \hat{t}_u^{dl(i)}, \hat{D}_u^{L(i)}, \hat{D}_{u,S}^{I(i)}, \hat{f}_s^{(i)}\}$ is obtained for (P1.1). Among all these obtained primal variables solutions, it takes the local computing

Table 5.3: Alternating optimisation algorithm for obtaining the optimum of (P1)

Require: $i = 0, \hat{E}_m^{(0)} = 0, \hat{E}_e^{(0)} = 0, \hat{f}_u^{(0)} = F/U, \forall u \in \mathcal{U}$, and the stopping criterion $\varepsilon = 10^{-3}$.

1: **repeat**
2: $i = i + 1$
3: solve (P1.1) given $\overline{f}_u = \hat{f}_u^{(i-1)}, \forall u \in \mathcal{U}$, obtain $\hat{E}_m^{(i)}, \hat{E}_e^{(i)}$, and $\{\hat{t}_{u,S}^{ul(i)}, \hat{t}_u^{ul(i)}, \hat{t}_{u,S}^{dl(i)}, \hat{t}_u^{dl(i)}, \hat{D}_u^{L(i)}, \hat{D}_{u,S}^{I(i)}, \hat{f}_s^{(i)}\}$;
4: $i = i + 1$
5: solve (P1.2) given $\overline{D}_u^L = \hat{D}_u^{L(i-1)}, \forall u \in \mathcal{U}$, obtain $\hat{E}_m^{(i)}, \hat{E}_e^{(i)}$, and $\{\hat{t}_{u,S}^{ul(i)}, \hat{t}_u^{ul(i)}, \hat{t}_{u,S}^{dl(i)}, \hat{t}_u^{dl(i)}, \hat{D}_{u,S}^{I(i)}, \hat{f}_s^{(i)}, \hat{f}_u^{(i)}\}$;
6: **until** $\frac{\hat{E}_{total}^{(i-1)} - \hat{E}_{total}^{(i)}}{\hat{E}_{total}^{(i-1)}} < \varepsilon$, where $\hat{E}_{total}^{(i)} = \Omega_1 \hat{E}_m^{(i)} + \Omega_2 \hat{E}_e^{(i)}$.
7: **return** $E_m^* \Leftarrow \hat{E}_m^{(i)}, E_e^* \Leftarrow \hat{E}_e^{(i)}, \{t_{u,S}^{ul*}, t_u^{ul*}, t_{u,S}^{dl*}, t_u^{dl*}, D_u^{L*}, D_{u,S}^{I*}, f_s^*, f_u^*\} \Leftarrow \{\hat{t}_{u,S}^{ul(i)}, \hat{t}_u^{ul(i)}, \hat{t}_{u,S}^{dl(i)}, \hat{t}_u^{dl(i)}, \hat{D}_u^{L(i)}, \hat{D}_{u,S}^{I(i)}, \hat{f}_s^{(i)}, \hat{f}_u^{(i)}\}$

input-bits $\hat{D}_u^{L(i)}, \forall u \in \mathcal{U}$, as the fixed primal variables value $\overline{D}_u^L, \forall u \in \mathcal{U}$, for (P1.2). Then (P1.2) undergoes the same problem-solving process as that of (P1.1), obtaining another set of optimal solutions including the allocated cloudlet computing capabilities as $\hat{f}_u^{(i)}, \forall u \in \mathcal{U}$. Afterwards, $\overline{f}_u$ for (P1,1) is set to be equal to $\hat{f}_u^{(i)}$ obtained in the previous iteration, the optimal solutions for (P1.1) is then solved again producing relevant solutions set. The process repeats until the stopping criterion is met.

5.3.4 Special Case: Negligible Computational Latency

With limited cloudlet computing capabilities the edge-computing latency is non-negligible, which will certainly impact on the overall latency, since the downloading of output-bits cannot start until the computation results are available in the cloudlet. In the previous non-negligible computational latency case, the obtained solutions for the primal variables are the sub-optimal.

If the edge computational latency is short enough that has no impact

on the overall latency, it only needs to consider the transmission latency, which greatly reduce the complexity of solving the optimisation problem and the global optimum can be readily obtained as well. Recalling (5.12), the negligible computation time reduce the expression of τ_1 to $\tau_1' = t_{u,S}^{ul} + t_u^{ul}$. After reducing the subsequent latency expressions based on the negligible computational latency assumption, the expression (5.15) is finally reduced to:

$$\tau'_u = t_{u,S}^{ul} + t_u^{ul} + \max_{u \in \mathcal{U}} \{t_{u,S}^{dl}\} + t_u^{dl} \tag{5.50}$$

In this case, the allocation of edge computing capability for shared and individual input bits f_s and f_u is also excluded from the optimisation problem of the special case. In this context, by replacing E_S^C and E_u^C in (5.33) with pre-defined energy consumption E_C, the energy consumption of mobile users' and edge server's sides are given as:

$$\begin{aligned} E'_m = & \sum_{u \in \mathcal{U}} \kappa_0 \frac{(\lambda_0 D_u^L)^3}{{t_{u,L}^C}^2} + \sum_{u \in \mathcal{U}} \frac{t_{u,S}^{ul}}{|h_u|^2} f(\frac{D_{u,S}^I}{t_{u,S}^{ul}}) \\ & + \sum_{u \in \mathcal{U}} \frac{t_u^{ul}}{|h_u|^2} f(\frac{D_u^I - D_S^I - D_u^L}{t_u^{ul}}) + \sum_{u \in \mathcal{U}} (t_{u,S}^{dl} + t_u^{dl}) \rho_u^{dl}, \end{aligned} \tag{5.51}$$

and

$$E'_e = E_C + \sum_{u \in \mathcal{U}} \frac{t_{u,S}^{dl}}{|g_u|^2} \Gamma(\frac{a_0 D_S^I}{t_{u,S}^{dl}}) + \sum_{u \in \mathcal{U}} \frac{t_u^{dl}}{|g_u|^2} f(\frac{a_0 (D_u^I - D_S^I - D_u^L)}{t_u^{dl}})]. \tag{5.52}$$

The optimisation problem associated with this special case as:

$$(\mathrm{P1}'): \underset{\{t^{ul}_{u,S}, t^{ul}_{u}, t^{C}_{u,L}, t^{dl}_{u}, t^{dl}_{u,S}, D^{L}_{u}, D^{I}_{u,S}\}}{\texttt{Minimise}} \quad \Omega_1 E'_m(t^{ul}_{u,S}, t^{ul}_{u}, t^{C}_{u,L}, t^{dl}_{u}, t^{dl}_{u,S}, D^{L}_{u}, D^{I}_{u,S}) + \Omega_2 E'_e(t^{dl}_{u,S}, t^{dl}_{u}, D^{L}_{u})$$

$$\texttt{Subject to}$$

$$t^{ul}_{u,S} + t^{ul}_{u} + \max_{u\in\mathcal{U}}\{t^{dl}_{u,S}\} + t^{dl}_{u} \leq T_{\max}, \forall u \in \mathcal{U}, \tag{5.53a}$$

$$0 \leq t^{C}_{u,L} \leq T_{\max}, \forall u \in \mathcal{U}, \tag{5.53b}$$

$$\lambda_0 D^{L}_{u} / t^{C}_{u,L} \leq f_{u,\max}, \forall u \in \mathcal{U}, \tag{5.53c}$$

$$0 \leq D^{L}_{u} \leq D^{I}_{u} - D^{I}_{S}, \forall u \in \mathcal{U}, \tag{5.53d}$$

$$\sum_{u\in\mathcal{U}} D^{I}_{u,S} = D^{I}_{S}, D^{I}_{u,S} \geq 0, \tag{5.53e}$$

$$t^{ul}_{u,S} \geq 0, t^{ul}_{u} \geq 0, t^{C}_{u,L} \geq 0, t^{dl}_{u} \geq 0, t^{dl}_{u,S} \geq 0, \forall u \in \mathcal{U}. \tag{5.53f}$$

The given optimisation problem is in the strictly convex form, which can be readily solved by classic convex optimisation methodology in a straightforward manner. The optimal solutions for $t^{C}_{u,L}$'s are $t^{C}_{u,L} = T_{\max}, \forall u \in \mathcal{U}$ according to (5.51), which is monotonically decreasing with respect to $t^{C}_{u,L}$'s. For obtaining the optimal solutions for the rest primal variables in this case, it needs to introduce one auxiliary variable t^{dl}_{S}, so that

$$t^{dl}_{u,S} \leq t^{dl}_{S}, \forall u \in \mathcal{U}, \tag{5.54}$$

and

$$t^{ul}_{u,S} + t^{ul}_{u} + t^{dl}_{S} + t^{dl}_{u} \leq T_{\max}, \forall u \in \mathcal{U}. \tag{5.55}$$

There need two sets of Lagrangian variables $\boldsymbol{\omega} = \{\omega_1, ..., \omega_U\}$ and $\boldsymbol{\theta} = \{\theta_1, ..., \theta_U\}$, which are associated to (5.54) and (5.55), respectively, to de-

rive the dual function:

$$
\begin{aligned}
&L(\boldsymbol{\omega},\boldsymbol{\theta},t_{u,S}^{ul},t_u^{ul},t_{u,S}^{dl},t_u^{dl},D_u^L,D_{u,S}^I,t_S^{dl}) = \\
&\Omega_1\Bigg[\sum_{u\in\mathcal{U}}\frac{t_{u,S}^{ul}}{|h_u|^2}f(\frac{D_{u,S}^I}{t_{u,S}^{ul}})+\sum_{u\in\mathcal{U}}\frac{t_u^{ul}}{|h_u|^2}f(\frac{D_u^I-D_S^I-D_u^L}{t_u^{ul}})+\sum_{u\in\mathcal{U}}\kappa\frac{(\lambda_0 D_u^L)^3}{{T_{\max}}^2} \\
&+\sum_{u\in\mathcal{U}}(t_{u,S}^{dl}+t_u^{dl})\rho_u^{dl}\Bigg]+\Omega_2\Bigg[\kappa_0\sum_{u\in\mathcal{U}}[\overline{f}_u{}^2\lambda_0(D_u^I-D_S^I-D_u^L)+\kappa_0 {f_s}^2\lambda_0 D_S^I \\
&+\sum_{u\in\mathcal{U}}\frac{t_{u,S}^{dl}}{|g_u|^2}f(\frac{a_0 D_S^I}{t_{u,S}^{dl}})+\sum_{u\in\mathcal{U}}\frac{t_u^{dl}}{|g_u|^2}f(\frac{a_0(D_u^I-D_S^I-D_u^L)}{t_u^{dl}})]\Bigg]+\sum_{u\in\mathcal{U}}\omega_u(t_{u,S}^{dl}-t_S^{dl}) \\
&+\sum_{u\in\mathcal{U}}\theta_u(t_{u,S}^{ul}+t_u^{ul}+t_S^{dl}+t_u^{dl}-T_{\max}),
\end{aligned}
\tag{5.56}
$$

Through applying dual analysis to the derived Lagrangian function of (P1′), the optimal solutions for the primal variables $\{t_{u,S}^{ul},t_u^{ul},t_{u,S}^{dl},t_u^{dl},D_u^L,D_{u,S}^I\}$ and auxiliary variable t_S^{dl} are given.

Proposition 3. *Given a determined set of dual variables $\boldsymbol{\omega},\boldsymbol{\theta}$, the optimal solution to the optimisation problem of the special case can be determined as follows.*

The optimal primal variables $t_{u,S}^{ul}$, t_u^{ul}, $t_{u,S}^{dl}$ and t_u^{dl}, are given by

$$
\hat{t}_{u,S}^{ul}=\frac{\hat{D}_{u,S}^I}{\frac{W}{ln2}\left[W_0(\frac{1}{e}(\frac{\theta_u|h_u|^2}{\Omega_1 N_0}-1)+1)\right]},\forall u\in\mathcal{U}. \tag{5.57}
$$

$$
\hat{t}_u^{ul}=\frac{D_u^I-D_S^I-\hat{D}_u^L}{\frac{W}{ln2}\left[W_0(\frac{1}{e}(\frac{\theta_u|h_u|^2}{\Omega_1 N_0}-1)+1)\right]},\forall u\in\mathcal{U}. \tag{5.58}
$$

$$
\hat{t}_{u,S}^{dl}=\frac{a_0 D_S^I}{\frac{W_{all}}{ln2}\left[W_0(\frac{1}{e}(\frac{(\Omega_1\rho_u^{dl}+\omega_u)|g_u|^2}{\Omega_2 N_0}-1)+1)\right]},\forall u\in\mathcal{U}. \tag{5.59}
$$

$$
\hat{t}_u^{dl}=\frac{a_0(D_u^I-D_S^I-\hat{D}_u^L)}{\frac{W}{ln2}\left[W_0(\frac{1}{e}(\frac{(\Omega_1\rho_u^{dl}+\theta_u)|g_u|^2}{\Omega_2 N_0}-1)+1)\right]},\forall u\in\mathcal{U}. \tag{5.60}
$$

where $W_0(x)$ is the principle branch of the Lambert W function defined as the solution for $W_0(x)e^{W_0(x)} = x$ [WXWC18], e is the base of the natural logarithm. The values of the optimal solutions of $\hat{r}_{u,S}^{ul}$'s should be obtained given $\hat{D}_{u,S}^{I}, \forall u \in \mathcal{U}$, which is provided by Lemma 4. The optimal values for local computing bits D_u^L's are given by:

$$\hat{D}_u^L = \min\left\{T_{\max}\sqrt{\left[\frac{N_0 ln2}{3\kappa_0{\lambda_0}^3}(\frac{2^{\frac{\hat{r}_u^{ul}}{W_u^{ul}}}}{W_u^{ul}|h_u|^2} + \frac{\Omega_2 a_0 \cdot 2^{\frac{a_0\hat{r}_u^{dl}}{W_u^{dl}}}}{\Omega_1 W_u^{dl}|g_u|^2})\right]^+}, \min\left\{\frac{T_{\max}f_{u,\max}}{\lambda_0}, D_u^I - D_S^I\right\}\right\}, \forall u \in \mathcal{U}. \tag{5.61}$$

The optimal auxiliary variable t_S^{dl} is obtained as:

$$\hat{t}_S^{dl} = \begin{cases} T_{\max}, & -\sum_{u\in\mathcal{U}}\omega_u + \sum_{u\in\mathcal{U}}\theta_u < 0, \\ 0, & otherwise, \end{cases}$$

Lemma 4. *The optimal offloaded shared data for user u is expressed as,*

$$\hat{D}_{u,S}^I = \begin{cases} D_S^I, & \hat{u} = arg\min_{1\le u\le U}\Delta_u, \\ 0, & otherwise, \end{cases} \tag{5.62}$$

where $\Delta_u = \frac{\Omega_1 f(\hat{r}_{u,S}^{ul})}{\hat{r}_{u,S}^{ul}|h_u|^2} + \frac{\theta_u}{\hat{r}_{u,S}^{ul}}, \forall u \in \mathcal{U}$.

The algorithm for obtaining the optimum of the special case is presented in Table 5.4.

5.3.5 Simulation Results and Performance Analysis

In this section, the simulation results of the proposed algorithm together with other baseline algorithms are presented. The effectiveness of the proposed joint optimisation of cooperative shared task data offloading, wireless transmit power allocation and computation resource allocation is validated by the comparison against the fixed-frequency scheme, neglecting shared data

Table 5.4: Algorithm for obtaining the optimum of the special case

Require: $(\boldsymbol{\beta^{(0)}}, \boldsymbol{\theta^{(0)}})$

1: **repeat**

2: Solve Lagrangian dual problem given $(\boldsymbol{\beta^{(i)}}, \boldsymbol{\theta^{(i)}})$ according to the process according to Proposition 3 and obtain $\{\hat{t}^{ul}_{u,S}, \hat{t}^{ul}_{u}, \hat{t}^{dl}_{u,S}, \hat{t}^{dl}_{u}, \hat{D}^{L}_{u}, \hat{D}^{I}_{u,S}, \hat{t}^{dl}_{S}\}$;

3: update the subgradient of $\boldsymbol{\beta}, \boldsymbol{\theta}$ respectively, i.e., $t^{dl}_{u,S} - t^{dl}_{S}$, $t^{ul}_{u,S} + t^{ul}_{u} + t^{dl}_{S} + t^{dl}_{u} - T_{max}$, in accordance with the ellipsoid method [Boy];

4: **until** the predefined accuracy threshold is satisfied.

Ensure: The optimal dual variables to the dual problem (5.40) $(\boldsymbol{\beta^{*}}, \boldsymbol{\theta^{*}})$

5: Solve teh Lagrangian dual problem again with $(\boldsymbol{\beta^{*}}, \boldsymbol{\theta^{*}})$

Ensure: $\{t^{ul*}_{u,S}, t^{ul*}_{u}, t^{dl*}_{u,S}, t^{dl*}_{u}, D^{L*}_{u}, D^{I*}_{u,S}, t^{dl*}_{S}\}$

scheme, local execution scheme and full offloading scheme, which are described in detail as follows.

- **Fixed Frequency** In order to address the non-convexity of the original optimisation problem, it can be seen that the alternating optimisation method is applied in the section 5.3.3.3 to find the solution that is very near to the optimal solution. It is achieved by fixing one primal variable, either D^{L}_{u} or f_u, at one time alternatively. The results of this baseline algorithm is obtained by fixing the allocated computing capabilities f_u to each u in cloudlet server. It is assumed that the CPU cycles are equally distributed to compute the allocated task bits from every mobile user.

- **Neglecting Shared Data** In this scenario, the collaborative property is ignored, every user makes the offloading decision without coordination among other users. The number of shared data bits is assumed to be zero, which is $D^{I}_{S} = 0$ in the optimisation problem. However, this assumption does not deny the de facto existence of the shared data. The shared data in different users are regarded as their exclusive

data, causing repetitive allocation of communication and computation resources for transmitting and computing the same data bits.

- **Local Execution** Local execution exploits the local computing capabilities of all u in mobile users side. The shared data are still undergoing cooperative offloading towards cloudlet server for remote computation. But the individual task data $D_u^I - D_S^I$ all remain at the mobile user to which they are exclusive. The computing power of the local CPU chips is exploited to complete the task execution within imposed time length constraint.

- **Full Offloading** Contrary to the local execution scheme, full offloading exploits that computing capabilities of the cloudlet server to the largest extent. Except for the cooperative offloading of the shared data, the individual task bits are not under optimisation in this case. Instead, it is assumed that all of them would be offloaded for remote computing, leaving the edge computing capabilities f_u to be optimised to complete the task execution within latency constraint.

In simulations, the mobile users are uniformly distributed in a geographical area over [0, 1000]m away from the BS with built-in cloudlet server, which is located in the center. The bandwidth available for uplink and downlink transmission are both set as $W^{ul} = W^{dl}$ =10MHz. The input data size for mobile users' AR/VR application is $D_u^I = 10kbits, \forall u \in \mathcal{U}$. The capacitance coefficients are set to be $\kappa_0 = 10^{-26}$. The energy expenditure per second in the downlink is ρ_u^{dl} =0.625 J/s [BBV09]. The maximum local computing capability is $f_{u,max}$ = 1GHz for all mobile users. The CPU cycles required to process every bit of input data is $\lambda_0 = 1 \times 10^3$ cycle/bit. For every bit of input data, there is one bit of corresponding output data, which means that

$a_0 = 1$. The pathloss model related to the transmission distance is given by $128.1 + 37.6log10(d_u)$ in dB, where d_u denotes the distance between the mobile user u and the BS in kilometers. The spectral density of the Additive White Gaussian Noise is -169 dBm/Hz.

5.3.5.1 Simulation Results for the General Case

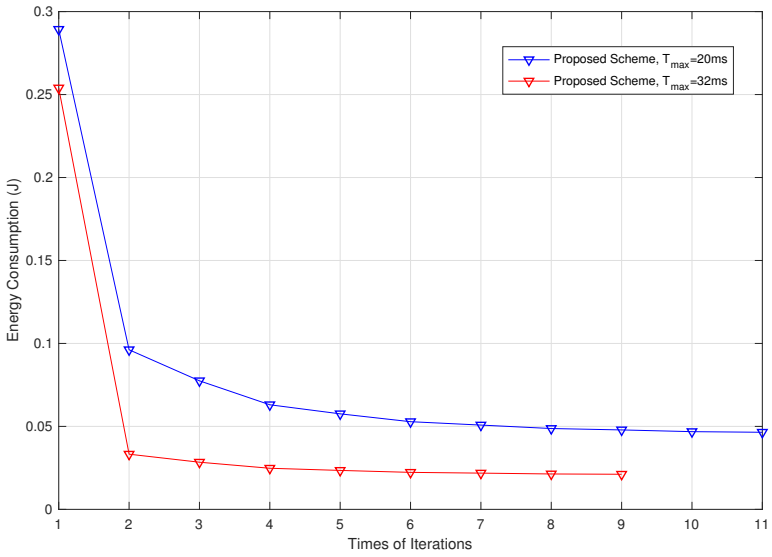


Figure 5.5: The illustration of convergence of the proposed algorithm.

The convergence of the proposed algorithm based on the alternative optimisation scheme is examined firstly. Given specific stopping criterion, the convergence speed under different latency constraints varies. As shown in Figure 5.5, when T_{max} =20ms. it takes 11 iterations to reach the stopping criterion and converges. While it takes 9 iterations to converge when T_{max} =32ms. The dramatic decreases after the first iterations are obvious in both cases, which proves that allocating cloudlet computing capabilities equally to each mobile user's offloaded computation task is not an energy

efficient solution. Further optimisation on allocation of cloudlet computing capabilities is necessary.

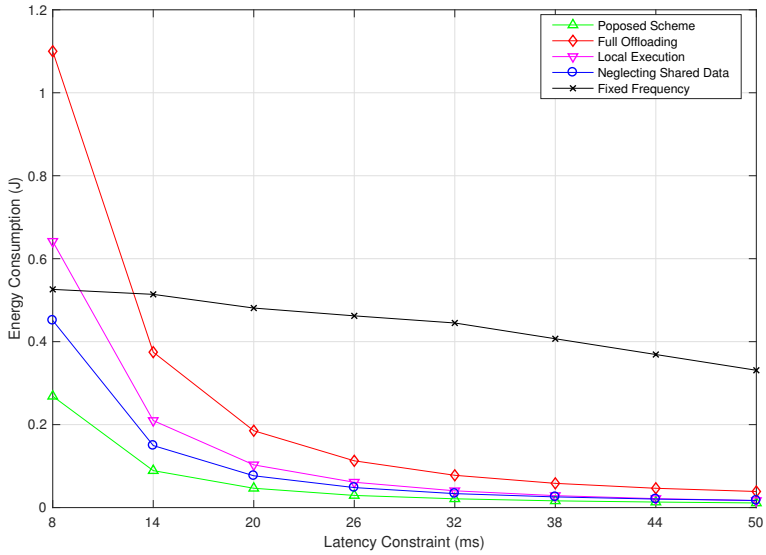
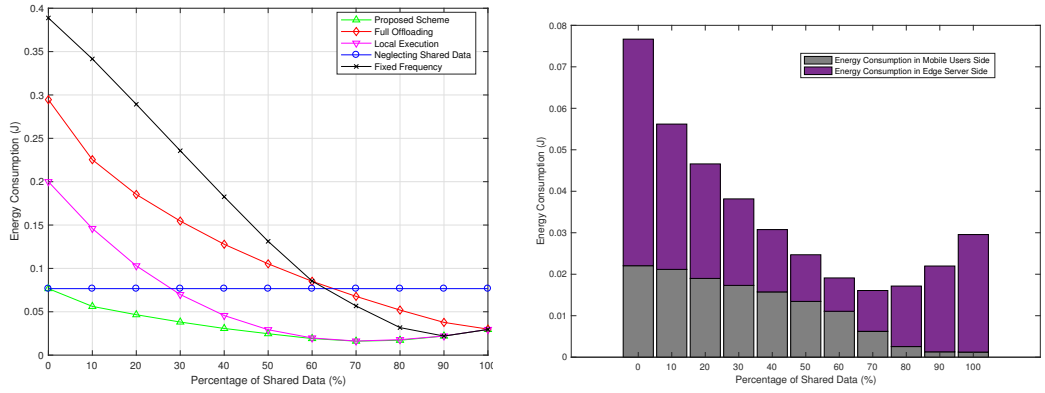


Figure 5.6: The overall energy consumption versus varied latency constraints.

The impact of imposed latency constraints on the energy consumption is presented in Figure 5.6, which shows the sum energy consumption versus different latency constraints assuming equal weights on mobile users side and BS side, $\Omega_1 = \Omega_2$ =0.5, and 20 percent of shared data as $\epsilon = 0.2$. There are 8 mobile users deployed in the simulation area. It is clear that although all algorithms give less energy consumption with the relaxed longer latency constraint ,the proposed energy efficient computation offloading and communication resources allocation scheme considering collaborative property of shared data outperforms all other baseline algorithms in terms of energy consumption. "Fixed Frequency" is the scheme without optimizing the cloudlet computing capabilities allocation. Except for the very stringent latency constraint such as 8ms, "Fixed Frequency" scheme gives the highest energy consumption compared to all other algorithms, which indicates the importance of

cloudlet computing capabilities allocation. Excluding the "Fixed Frequency" scheme, "Full Offloading" and "Local Execution" schemes give highest and second highest energy consumption. The result of "Full Offloading" manifests that it is not an energy efficient solution to offloads all input data to the cloudlet server for remote execution. Huge amount of offloaded data will take longer transmission time in both uplink and downlink, in turn requiring increased CPU frequencies for faster remote computation which introduces higher energy expenditure. The "Local Execution" keeps all individual input data locally in the mobile users. The local computing capabilities are exploited to its highest extend, which is not a energy efficient solution as well. "Neglecting Shared Data" wasted some energy because of repetitive allocation of communication and computation resources as mention previously. In a word, the best energy saving improvement can only be achieved through the joint participation of local computation and remote computation, with the shared data being offloaded by the coordination among mobile users.



(a) The overall energy consumption versus different percentages of shared data.

(b) The overall energy consumption distribution between two sides under different percentages of shared data.

Figure 5.7: The energy consumption illustration versus different percentages of shared data.

Figure 5.7a shows the energy consumption versus different percentage of

shared data ranging from 0 to 100 percent. There are 8 mobile users in the simulation, $\Omega_1 = \Omega_2 = 0.5$, and the latency constraint is $T_{max} = 20$ms. The best energy consumption performance of the proposed algorithm compared to other baseline schemes is obvious in the first half of the figure. However, The overall energy consumption of the proposed scheme is not monotonically decreasing with higher proportion of shared data. In the algorithm design, it is assumed that the shared data is always offloaded to the cloudlet server for remote execution. Recalling the computation energy consumption model (5.10), the energy consumption increases with the computing frequency to the second power. In addition, more shared data bits requires longer transmission time between mobile users and AP, which squeezes available time for remote computation. Consequently, the overall energy consumption of the proposed scheme undergoes an slight increasing at the last few points of the line. It is presented that when the percentage approaches 60% onwards, the "Local Execution" scheme almost overlaps with the proposed optimal scheme. Higher proportion of shared data reduces the amount of the rest input data subject to offloading optimisation. As a result, when the percentage of shared data exceeds certain point, keeping all the rest input data for local computing becomes an energy efficient choice, which explains the overlapping of the two lines after 60% of shared data in the figure. In the extreme case where 100% of input data are shard among all users, there is no individual data for local computing, all input data are fully offloaded to the cloudlet server, and the cloudlet computing capabilities are not subject for allocation to compute individual data bits. In this case, "Local Execution", "Full Offloading", "Fixed Frequency" and the proposed scheme are equivalent to each other. In Figure 5.7a, it can be observed that the four lines merge into one point in this extreme case.

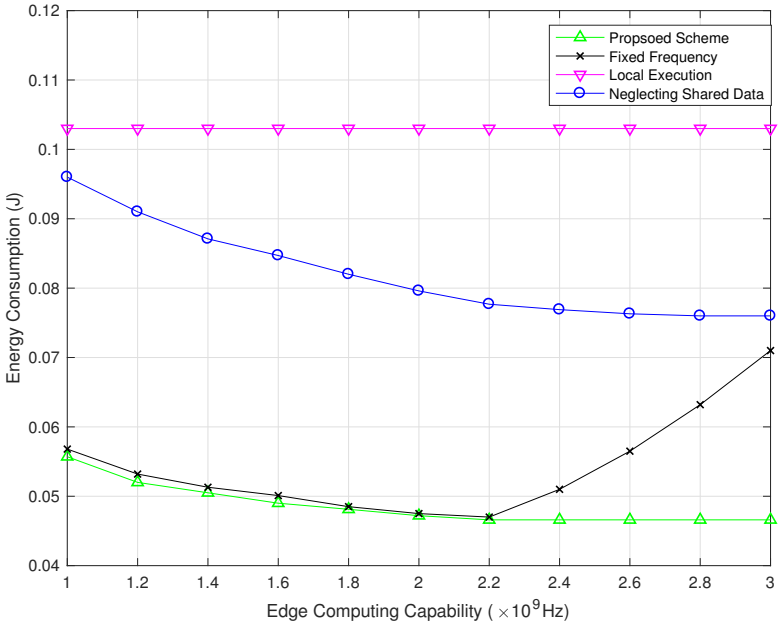
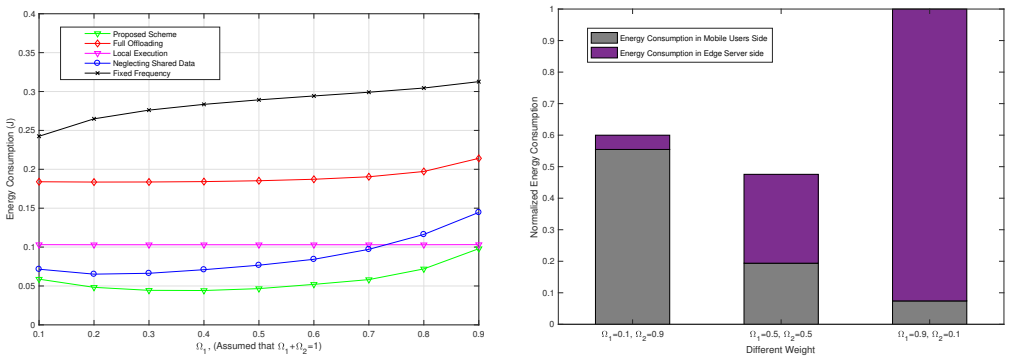


Figure 5.8: Energy consumption versus different edge computing capacities.

The effect of edge computing capacity on the energy consumption is revealed in Figure 5.8. Note that there is no "Full Offloading" scheme presented in this figure. The reason is that under the given system parameters if all the input data bits of every mobile user are offloaded, there requires more than 6GHz computing capabilities in the cloudlet server to satisfy the latency constraint according to the simulation results. "Local Execution" scheme gives flat line since all the individual data are remained in the mobile users where edge computing capabilities exercise no effect. "Neglecting Shared Data" scheme and the proposed scheme both give decreasing lines as the capabilities enhanced. It is easy to understand that higher the CPU cycles per second in the cloudlet server, the more computing capabilities the edge nodes can lend to the mobiles users so that they can offload more input data bits for remote computation to save energy consumption. There are two eye-

catching features of the "Fixed Frequency" plot. In the points before 2.2GHz, the energy consumption goer lower as the computing capacity increases, and almost overlap the with the proposed scheme plot in 2.2GHz point. After that, the trend turned adverse. In the simulation, the sum computing capability given by the proposed computing resource allocation algorithm for processing individual data offloaded by all mobiles users is around 2.18GHz. The edge computing capability less than this value would prevent the mobile users from reaping the advantage of remote computing to the largest extent by offloading their input data. Through the further optimisation of remote computing resources allocation, the proposed scheme gives slightly better energy consumption performance than "Fixed Frequency" scheme in these points. After the turning point 2.2GHz, spare computing capability exploited by "Fixed Frequency" scheme causes higher energy consumption, which is reflected in the increasing curve in Figure 5.8.



(a) Energy consumption versus different weights.

(b) Energy consumption distribution between two sides under different weights.

Figure 5.9: The energy consumption illustration versus different weights.

Both of the figures above illustrate the effects of different factors on the energy consumption based on equal weight weights put on both sides, $\Omega_1 = \Omega_2 = 0.5$. It is natural to wonder what results would be obtained if the weights put on mobile users side and the BS side are changed. Figure 5.9a depicts

the energy consumptions of all algorithms with changed weights based on T_{max} =20ms, and 20% of shared data. "Local Execution" scheme gives a flat line since there is no computing resources allocation needed in the cloudlet server. The line representing the results of the proposed scheme gives an "U" shape, decreasing with larger Ω_1 in the first few points and then increasing with more weight put on mobile users side. It is illustrated in Figure 5.9b that put too much weight on either AP side or mobile users side would result in huge amount of energy consumption of the other. The increased energy consumption under unbalanced weight is very obvious when most weight is put on mobile users side, which incurs large amount of energy consumption in AP since the mobile users may extend the offloading transmission length at the cost of more computing resources exploited in the cloudlet server to meet the latency constraint.

E_{total} (J)	0.940	0.509	0.403	0.361	0.340	0.329	0.322	0.317
T_{max} (ms)	8	14	20	26	32	38	44	50

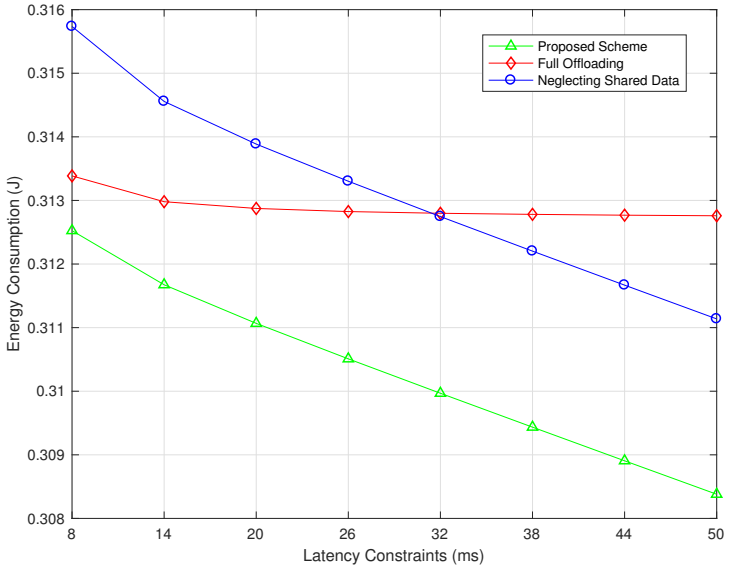


Figure 5.10: The total energy consumption versus the latency constraint in negligible computation time.

5.3.5.2 Simulation Results for Negligible Computing Latency

Then it comes to the energy consumption of the negligible computation time, it is assumed that the energy consumption in remote computing is constant as $0.3J$. The advantage of making this assumption is that the energy-efficient offloading and resource allocation algorithm can give lower complexity since there is no need for alternating optimisation any more. The results of "Local Computing" scheme are given in a table above the plot since it is much larger then other scenarios, especially in low latency constraint cases. This proves the inefficiency of the "Local Computing" scheme in terms of energy consumption. The proposed scheme considering shared data still gives the lowest energy consumption.

E_{total} (10^{-3}J)	500.0	446.0	402.7	369.1	343.9	325.9	313.8	306.6	303.0	301.8	301.7
ϵ (%)	0	10	20	30	40	50	60	70	80	90	100

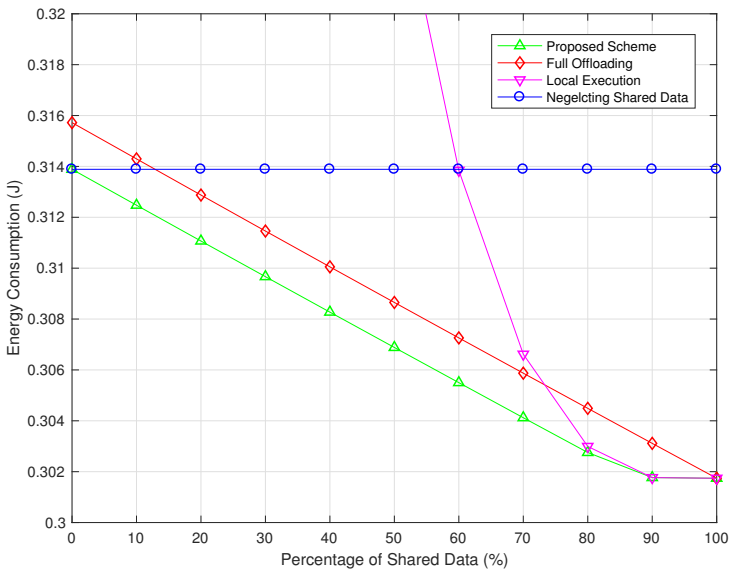


Figure 5.11: The total energy consumption versus the percentage of shared data in negligible computation time.

Unlike the simulation results for the non-negligible computation time, the energy consumption of all the presented schemes are monotonically decreas-

ing as the percentage of the shared data increases. The reason lies in the fact that through assuming constant remote computation energy, only the energy expenses for data transmission in both mobile users and the AP sides are subject to optimisation. In this context, higher percentage of shared data means less amount of transmitted data, which results in lower overall energy consumption.

5.4 Summary

In this chapter, a multi-user fog computing system was considered, in which multiple single-antenna mobile users running applications featuring shared data can partially offload their individual computation tasks to a nearby single-antenna cloudlet and then download the results from it. The mobile users' energy consumption minimisation problem subject to the total latency, the total downlink transmission energy and the local computing constraints was formulated as a convex problem with the optimal solution obtained by classical Lagrangian duality method. Based upon the semi-closed form solution, it was proved that the shared data is optimally transmitted by only one of the mobile users instead of multiple ones collaboratively. The proposed joint computation offloading and communications resource allocation was verified by simulations against other baseline algorithms that ignore the shared data property or the mobile users' own computing capabilities.

Chapter 6

Conclusions and Future Work

This chapter summarises the conclusions drawn from this thesis and presents the research directions of the future work.

6.1 Conclusions

This thesis is dedicated to the energy efficient resource allocation schemes for different scenarios under novel radio access networks architectures.

In hybrid energy source powered heterogeneous cloud radio access networks, a joint user association and resource allocation scheme is proposed in Chapter 3. The optimal user association and power allocation policy is obtained via resolving the formulated optimisation problem. Based on the numerical results analysis, the proposed algorithm can greatly reduce the system's overall grid power consumption with the help of energy harvesting technology compared to other algorithms. In addition, another policy is designed for maximising the grid power utility, which is defined as the ratio of the overall data throughput over the consumed grid power.

In hybrid energy source powered F-RAN, a delay-aware energy-efficient computation offloading algorithm is presented in Chapter 4. Through enabling the forwarding of the computation tasks offloaded by mobile users

from their corresponding serving F-AP to its neighbouring F-APs, the proposed algorithm can better improve the utilisation of renewable energy supply compared to existing algorithms. Numerical results reveal that such energy saving performance can be achieved without violating imposed latency constraint.

The computation offloading scheme where mobile users run applications featuring intrinsic collaborative properties is investigated in Chapter 5. Unlike conventional computation offloading schemes, the proposed algorithm takes the shared input and output data into consideration. The shared input data can be jointly offloaded and processed with the help of coordination among mobile users. It is proved that with the help of mobile users' local computing capabilities, the fog-computing based system considering shared input data can better save the energy consumption compared to all other algorithms.

For all algorithms proposed in this thesis, great attention is given to reduce the energy consumption on satisfying users' service requirements. The proposed schemes provide useful guidelines and potential solutions for future greener mobile networks in terms of energy saving.

6.2 Future Work

In this section, extensions to current work and some future research directions are proposed.

6.2.1 Energy Efficient Resource Allocation Scheme for Fronthaul-Constrained H-CRAN and F-RAN

In H-CRAN architectures, there are fronthaul connections between access points (APs) to the cloud center [PCSD15]. Non-ideal wireless fronthaul or capacity-constrained optical fiber are envisioned to be the prominent fronthaul solutions to enhance the feasibility of H-CRAN deployment [PWLP15b]. The deficient capacity of non-ideal fronthaul links hinders APs from making full use of available radio resources in providing high data rate services. Regardless of what is the theoretically achievable data rate supported by the air interface, it is the fronthaul capacity that defines the practical service rate.

In the future work, the fronthaul-aware user association and resource allocation will be investigated. The UA/RA algorithm suitable for RAN with constrained fronthaul will be proposed, and the influence of the varied fronthaul capacity on the supported data throughput will be presented.

6.2.2 MEC Computation Offloading for Applications Featuring Shared Data in Multi-cell Scenario

In this thesis, the computation offloading for applications featuring shared data in single-cell scenario is presented in Chapter 6. It was revealed that the proposed algorithm achieves much improvement in terms of energy consumption in that single-cell case. In the the previous Chapter, the investigation on computation offloading for multi-cell MEC system with green energy supply reveals that in such scenario, the existence of multiple offloading path enables the further reduction of non-renewable energy dissipation.

In the future work, computational resources will be allocated to the offloaded computation tasks from mobile users in a coordinated manner among

cooperating edge computing nodes. The shared data in mobile users makes the coordination not so straightforward as that in investigated case, since intuitively the shared input-bits are better to be aggregated in one computing node. Further conclusions need to be drew through deeper researches.

6.2.3 Computation Optimisation in Three-layer User-Fog-Cloud Scenarios

The practical deployment of the computation offloading system may include fog servers and cloud servers at that same time, which forms a three-layer user-fog-cloud structure. In this case, the fog-computing server may act as an intermediate layer between the mobile users and the cloud-computing servers. The computation tasks offloaded from the mobile users to the fog-computing nodes can be either executed by the fog server or be further offloaded to the cloud-computing nodes. This strategy is of practical importance especially for the case where there exists shared input-bits for the computation tasks as that of Chapter 5.

The participation of cloud-computing nodes in processing shared input-bits frees up the computing capabilities of fog-computing nodes so that the shared input-bits and individual input-bits can be computed simultaneously. However, imperfect interconnections and other factors may affect the performance of this strategy. In the future work, the optimal computation offloading strategy to maximise the computation offloading performance under specific constraints can be investigated.

References

[AAA+19] M. Awais, A. Ahmed, S. A. Ali, M. Naeem, W. Ejaz, and A. Anpalagan. Resource management in multicloud iot radio access network. *IEEE Internet of Things Journal*, 6(2):3014–3023, April 2019.

[AGD+11] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske. How much energy is needed to run a wireless network? *IEEE Wireless Communications*, 18(5):40–49, October 2011.

[AJ16] A. Abrol and R. K. Jha. Power optimization in 5g networks: A step towards green communication. *IEEE Access*, 4:1355–1374, 2016.

[ASS17] A. Al-Shuwaili and O. Simeone. Energy-efficient resource allocation for mobile edge computing-based augmented reality applications. *IEEE Wireless Communications Letters*, 6(3):398–401, June 2017.

[BBV09] Niranjan Balasubramanian, Aruna Balasubramanian, and Arun Venkataramani. Energy consumption in mobile phones:

A measurement study and implications for network applications. pages 280–293, 01 2009.

[BdlOS+14] C. J. Bernardos, A. de la Oliva, P. Serrano, A. Banchs, L. M. Contreras, H. Jin, and J. C. Zuniga. An architecture for software defined wireless networking. *IEEE Wireless Communications*, 21(3):52–61, June 2014.

[BH03] James C. Bezdek and Richard J. Hathaway. Convergence of alternating optimization. *Neural, Parallel Sci. Comput.*, 11(4):351–368, December 2003.

[BHZ15] S. Bi, C. K. Ho, and R. Zhang. Wireless powered communication: opportunities and challenges. *IEEE Communications Magazine*, 53(4):117–125, April 2015.

[BLM+14] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer. Network densification: the dominant theme for wireless evolution into 5g. *IEEE Communications Magazine*, 52(2):82–89, February 2014.

[BMZA12] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli. Fog computing and its role in the Internet of Things. In *Proc. ACM SIGCOMM Workshop on Mobile Cloud Computing (MCC)*, Helsinki, Finland, Aug. 2012.

[Boy] Stephen Boyd. Lecture notes for EE364b: Convex Optimization II.

[BSD13] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo. Joint allocation of computation and communication resources in mul-

tiuser mobile cloud computing. In *2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 26–30, June 2013.

[CCY+15] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann. Cloud ran for mobile networks - a technology overview. *IEEE Communications Surveys Tutorials*, 17(1):405–426, Firstquarter 2015.

[Che15] X. Chen. Decentralized computation offloading game for mobile cloud computing. *IEEE Transactions on Parallel and Distributed Systems*, 26(4):974–983, April 2015.

[Chi11] C-ran the road towards green ran. Technical report, China Mobile Research Institute, Beijing, China, Oct. 2011.

[CSBC16] J. Cheng, Y. Shi, B. Bai, and W. Chen. Computation offloading in cloud-RAN based mobile cloud computing system. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2016.

[CSYD17] M. Chen, W. Saad, C. Yin, and M. Debbah. Echo state transfer learning for data correlation aware resource allocation in wireless virtual reality. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 1852–1856, Oct 2017.

[DANA15] H. Dahrouj, T. Y. Al-Naffouri, and M. Alouini. Distributed cloud association in downlink multicloud radio access networks. In *2015 49th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–3, March 2015.

[DDANA15] A. Douik, H. Dahrouj, T. Y. Al-Naffouri, and M. Alouini. Coordinated scheduling for the downlink of cloud radio-access networks. In *2015 IEEE International Conference on Communications (ICC)*, pages 2906–2911, June 2015.

[DDD+15] H. Dahrouj, A. Douik, O. Dhifallah, T. Y. Al-Naffouri, and M. S. Alouini. Resource allocation in heterogeneous cloud radio access networks: advances and challenges. *IEEE Wireless Communications*, 22(3):66–73, June 2015.

[DMW+11] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi. A survey on 3gpp heterogeneous networks. *IEEE Wireless Communications*, 18(3):10–21, June 2011.

[GLN+12] Iñigo Goiri, Kien Le, Thu Nguyen, Jordi Guitart, Jordi Torres, and Ricardo Bianchini. Greenhadoop: Leveraging green energy in data-processing frameworks. In *EuroSys'12 - Proceedings of the EuroSys 2012 Conference*, 04 2012.

[GZQL12] Yang Ge, Yukan Zhang, Qinru Qiu, and Yung-Hsiang Lu. A game theoretic resource allocation for overall energy minimization in mobile cloud computing system. *Proceedings of the International Symposium on Low Power Electronics and Design*, 07 2012.

[HA13] T. Han and N. Ansari. Green-energy aware and latency aware user associations in heterogeneous cellular networks. In *2013 IEEE Global Communications Conference (GLOBECOM)*, pages 4946–4951, Dec 2013.

[HBB11] Z. Hasan, H. Boostanimehr, and V. K. Bhargava. Green cellular networks: A survey, some research issues and challenges. *IEEE Communications Surveys Tutorials*, 13(4):524–540, Fourth 2011.

[HSS13] I. Hwang, B. Song, and S. S. Soliman. A holistic view on hyper-dense heterogeneous and small cell networks. *IEEE Communications Magazine*, 51(6):20–27, June 2013.

[IRH+14] C. L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan. Toward green and soft: a 5g perspective. *IEEE Communications Magazine*, 52(2):66–73, February 2014.

[JYM+14] Jian Song, Yong Cui, Minming Li, Jiezhong Qiu, and R. Buyya. Energy-traffic tradeoff cooperative offloading for mobile cloud computing. In *2014 IEEE 22nd International Symposium of Quality of Service (IWQoS)*, pages 284–289, May 2014.

[KBTV10] A. Khandekar, N. Bhushan, J. Tingfang, and V. Vanghi. Lte-advanced: Heterogeneous networks. In *2010 European Wireless Conference (EW)*, pages 978–982, April 2010.

[KGH03] D. Kivanc, Guoqing Li, and Hui Liu. Computationally efficient bandwidth allocation and power control for ofdma. *IEEE Transactions on Wireless Communications*, 2(6):1150–1158, Nov 2003.

[KH00] D. Kivanc and Hui Liu. Subcarrier allocation and power control for ofdma. In *Conference Record of the Thirty-Fourth Asilomar Conference on Signals, Systems and Computers (Cat. No.00CH37154)*, volume 1, pages 147–151 vol.1, Oct 2000.

[KKLC15] J. Kwak, Y. Kim, J. Lee, and S. Chong. Dream: Dynamic resource and task allocation for energy minimization in mobile cloud systems. *IEEE Journal on Selected Areas in Communications*, 33(12):2510–2523, Dec 2015.

[KLL+17] Y. J. Ku, D. Y. Lin, C. F. Lee, P. J. Hsieh, H. Y. Wei, C. T. Chou, and A. C. Pang. 5g radio access network design with the fog paradigm: Confluence of communications and computing. *IEEE Communications Magazine*, 55(4):46–52, April 2017.

[KMM95] R. Kohno, R. Meidan, and L. B. Milstein. Spread spectrum access methods for wireless communications. *IEEE Communications Magazine*, 33(1):58–67, Jan 1995.

[KNWH13] R. Kaewpuang, D. Niyato, P. Wang, and E. Hossain. A framework for cooperative resource management in mobile cloud computing. *IEEE Journal on Selected Areas in Communications*, 31(12):2685–2700, December 2013.

[LCC+15] D. Liu, Y. Chen, K. K. Chai, T. Zhang, and M. Elkashlan. Two-dimensional optimization on user association and green energy allocation for hetnets with hybrid energy sources. *IEEE Transactions on Communications*, 63(11):4111–4124, Nov 2015.

[LCLH15] Y. D. Lin, E. T. H. Chu, Y. C. Lai, and T. J. Huang. Time-and-energy-aware computation offloading in handheld devices to coprocessors and clouds. *IEEE Systems Journal*, 9(2):393–405, June 2015.

[LMZL16] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief. Delay-optimal computation task scheduling for mobile-edge computing sys-

tems. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1451–1455, July 2016.

[LWC+15] D. Liu, L. Wang, Y. Chen, T. Zhang, K. K. Chai, and M. Elkashlan. Distributed energy efficient fair user association in massive mimo enabled hetnets. *IEEE Communications Letters*, 19(10):1770–1773, Oct 2015.

[LZL15] S. Luo, R. Zhang, and T. J. Lim. Downlink and uplink energy minimization through user association and beamforming in c-ran. *IEEE Transactions on Wireless Communications*, 14(1):494–508, Jan 2015.

[MHLB08] G. Miao, N. Himayat, Y. Li, and D. Bormann. Energy efficient design in wireless ofdma. In *2008 IEEE International Conference on Communications*, pages 3307–3312, May 2008.

[MKGM+15] M. A. Marotta, N. Kaminski, I. Gomez-Miguelez, L. Z. Granville, J. Rochol, L. DaSilva, and C. B. Both. Resource sharing in heterogeneous cloud radio access networks. *IEEE Wireless Communications*, 22(3):74–82, June 2015.

[MLZL15] Y. Mao, Y. Luo, J. Zhang, and K. B. Letaief. Energy harvesting small cell networks: feasibility, deployment, and operation. *IEEE Communications Magazine*, 53(6):94–101, June 2015.

[MOP+14] M. Matinmikko, H. Okkonen, M. Palola, S. Yrjola, P. Ahokangas, and M. Mustonen. Spectrum sharing using licensed shared access: the concept and its workflow for lte-advanced networks. *IEEE Wireless Communications*, 21(2):72–79, April 2014.

[MSS15] S. E. Mahmoodi, K. P. Subbalakshmi, and V. Sagar. Cloud offloading for multi-radio enabled mobile devices. In *2015 IEEE International Conference on Communications (ICC)*, pages 5473–5478, June 2015.

[MYZ+17] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief. A survey on mobile edge computing: The communication perspective. 19(4):2322–2358, Fourth Quarter 2017.

[MZSL17] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief. Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems. *IEEE Transactions on Wireless Communications*, 16(9):5994–6009, Sep. 2017.

[OBB+14] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren. Scenarios for 5g mobile and wireless communications: the vision of the metis project. *IEEE Communications Magazine*, 52(5):26–35, May 2014.

[OSSB15] J. Oueis, E. C. Strinati, S. Sardellitti, and S. Barbarossa. Small cell clustering for efficient distributed fog computing: A multi-user case. In *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, pages 1–5, Sept 2015.

[OTUY15] O. Ozel, K. Tutuncuoglu, S. Ulukus, and A. Yener. Fundamental limits of energy harvesting communications. *IEEE Communications Magazine*, 53(4):126–132, April 2015.

[PCSD15] A. Pizzinat, P. Chanclou, F. Saliou, and T. Diallo. Things you should know about fronthaul. *Journal of Lightwave Technology*, 33(5):1077–1083, March 2015.

[PE13] T. Pamuklu and C. Ersoy. Optimization of renewable green base station deployment. In *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, pages 59–63, Aug 2013.

[PLJ+14] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang. Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies. *IEEE Wireless Communications*, 21(6):126–135, December 2014.

[PLZW15] M. Peng, Y. Li, Z. Zhao, and C. Wang. System architecture and key technologies for 5g heterogeneous cloud radio access networks. *IEEE Network*, 29(2):6–14, March 2015.

[PPS18] J. Park, P. Popovski, and O. Simeone. Minimizing latency to support VR social interactions over wireless cellular systems via bandwidth allocation. *IEEE Wireless Communications Letters*, pages 1–1, 2018.

[PSSS13] S. Park, O. Simeone, O. Sahin, and S. Shamai. Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks. *IEEE Transactions on Signal Processing*, 61(22):5646–5658, Nov 2013.

[PSSS14] S. Park, O. Simeone, O. Sahin, and S. Shamai. Inter-cluster design of precoding and fronthaul compression for cloud ra-

dio access networks. *IEEE Wireless Communications Letters*, 3(4):369–372, Aug 2014.

[PWLP15a] M. Peng, C. Wang, V. Lau, and H. V. Poor. Fronthaul-constrained cloud radio access networks: insights and challenges. *IEEE Wireless Communications*, 22(2):152–160, April 2015.

[PWLP15b] M. Peng, C. Wang, V. Lau, and H. V. Poor. Fronthaul-constrained cloud radio access networks: insights and challenges. *IEEE Wireless Communications*, 22(2):152–160, April 2015.

[PXC+15] M. Peng, H. Xiang, Y. Cheng, S. Yan, and H. V. Poor. Inter-tier interference suppression in heterogeneous cloud radio access networks. *IEEE Access*, 3:2441–2455, Nov 2015.

[PYZW16] M. Peng, S. Yan, K. Zhang, and C. Wang. Fog-computing-based radio access networks: issues and challenges. *IEEE Network*, 30(4):46–53, July 2016.

[PZJ+15] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang. Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks. *IEEE Transactions on Vehicular Technology*, 64(11):5275–5287, Nov 2015.

[Sch04] Mischa Schwartz. Mobile wireless communications, 2004.

[SML+18] Z. Sheng, C. Mahapatra, V. C. M. Leung, M. Chen, and P. K. Sahu. Energy efficient cooperative computing in mobile wireless sensor networks. *IEEE Transactions on Cloud Computing*, 6(1):114–126, Jan 2018.

[SPM19] Y. Sun, M. Peng, and S. Mao. Deep reinforcement learning-based mode selection and resource management for green fog radio access networks. *IEEE Internet of Things Journal*, 6(2):1960–1971, April 2019.

[SSB15] S. Sardellitti, G. Scutari, and S. Barbarossa. Joint optimization of radio and computational resources for multicell mobile-edge computing. *IEEE Transactions on Signal and Information Processing over Networks*, 1(2):89–103, June 2015.

[SY14] K. Shen and W. Yu. Distributed pricing-based user association for downlink heterogeneous cellular networks. *IEEE Journal on Selected Areas in Communications*, 32(6):1100–1113, June 2014.

[SZL14] Y. Shi, J. Zhang, and K. B. Letaief. Group sparse beamforming for green cloud-ran. *IEEE Transactions on Wireless Communications*, 13(5):2809–2823, May 2014.

[TL17] N. T. Ti and L. B. Le. Computation offloading leveraging computing resources from edge cloud and mobile peers. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2017.

[TSA+15] J. Tang, D. K. C. So, E. Alsusa, K. A. Hamdi, and A. Shojaeifard. Resource allocation for energy efficiency optimization in heterogeneous networks. *IEEE Journal on Selected Areas in Communications*, 33(10):2104–2117, Oct 2015.

[WGKN08] R. Wolski, S. Gurun, C. Krintz, and D. Nurmi. Using bandwidth data to make computation offloading decisions. In *2008*

IEEE International Symposium on Parallel and Distributed Processing, pages 1–8, April 2008.

[WXD17] F. Wang, J. Xu, and Z. Ding. Optimized multiuser computation offloading with multi-antenna noma. In *2017 IEEE Globecom Workshops (GC Wkshps)*, pages 1–7, Dec 2017.

[WXWC18] F. Wang, J. Xu, X. Wang, and S. Cui. Joint offloading and computing optimization in wireless powered mobile-edge computing systems. *IEEE Transactions on Wireless Communications*, 17(3):1784–1797, March 2018.

[WZL12] Y. Wen, W. Zhang, and H. Luo. Energy-optimal mobile application execution: Taming resource-poor mobile devices with cloud clones. In *2012 Proceedings IEEE INFOCOM*, pages 2716–2720, March 2012.

[XCE+16] B. Xu, Y. Chen, M. Elkashlan, T. Zhang, and K. Wong. User association in massive mimo and mmwave enabled hetnets powered by renewable energy. In *2016 IEEE Wireless Communications and Networking Conference*, pages 1–6, April 2016.

[XH12] Xue Chen and R. Q. Hu. Joint uplink and downlink optimal mobile association in a wireless heterogeneous network. In *2012 IEEE Global Communications Conference (GLOBECOM)*, pages 4131–4137, Dec 2012.

[XLXN18] Hong Xing, Liang Liu, Jie Xu, and Arumugam Nallanathan. Joint task assignment and wireless resource allocation for cooperative mobile-edge computing. 02 2018.

[XLXN19] H. Xing, L. Liu, J. Xu, and A. Nallanathan. Joint task assignment and resource allocation for d2d-enabled mobile-edge computing. *IEEE Transactions on Communications*, 67(6):4193–4207, June 2019.

[YA18] J. Yao and N. Ansari. Qos-aware joint bbu-rrh mapping and user association in cloud-rans. *IEEE Transactions on Green Communications and Networking*, 2(4):881–889, Dec 2018.

[YCL09] J. Yeh, J. Chen, and C. Lee. Comparative analysis of energy-saving techniques in 3gpp and 3gpp2 systems. *IEEE Transactions on Vehicular Technology*, 58(1):432–448, Jan 2009.

[YHC16] C. You, K. Huang, and H. Chae. Energy efficient mobile cloud computing powered by wireless energy transfer. *IEEE Journal on Selected Areas in Communications*, 34(5):1757–1771, May 2016.

[YHCK17] C. You, K. Huang, H. Chae, and B. H. Kim. Energy-efficient resource allocation for mobile-edge computation offloading. *IEEE Transactions on Wireless Communications*, 16(3):1397–1411, March 2017.

[YWJ+16] Z. Yu, K. Wang, H. Ji, X. Li, and H. Zhang. Dynamic resource allocation in tdd-based heterogeneous cloud radio access networks. *China Communications*, 13(6):1–11, June 2016.

[ZCG+15] N. Zhang, N. Cheng, A. T. Gamage, K. Zhang, J. W. Mark, and X. Shen. Cloud assisted hetnets toward 5g wireless networks. *IEEE Communications Magazine*, 53(6):59–65, June 2015.

[ZDO+16] Z. Zhou, M. Dong, K. Ota, G. Wang, and L. T. Yang. Energy-efficient resource allocation for d2d communications underlaying cloud-ran-based lte-a networks. *IEEE Internet of Things Journal*, 3(3):428–438, June 2016.

[ZQL13] J. Zhao, T. Q. S. Quek, and Z. Lei. Coordinated multipoint transmission with limited backhaul data transfer. *IEEE Transactions on Wireless Communications*, 12(6):2762–2775, June 2013.

[ZWG+13] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu. Energy-optimal mobile cloud computing under stochastic wireless channel. *IEEE Transactions on Wireless Communications*, 12(9):4569–4581, September 2013.

[ZXL+15] T. Zhang, H. Xu, D. Liu, N. C. Beaulieu, and Y. Zhu. User association for energy-load tradeoffs in hetnets with renewable energy supply. *IEEE Communications Letters*, 19(12):2214–2217, Dec 2015.

[ZY14] Y. Zhou and W. Yu. Optimized backhaul compression for uplink cloud radio access network. *IEEE Journal on Selected Areas in Communications*, 32(6):1295–1307, June 2014.

[ZZY+16] J. Zuo, J. Zhang, C. Yuen, W. Jiang, and W. Luo. Energy efficient user association for cloud radio access networks. *IEEE Access*, 4:2429–2438, June 2016.