Resource Allocation for Heterogeneous Wireless Networks

by

Amila Pradeep Kumara Tharaperiya Gamage

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2015

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Demand for high volumes of mobile data traffic with better quality-of-service (QoS) support and seamless network coverage is ever increasing, due to growth of the number of smart mobile devices and the applications that run on these devices. Also, most of these high volumes of data traffic demanding areas are covered by heterogeneous wireless networks, such as cellular networks and wireless local area networks (WLANs). Therefore, interworking mechanisms can be used in these areas to enhance the network capacity, QoS support and coverage. Interworking enhances network capacity and QoS support by jointly allocating resources of multiple networks and enabling user multi-homing, where multi-homing allows users to simultaneously communicate over multiple networks. It widens network coverage by merging coverage of individual networks. However, there are areas where interworking cannot improve network capacity or QoS support, such as the areas with coverage of only one networks. Therefore, to achieve network-wide uniform capacity and QoS support enhancements, interworking can be integrated with device-to-device (D2D) communication and small cell deployment techniques. One of the challenging issues that need to be solved before these techniques can be applied in practical networks is the efficient resource allocation, as it has a direct impact on the network capacity and QoS support. Therefore, this thesis focuses on studying and developing efficient resource allocation schemes for interworking heterogeneous wireless networks which apply D2D communication and small cell deployment techniques.

First, uplink resource allocation for cellular network and WLAN interworking to provide multi-homing voice and data services is investigated. The main technical challenge, which makes the resource allocation for this system complicated, is that resource allocation decisions need to be made capturing multiple physical layer (PHY) and medium access control layer (MAC) technologies of the two networks. This is essential to ensure that the decisions are feasible and can be executed at the lower layers. Thus, the resource allocation problem is formulated based on PHY and MAC technologies of the two networks. The optimal resource allocation problem is a multiple time-scale Markov decision process (MMDP) as the two networks operate at different time-scales, and due to voice and data service requirements. A resource allocation scheme consisting of decision policies for the upper and the lower levels of the MMDP is derived. To reduce the time complexity, a heuristic resource allocation algorithm is also proposed.

Second, resource allocation for D2D communication underlaying cellular network and

WLAN interworking is investigated. Enabling D2D communication within the interworking system further enhances the spectrum efficiency, especially at areas where only one network is available. In addition to the technical challenges encountered in the first interworking system, interference management and selection of users' communication modes for multiple networks to maximize hop and reuse gains complicate resource allocation for this system. To address these challenges, a semi-distributed resource allocation scheme that performs mode selection, allocation of WLAN resources, and allocation of cellular network resources in three different time-scales is proposed.

Third, resource allocation for interworking macrocell and hyper-dense small cell networks is studied. Such system is particularly useful for interference prone and high capacity demanding areas, such as busy streets and city centers, as it uses license frequency bands and provides a high spectrum efficiency through frequency reuse and bringing network closer to the users. The key challenge for allocating resources for this system is high complexity of the resource allocation scheme due to requirement to jointly allocate resources for a large number of small cells to manage co-channel interference (CCI) in the system. Further, the resource allocation scheme should minimize the computational burden for low-cost small cell base stations (BSs), be able to adapt to time-varying network load conditions, and reduce signaling overhead in the small cell backhauls with limited capacity. To this end, a resource allocation scheme which operates on two time-scales and utilizes cloud computing to determine resource allocation decisions is proposed. Resource allocation decisions are made at the cloud in a slow time-scale, and are further optimized at the BSs in a fast time-scale in order to adapt the decisions to fast varying wireless channel conditions. Achievable throughput and QoS improvements using the proposed resource allocation schemes for all three systems are demonstrated via simulation results.

In summary, designing of the proposed resource allocation schemes provides valuable insights on how to efficiently allocate resources considering PHY and MAC technologies of the heterogeneous wireless networks, and how to utilize cloud computing to assist executing a complex resource allocation scheme. Furthermore, it also demonstrates how to operate a resource allocation scheme over multiple time-scales. This is particularly important if the scheme is complex and requires a long time to execute, yet the resource allocation decisions are needed to be made within a short interval.

Acknowledgements

I would like to express my utmost appreciation to my PhD supervisor, Prof. Sherman (Xuemin) Shen, for his exemplary supervision, tremendous support, and valuable advice throughout my PhD program.

I sincerely would like to thank Prof. Weihua Zhuang, Prof. Jon W. Mark, and all my colleagues at the Broadband Communications Research (BBCR) group for the research collaboration, beneficial discussions, and continuous exchange of knowledge.

I gratefully acknowledge my PhD committee members, Prof. Liang-liang Xie, Prof. Oussama Damen, and Prof. Wei-Chau Xie, for their constructive comments and suggestions, which helped to improve the quality of the thesis.

I would like to thank Prof. Nandana Rajatheva, Centre for Wireless Communications, University of Oulu and Prof. Poompat Saengudomlert, Center of Research in Optoelectronics, Communications and Control Systems, Bangkok University for their valuable comments and suggestions throughout the PhD program. Also, I would like to thank Prof. Tan Ai Hui, Faculty of Engineering, Multimedia University and Dr. Sim Moh Lim, Motorola Solutions, Malaysia for providing supportive recommendation letters at the time of my application for a PhD program at the University of Waterloo.

I would like to express my deepest and heartfelt gratitude to my family and friends for their support and assistance.

This PhD thesis is dedicated to my parents.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3GPP	3rd Generation Partnership Project
AAA	Authorization, Authentication and Accounting
ABS	Almost Blank Subframes
A-GNSS	Assisted Global Navigation Satellite Systems
AP	Access Point
ASN	Access Service Network
BM1	Benchmark resource allocation algorithm-1
BM2	Benchmark resource allocation algorithm-2
BS	Base Station
CCI	Co-Channel Interference
CCS	Centralized Control Server
CDMA	Code Division Multiple Access
CFP	Contention-Free Period
CoA	Care of Address
CoMP	Coordinated Multi-point
CP	Contention Period
CSI	Channel State Information
CTS	Clear To Send
D2D	Device-to-Device
DCF	Distributed Coordination Function
DL	Downlink
DSL	Digital Subscribed Line
EAP	Extensible Authentication Protocol
eNB	Enhanced NodeB

EPC	Evolved Packet Core
ePDG	Evolved Packet Data Gateway
FDD	Frequency-Division Duplexing
FFR	Fractional Frequency Reuse
FSMC	Finite State Markov Channel
GAN	Generic Access Networks
GANC	Generic Access Networks Controller
HA	Home Agent
HCF	Hybrid Coordination Function
HM	Huristic resource allocation algorithm
ICI	Intercarrier Interference
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IPsec	Internet Protocol Security
ISP	Internet Service Provider
I-WLAN	Interworking-WLAN
L2TP	Layer 2 Tunneling Protocol
LPP	LTE Positioning Protocol
LTE	Long Term Evolution
LTE-A	LTE-Advanced
MAC	Medium Access Control Layer
MIH	Media Independent Handover
MIMO	Multipl-Input and Multiple Output
MIP	Mobile IP
MIPv6	Mobile IP Version 6
MM	Optimal MMDP based resource allocation algorithm
MMDP	Multiple time-scale Markov Decision Process
MMSE	Minimum Mean Squared Error
MRC	Maximal Ratio Combining
NTPv4	Network Time Protocol version 4
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OTDOA	Observed Time Difference Of Arrival

PDG	Packet Data Gateway
PDN	Packet Data Network
PDN-GW	PDN Gateway
PHY	Physical Layer
PMIP	Proxy Mobile IP
QoS	Quality of Service
RB	Resource Block
RTS	Request To Send
SAE	System Architecture Evolution
SDT	Sum of Discounted Throughputs
S-GW	Serving Gateway
SI	Satisfaction Index
SINR	Signal-to-Interference plus Noise Ratio
SIP	Session Initiation Protocol
TDD	Time-Division Duplexing
TXOP	Transmission Opportunity
UE	Users' Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
WAG	WLAN Access Gateway
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network

List of Symbols

$A^L_{u,l}$	Resource allocation decisions made at lower-level at the beginning of the (u, l)th time slot
A^U_u	Resource allocation decisions made at upper-level at the beginning of the uth time slot
B	Bandwidth of a network
B^W	Bandwidth of WLAN
C	Number of clusters in the network
D	Maximum allowed packet size in contention-based channel access
d_l	The distance within which D2D users are allocated orthogonal cellular network resources
d_p	Processing time at the cloud
d_q	Queueing delay at the cloud
d_t	Transmission delay when access cloud
d_{Total}	Total delay when access cloud
$\mathbb{D}^L$	Lower-level policy
$\mathbb{D}^U$	Upper-level policy
$\mathbf{E}_1(\cdot)$	Exponential integral
$\boldsymbol{F}^s$	Linear polyhedron at sth iteration
h	Channel gain
$\tilde{h}_t$	Normalized complex channel gain at tth time slot
H_{ukt}	$\lvert\tilde{h}_t\rvert^2$ of uth user over kth subcarrier
H^u_{vkt}	Normalized power gain of the channel over kth subcarrier between vth user and the BS to which uth user is connected to, during tth fast time-scale time slot

I_c	Maximum CCI received by the eNB
I_{uk}	Normalized average interference to uth user over kth subcarrier
I_{ukt}	Normalized interference to uth user over kth subcarrier during tth fast time-scale time slot
$\mathcal{K}$	Set of subcarriers available for the entire network
$\mathcal{K}^C$	Set of OFDM subcarriers available in cellular network
$\mathcal{K}^{CF}$	Set of contention-free TXOPs available in WLAN
K_S	Number of possible states for each channel
L	Number of fast time-scale time slots within one slow time-scale time slot
$\mathbf{L}$	A vector consist of L_i's
L_0	Number of fast time-scale time slots in T_0
L_i	Duration of a data packet transmitted by the ith user
$L^U(\cdot)$	Lagrangian for the problem $\mathcal{P}2$
$L^{U(2)}(\cdot)$	Lagrangian for the problem $\mathcal{P}3$
$\mathcal{M}^{(c)}$	Set of macrocells in cth cluster
n	Total noise plus interference power
N	Number of users in the interworking system
N_0	Single sided power spectral density of additive white Gaussian noise
N_W	Number of users allocated for the contention-based channel access of WLAN
$\mathcal{N}^{(c)}$	Set of small cells in cth cluster
p_{ukt}	Transmit power of uth user during tth fat time-scale time slot over kth subcarrier
$\mathbf{P}$	A vector consists of P_{uk}'s
$P_{T,i}$	Total average power available for the ith user
P_{uk}	Average transmit power of uth user over kth subcarrier during next slow time-scale time slot
P_i^C	Total transmit power used by ith user for communications over cellular network
$P_{i,k}^C$	Transmit power level of ith user over kth subcarrier
$P_{avg,i}^C$	Average power usage of ith user for cellular network during one time slot in slow time-scale
$P_{tot,i,l}^C(\cdot)$	Total power allocated by the ith user for cellular network during the

	(u,l)th time slot
$\mathbf{P}^{CB}$	A vector which consists of transmit power levels of users in $\mathcal{S}^{CB}$ during contention-based channel access
$\mathbf{P}_{-1}^{CB}$	A vector which consists of transmit power levels of users in $\mathcal{S}^{CB}$ except the ith user, during contention-based channel access
P_i^{CB}	Transmit power level of ith user during contention-based channel access
$P_{avg,i}^{CB}(\cdot)$	Average power usage of ith user for contention-based channel access during one time slot in slow time-scale
$P_{i,j}^{CF}$	Transmit power level of ith user over jth contention-free TXOP
$P^U_{\psi_0^U \psi_1^U}$	Probability of upper level state change from ψ_0^U to ψ_1^U
$P^L_{\psi_{0,0}^L \psi_{1,0}^L}$	Probability of lower level state change from $\psi_{0,0}^L$ to $\psi_{1,0}^L$
$P^{L(2)}_{\psi_{u,0}^L \psi_{u,1}^L}$	Probability of lower level state change from $\psi_{u,0}^L$ to $\psi_{u,1}^L$
P_i^W	Total transmit power used by ith user for communications over WLAN
r_{ukt}	Throughput achieved by uth user during tth fat time-scale time slot over kth subcarrier
$r_{i,u,l}^L(\cdot)$	Throughput achieved by the ith user at the lower-level during (u,l)th time slot in fast time-scale
$r_{i,u}^U(\cdot)$	Throughput achieved by the ith user at the upper-level during uth time slot in slow time-scale
$R_{Dmin,i}$	Minimum data rate required for data traffic services of ith user
$R_{Vmin,i}$	Minimum data rate required for voice traffic services of ith user
R_{min}	Minimum required data rate
R_{uk}	Average throughput of uth user over kth subcarrier during next slow time-scale time slot
R_i^C	Throughput achieved by ith user via cellular network
$R_{i,k}^C(\cdot)$	Throughput achieved by ith user over kth subcarrier
$R_i^{CB}(\cdot)$	Average throughput achieved by ith user via contention-based channel access
$R_i^{C(D)}$	Throughput achieved by ith user via cellular network using D2D mode
$R_{i,j}^{CF}(\cdot)$	Throughput achieved by ith user over jth contention-free TXOP
$R_i^{C(T)}$	Throughput achieved by ith user via cellular network using traditional mode

$R^{D2D}_{min,i}$ Minimum data rate required for ith user's D2D communications

$R^{L}_{i,u}(\cdot)$ SDT achieved by the ith user at the lower-level over the uth time slot in slow time-scale

$R^{ND}_{min,i}$ Minimum data rate required for ith user's non-D2D communications

$R^{U}_{i}(\cdot)$ SDT achieved by the ith user at the upper-level

R^{W}_{i} Throughput achieved by ith user via WLAN

$R^{W(D)}_{i}$ Throughput achieved by ith user via WLAN using D2D mode

$R^{W(T)}_{i}$ Throughput achieved by ith user via WLAN using traditional mode

$\mathcal{S}_1$ Set of users who are allocated resource during the first step

$\mathcal{S}_2$ Set of users who are allocated resource during the second step

$\mathcal{S}_M$ Set of low-mobility users within WLAN coverage

$\mathcal{S}_N$ Set of all users

$\mathcal{S}_S$ Set of all users except the users in $\mathcal{S}_M$

$\mathcal{S}^{CB}$ Set of users communicate through contention-based channel access

$\mathcal{S}^{CF}$ Set of users communicate through contention-free channel access

SI_D Satisfaction index for data traffic

SI_V Satisfaction index for voice traffic

$SI_{x,y}$ Satisfaction index achieved by y algorithm for x traffic class

T_0 Duration from a beginning of a time slot to the point at which the BSs send CSI to the cloud

T_{ACK} Duration of a acknowledgment in contention-based channel access

T_{AIFS} Duration of arbitration interframe space in contention-based channel access

T_{coh} Coherence time of a wireless channel

T_{CF} Durations of a contention-free TXOP

T_{CFP} Durations of CFP

T_{CP} Durations of CP

T_{CTS} Duration of CTS message in contention-based channel access

T_F Duration of a fast time-scale time slot

T_P CFP repetition period

T_{RTS} Duration of RTS message in contention-based channel access

T_S Duration of a slow time-scale time slot

T_{SIFS} Duration of short interframe space in contention-based channel access

T^L	Duration of a time slot in fast time-scale
T^U	Duration of a time slot in slow time-scale
$\mathcal{U}^{(c)}$	Set of users in cth cluster
$\mathcal{U}_b^{(c)}$	Set of users connected to bth cell in cth cluster
V_L	Number of fast time-scale time slots within a slow time-scale time slot
$\boldsymbol{V}^s$	Set of vertices of the linear polyhedron $\boldsymbol{F}^s$
x_{uk}	Auxiliary variables
$\mathbf{x}$	A vector consists of x_{uk}'s
Y	Percentage of users who can communicate using D2D mode
$\alpha_{i,k}^C$	SINR of the channel between ith user and cellular BS over kth subcarrier with unit transmit power
α_i^W	SINR of the channel between ith user and WLAN AP with unit transmit power
β	Discount factor for the lower-level
$\boldsymbol{\gamma}$	Vector of dual variables corresponding to contention-free TXOP allocation constraints
γ_j	Dual variable corresponds to jth contention-free TXOP allocation constraint
Δf	Bandwidth of a OFDM subcarrier
θ	Discount factor for the upper-level
$\boldsymbol{\lambda}$	Vector of dual variables corresponding to data traffic constraints
λ_i	Dual variable corresponds to data traffic constraint of ith user
$\boldsymbol{\mu}$	Vector of dual variables corresponding to total power constraints
μ_i	Dual variable corresponds to total power constraint of ith user
μ_{uk}	Water level for uth user over kth subcarrier
$\boldsymbol{\xi}$	Vector of dual variables corresponding to voice traffic constraints
ξ_i	Dual variable corresponds to voice traffic constraint of ith user
ρ	Correlation coefficient
$\rho_{i,k}^C$	$\rho_{i,k}^C = 1$ if kth subcarrier is allocated for ith user or $\rho_{i,k}^C = 0$ otherwise
$\rho_{i,j}^{CF}$	$\rho_{i,j}^{CF} = 1$ if jth contention-free TXOP is allocated for ith ($i \in \mathcal{S}_M$) user or $\rho_{i,j}^{CF} = 0$ otherwise
σ_0	Duration an empty slot in contention-based channel access
σ^2	Average power gain of a channel divided by the noise power

σ_{uk}^2	Average normalized power gain of the channel over kth subcarrier between uth user and the BS to which uth user is connected to
$(\sigma_{vk}^u)^2$	Average normalized power gain of the channel over kth subcarrier between vth user and the BS to which uth user is connected to
τ	Probability of a user transmits a packet in a randomly chosen time slot during contention-based channel access
$\psi_{u,l}$	System state, i.e., $\{\psi_u^U, \psi_{u,l}^L\}$
$\psi_{u,l}^L$	State of the lower-level during (u, l)th time slot in fast time-scale
ψ_u^U	State of the upper-level during uth time slot in slow time-scale
Ψ^L	Set of all the possible states of lower level
Ψ^U	Set of all the possible states of upper level
$\tilde{w}$, $\tilde{w}_t$	Complex Gaussian random variables
Ω	Average of square channel gain

Chapter 1

Introduction

Recent advancements in mobile industry have dramatically increased the number of smart mobile devices, such as smart phones, tablets and PDAs, operating in any geographical region. Furthermore, the number of data hungry applications that run on these devices, such as video streaming, YouTube, Google Maps and Facebook, has also been increased. Consequently, the demand for higher data rates with seamless service coverage and support for various applications' diverse quality-of-service (QoS) requirements has been increased than ever before. By year 2020, the volume of mobile data traffic that wireless networks should support is expected to be as high as 1000 times of its value in year 2013 [1, 2]. Therefore, to satisfy the future demands of mobile users, it is essential to develop techniques to enhance the network capacity, QoS support and coverage.

The key techniques that can enhance the wireless network capacity, QoS and coverage performance are adding more spectrum to the networks, increasing the spectrum efficiency and adding more small cells to the networks. Though adding more spectrum to the networks is the simplest and most straight forward method to increase the network capacity, cellular communication-friendly low-frequency spectrum is a scarce resource. Most of the underutilized spectrum lies in very high frequency ranges, and high frequency communication signals travel only a limited distance due to their propagation characteristics. For example, millimeter waves which occupy frequencies between 30GHz and 300GHz are highly susceptible to penetration loss due to their short wave lengths [3, 4]. As a result, they are highly attenuated even by very thin obstacles, such as rain and very thin walls, leaving them only suitable for indoor communications and for small cells. Therefore, it is crucial to efficiently utilize the spectrum available at the lower frequency ranges, such as 800MHz, 1.7GHz and 2.6GHz communications bands, to facilitate long range and robust outdoor communications.

Spectrum efficiency can be enhanced by several techniques, such as interworking of heterogeneous networks, device-to-device (D2D) communication, frequency reuse, and multiple-input and multiple-output (MIMO) communication techniques. Interworking enhances the spectrum efficiency by satisfying the user QoS requirements utilizing resources (i.e., frequency and transmit power) available at multiple networks in an efficient manner. For that, it jointly allocates resources of multiple networks and enables user multi-homing, where multi-homing allows users to simultaneously communicate over multiple networks. D2D communication allows direct communication among the users in proximity [5]. It achieves high throughputs due to short communication distances between the transmitters and the receivers, and saves network resources as it uses only a single hop communication. In traditional communication, two hops are used; first hop is between the transmitter and the base station (BS), and second hop is between the BS and the receiver. Frequency reuse enhances the spectrum efficiency as it reuses the same frequency multiple times at different cells, increasing the number of transmitted bits per frequency [6, 7]. MIMO techniques, such as beamforming and coordinated multi-point (CoMP), enhance the spectrum efficiency as well as the robustness of communications by exploiting spatial diversity (or antenna diversity) [8].

Adding small cells to a wireless network increases the network performance by bringing the network closer to the users [9]. In small cells, the distance between users and the BS is short. Thus, the wireless channels are strong, and the users are able to receive high throughputs. In addition to that, small cells also increase the frequency reuse in the network, due to deployment of a large number of low-powered small-radius cells.

Furthermore, most of the high volumes of mobile traffic demanding areas are covered by heterogeneous wireless networks. For example, office buildings, airports and hotspots are covered by cellular networks and wireless local area networks (WLANs). Therefore, to enhance the spectrum efficiency to increase network capacity and QoS performance, interworking can be used in these areas. Furthermore, use of interworking in these areas widens the network coverage by merging coverage of individual networks. However, interworking cannot improve the network performance when there is coverage of only one network, e.g., at places such as rural areas and cell edges. This shortcoming would significantly affect the cell edge users, as the available coverage is also very weak. To overcome this shortcoming and achieve network-wide uniform performance improvements, interworking can be integrated with D2D communication, frequency reuse, and small cell deployment techniques.

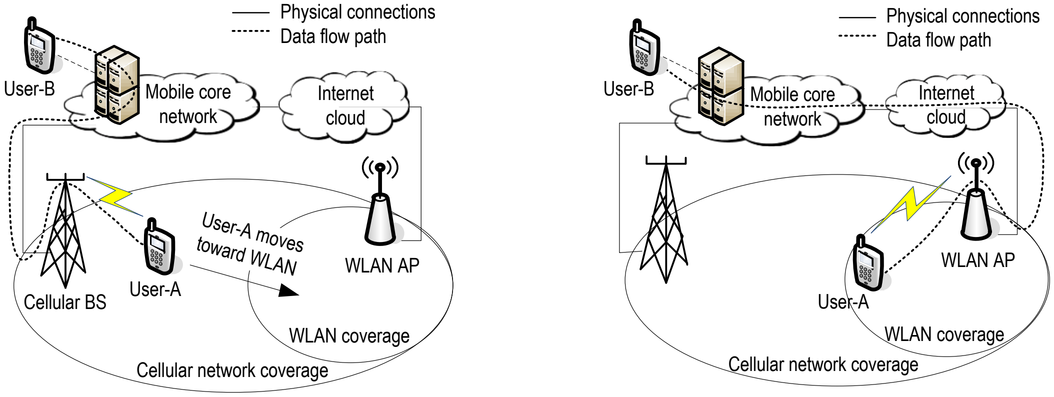
Rest of the chapter is organized as follows. Advantages of and challenges for interworking of heterogeneous wireless networks are discussed in Section 1.1. Section 1.2 discusses integration of D2D communication with interworking. Section 1.3 discusses interworking of macrocell and hyper-dense small cell networks (i.e., highly dense small cell deployments). Motivations and thesis objectives are presented in Section 1.4, while thesis outline is presented in Section 1.5.

1.1 Interworking of Heterogeneous Wireless Networks

Most of the high data capacity demanding areas, such as office buildings, hotspots and airports, are covered by multiple wireless networks with diverse radio access technologies, such as cellular networks and WLANs. These different networks can be interconnected via interworking mechanisms in order to provide users with better network coverage, higher throughputs and better QoS support [10].

Interworking provides users with a seamless network coverage by merging the coverages of individual networks. This is achieved by allowing the users to access the services they need utilizing resources available in multiple networks. A typical coverage extension achieved by cellular/WLAN interworking is demonstrated in Fig. 1.1, where user-A initiates a call using the cellular network, and the call is continued via the WLAN without any interruption as user-A moves away from the cellular BS.

User throughputs and QoS are enhanced as interworking jointly allocates resources of multiple networks, enables user multi-homing, and overcomes the shortcomings of individual networks. Interworking provides a way to pool the resources available in multiple networks. Jointly allocating these pooled resources among all the users in the interworking system increases the efficiency of resource utilization, compared to individually allocating resources available in each network. Multi-homing capability of users' equipment (UEs) with multiple radio interfaces allows the users to simultaneously communicate over multiple networks, as shown in Fig. 1.2. Thus, UEs can access the services simultaneously utilizing resources available in multiple networks in an optimal manner. Therefore, multi-homing further improves the efficiency of resource utilization in the interworking



(a) User-A moves while calling user-B

(b) When the cellular signal strength becomes weak, user-A continues the call via WLAN

Figure 1.1: A coverage extension scenario of cellular/WLAN interworking.

system [11]. Moreover, interworking can overcome the shortcomings of individual networks, e.g., cellular networks provide services at higher data charges with support for mobility while WLANs provide services at much lower charges with limited support for mobility [12, 13, 14, 15]. Also, interworking can solve one of the WLANs' inherent problems of significant throughput degradation when a user with a weaker channel or a low transmit power level joining the WLAN, by reallocating that user to another network (e.g., cellular network) in the interworking system [16].

One of the key challenges for interworking is high complexity of the resource allocation schemes for interworking systems, due to existence of multiple physical layer (PHY) and medium access control layer (MAC) technologies within an interworking system. For example, when interworking of Long Term Evolution (LTE) or LTE-Advanced (LTE-A) cellular networks and IEEE 802.11n WLANs is considered, the cellular network has a centrally coordinated MAC and a orthogonal frequency division multiple access (OFDMA) based PHY, while the WLAN has a hybrid coordination function (HCF) based MAC and a orthogonal frequency division multiplexing (OFDM) based PHY [13, 14]. The region of feasible transmission rates, which includes all the possible sets of users' transmission rates that are supportable by the networks considering all the possible resource allocations, depends on the PHY and MAC technologies of different networks in the interworking

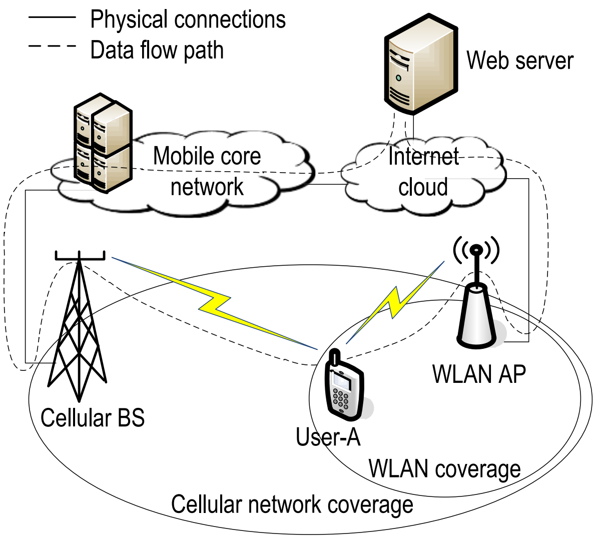


Figure 1.2: User-A uses multi-homing capability to simultaneously communicate via cellular network and WLAN.

system [17]. Therefore, if resources are allocated to the users without considering the underlying PHY and MAC technologies, the resource allocation decisions may be infeasible to carry out; hence, efficiency of the interworking system decreases. Thus, resource allocation schemes should be designed capturing diverse PHY and MAC technologies used in different types of networks in the interworking system. Furthermore, resource allocation intervals (i.e., interval between two successive resource allocations) of existing cellular networks are usually shorter than those of existing WLANs [13, 14]. Therefore, the resource allocation schemes for cellular/WLAN interworking should be designed to periodically allocate cellular network and WLAN resources with a shorter and a longer period respectively, where the periods correspond to the resource allocation intervals of the networks. That is, resources of the two networks are allocated at a faster and a slower time-scale, respectively [18]. In addition to that, most of the wireless networks are able to support multiple classes of QoS. For example, cellular networks and IEEE 802.11n WLANs support both constant bit rate voice traffic and variable bit rate data traffic. Therefore, resources of the interworking system should be jointly allocated exploiting multi-homing capability of the UEs to enhance performance of the interworking system.

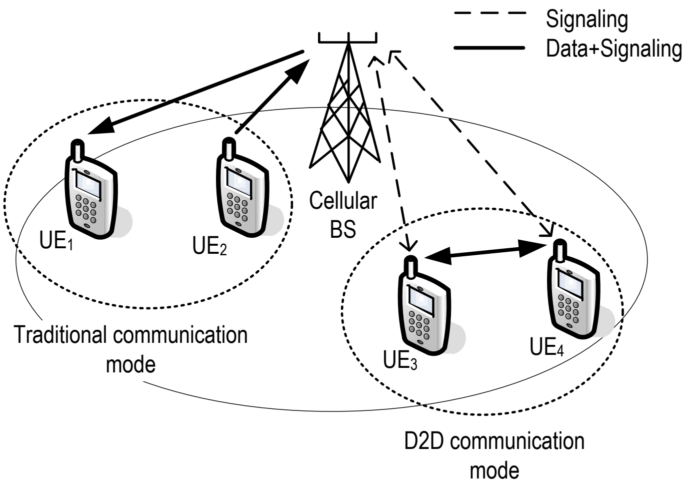


Figure 1.3: Traditional and D2D communications in a cellular network.

Furthermore, to achieve optimal uplink user throughputs, the uplink resource allocation schemes should optimally distribute the transmit power available at multi-homing capable UEs among the multiple network interfaces of these UEs [19].

1.2 D2D Communication Underlaying Interworking Systems

Device-to-device communication allows users in proximity to directly communicate among themselves using direct communication links, without having to send data through a BS. The users who communicate using D2D communication links are referred to as using D2D communication mode, while the users who do not use D2D communication mode (i.e., communications from source to destination are routed through a BS) are referred to as using traditional communication mode. Fig. 1.3 shows a cellular network, where UE_1 and UE_2 communicate using traditional communication mode while UE_3 and UE_4 communicate using D2D communication mode.

Enabling D2D communication in an interworking system provides several important benefits. First, an interworking system may not provide the enhanced network performance at areas such as cell edges or rural areas where only one network is available. D2D communication can be applied in these areas to improve the network performance, as it allows direct communication between source and destination UEs which are in proximity, and incorporates hop and reuse gains to the network [20, 21, 22, 23, 24]. Hop gain is a result of D2D communication links using either uplink (UL) or downlink (DL) resources only. Reuse gain is achieved by simultaneously using the same set of resources for both traditional and D2D communication links [20, 22, 25]. Second, D2D communication reduces the cost of service as it allows network operators to offload traffic from the mobile core network. Third, enabling D2D communication does not require additional hardware deployments.

Similar to interworking, D2D communication also cannot improve the network performance throughout the network, as the probability of communications among users in proximity is small [26]. Therefore, integration of D2D communication and interworking would complement the areas where each of these technologies can improve the network performance; hence, the integrated system would provide uniform performance improvements throughout the networks.

D2D communication between two UEs can be initiated by UEs without any involvement of the mobile operators, or it can be controlled and initiated by the mobile operator. The latter is called network assisted D2D communication mode. It provides security and mobility support for the UEs, and reduces the interference caused by non-synchronized UEs [22].

There are two types of resource allocations for D2D communication links: 1) a fixed set of resources are allocated for the D2D communication links, and from which resources for each D2D communication link is allocated; and 2) resources for D2D and traditional communication links are jointly allocated. Each of these two resource allocation types can be again divided into two categories based on the nature of the resource sets that are allocated for D2D and traditional communication links. When the same resource set is shared by D2D and traditional communication links, such resource sharing is called non-orthogonal resource sharing. When the allocated resource sets for D2D and traditional communication links are non-overlapping, it is called orthogonal resource sharing. Non-orthogonal resource sharing is more efficient than orthogonal resource sharing, as it allows both D2D and traditional communication links to utilize all the available resources. However, use of non-orthogonal resources causes co-channel interference (CCI) among D2D and traditional communication links.

There are several technical challenges which make the resource allocation for D2D communication underlaying interworking systems complicated: 1) involvement of multiple PHY and MAC technologies of different networks [14, 13, 11], 2) selection of users' communication modes (i.e., traditional or D2D) for multiple networks to maximize hop and reuse gains while considering the resources available in individual networks, and 3) interference management [22].

1.3 Interworking of Macrocell and Hyper-Dense Small Cell Networks

Small cells can be densely deployed to cater for high service demands in areas, such as busy streets, city centers and shopping malls. They can be deployed by the network operators as well as the subscribers, in both planned and unplanned manner [2]. Backhauls to the small cell BSs could be fiber-optic cables or digital subscriber lines (DSL) [27].

There are four main benefits of hyper-dense small cell deployments. First, hyper-dense small cell deployments increase the network capacity by bringing network closer to the users [28, 9]. Second, the spectral efficiency is significantly enhanced by reusing the same set of spectrum resources in a large number of small cells [29]. Third, less deployment and operational costs due to low-cost of small cell BSs, DSL based backhauls and that most of the small cell BSs are installed and operated by the subscribers [27]. Fourth, higher carrier frequencies can also be used for small cells due to small targeted coverage areas.

However, there may be coverage holes in hyper-dense small cell networks due to unplanned deployment of small cells with small coverage areas. In addition to that, highly mobile users may experience call drops in these networks, as they have to be frequently handed over from one small cell to another. These two shortcomings can be eliminated by enabling interworking between hyper-dense small cell networks and macrocell networks, as the users at coverage holes of hyper-dense small cell networks and the highly mobile users can be allocated to the macrocell network. Therefore, interworking of macrocell and hyper-dense small cell networks is a promising technique that can enhance the network performance to satisfy the ever increasing demand for high volumes of mobile data traffic. Interworking of macrocell and hyper-dense small cell networks is shown in Fig. 1.4.

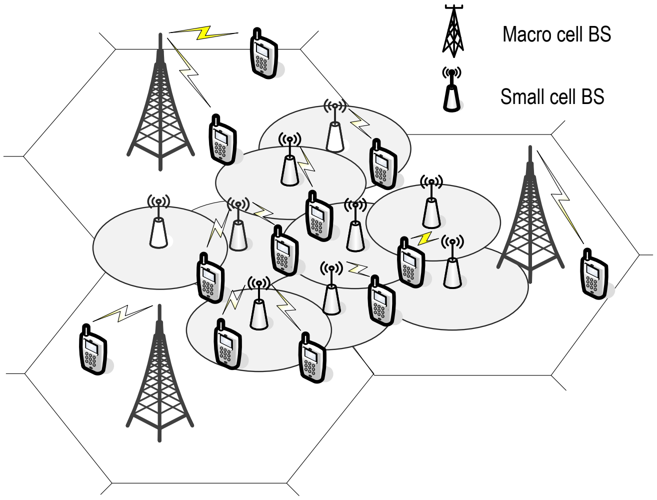


Figure 1.4: Interworking of macrocell and hyper-dense small cell networks.

There are several challenges for allocating resources for interworking macrocell and hyper-dense small cell networks. First, such network is prone to severe CCI, mainly due to short distances between the densely deployed small cells and that many small cell BSs are deployed by the subscribers in an unplanned manner [9]. Therefore, resources of all the small cells in vicinity should be jointly allocated such that CCI remains within a tolerable level. Second, network load significantly varies and moves across the network with time. For example, network load will be high near restaurants during lunch time, while it will be high near coffee shops in the afternoon. Thus, variations of the network load should be considered in order to optimally utilize the available spectrum resources. Third, small cell backhaul capacities may be limited due to the use of DSLs. Thus, it may be a bottleneck if the volume of user data and control signaling is high. Fourth, limited computational capacity available at low-cost small cell BSs. Therefore, small cell BSs cannot execute computationally expensive resource allocation algorithms.

In addition to jointly allocating resources of the small cells, CCI can be managed by allocating almost blank subframes (ABS) [30, 31], fractional frequency reuse (FFR) techniques [32, 33], and MIMO techniques, such as beamforming and joint decoding [34, 35, 36, 28]. By allocating ABSs, CCI among the small cells cannot be managed as ABSs are allocated by the macrocells to mitigate CCI between the macrocell and the small cells. The FFR techniques will significantly reduce the frequency reuse in the network due to highly dense small cell deployments. The MIMO techniques are inefficient for managing CCI in a densely deployed small cell network due to three reasons: 1) a dense unplanned small cell deployment causes each cell to receive CCI not only from the neighbouring cells, but also from neighbours of the neighbouring cells; 2) requirement to exchange a large volume of channel state information (CSI) among a large number of BSs in proximity, occupying a large part of the small cells' limited capacity backhauls, as the MIMO techniques coordinate BSs based on the instantaneous CSI; and 3) high cost of antenna arrays.

1.4 Motivations and Thesis Objectives

Demand for high volumes of mobile data traffic with better QoS support and seamless network coverage is ever increasing. Most of these high volumes of data traffic demanding areas are covered by heterogeneous wireless networks. Therefore, through interworking of these heterogeneous wireless networks, network capacity, QoS support and coverage can be improved. However, interworking cannot improve the network performance when coverage of only one network is available. Thus, to achieve network-wide uniform performance improvements, interworking can be integrated with D2D communication, frequency reuse, and small cell deployment techniques. There are several challenging issues that need to be solved before these techniques can be applied in practical networks. One of such issues is the efficient resource allocation when these techniques are applied. Efficiently allocating resources, such as frequency and transmit power, is crucial as it has a direct impact on the spectrum efficiency. Therefore, in this thesis, we study and develop efficient resource allocation schemes for interworking heterogeneous wireless networks which also apply D2D communication and small cell deployment techniques for different areas of the network.

The objectives of the thesis can be stated as follows:

1. to design an efficient resource allocation scheme for cellular/WLAN interworking system, based on the PHY and MAC technologies and the different resource allocation time-scales of the two networks;

2. to present a resource allocation scheme for D2D communication underlaying cellular/WLAN interworking system, to determine effective user communication modes and allocate resources of the two networks; and
3. to develop a resource allocation scheme for interworking macrocell and hyper-dense small cell networks, based on the characteristics of the hyper-dense small cell networks.

1.5 Thesis Outline

The remainder of the thesis is organized as follows.
Chapter 2 describes implementation of interworking of heterogeneous wireless networks.
Chapter 3 presents the system model.
Chapter 4 proposes an optimal multiple time-scale Markov decision process (MMDP) based and a low-complex heuristic resource allocation schemes for cellular/WLAN interworking system. The resource allocation schemes are designed considering multi-homing capable users with voice and data traffic requirements, two time-scales, and underlying PHY and MAC layer technologies of an OFDMA based cellular network and a WLAN which operates on contention-based and contention-free channel access mechanisms.
Chapter 5 first presents the technical challenges for selecting communication modes and allocating resources in a D2D communication underlaying cellular/WLAN interworking system, and discusses existing and new solutions. Second, a semi-distributed resource allocation scheme is proposed to address these challenges, and the related implementation issues are investigated.
Chapter 6 focuses on resource allocation for interworking macrocell and hyper-dense small cell networks. First, it discusses potentials of using cloud computing to assist resource allocation decision making process, to reduce the computational burden for low-cost small cell BSs. Second, a cloud assisted resource allocation scheme which operates on two time-scales is proposed. Two time-scales are used to overcome the negative effect of high delay, when access cloud computing facilities, on the resource allocation decisions.
Chapter 7 presents the conclusions and the future works.

Chapter 2

Implementation of Interworking

This chapter provides implementation details of interworking of heterogeneous wireless networks. Section 2.1 describes different categories of interworking architectures and the mechanisms that can be used for designing interworking architectures belonging to those categories. Section 2.2 presents the interworking architectures that have been proposed in literature. The chapter is summarized in Section 2.3.

2.1 Classification of Interworking Architectures

Several architectures which enable interworking by interconnecting multiple networks have been proposed in literature. These interworking architectures can be classified into four main categories, based on the supported interworking levels (i.e., service levels). The main interworking levels are as follows [10],

- **Level A**: Mobile users are only allowed to access the services provided by the visited network.
- **Level B**: Mobile users are able to access the services provided by its home network through the visited network. However, users have to re-establish the sessions through the visited network.
- **Level C**: This level provides service continuity when users move between different networks, unlike in level B. Users do not have to re-establish the active sessions.

However, there may be temporary QoS degradation during the transition time (i.e., handover).

- **Level D**: This level provides seamless mobility with no QoS degradation during and after the transition.

The key mechanisms (i.e., concepts and technologies) which can be used for achieving different interworking levels can be summarized as follows [10],

- **Mechanisms that achieve interworking level A**: This mechanism mainly extends the home network's authorization, authentication and accounting (AAA) functionalities to the visited network. In addition to that, enhanced network discovery mechanisms are also employed in order for the mobile users to know the networks which they are allowed to connect (i.e., which network is going to accept its credentials, etc.). E.g., Interworking-WLAN (I-WLAN) architecture [37].
- **Mechanisms that achieve interworking level B**: Data between the mobile user and its home network is transferred via a layer-2 tunneling protocol (L2TP)/internet protocol security (IPsec) tunnel. E.g., I-WLAN architecture.
- **Mechanisms that achieve interworking level C**: Service continuity when handover occurs is achieved by employing network layer handovers with the aid of mobile Internet protocol (MIP) or proxy mobile Internet protocol (PMIP) techniques [38]. These MIP based mobility solutions have been adopted by the LTE/system architecture evolution (SAE) architecture to provide service continuity between LTE/SAE networks and other networks [37]. The main advantage of using PMIP over MIP is that PMIP does not require a UE to update its new Internet protocol (IP) address (i.e., care of address (CoA)) at the home agent (HA) whenever the UE moves to a new network. In addition to these solutions, if the service is available at both visited and home networks, service continuity can also be achieved using application based mobility solutions, such as session initiation protocol (SIP), without using network layer handover mechanisms.
- **Mechanisms that achieve interworking level D**: Seamless handovers between networks with different radio access technologies are achieved by optimizing the network layer and the data link layer handover mechanisms. Three methods can

be used for reducing the latency in the network layer during a handover: 1) acquire CoA before the handover occur, using MIP version 6 (MIPv6); 2) pre-authenticate the UE; and 3) preconfigure the network layer parameters in the targeted network prior to handover occur, using context transfer protocols. In the data link layer, latency can be reduced by providing UEs with information about the available networks, such as, configurations, signal level measurements and resource availability, prior to handover occur. Furthermore, when UEs are capable of multi-homing, UEs can simultaneously connect to two networks; thus, optimization procedures for these two layers are not required. Generic access network (GAN) architecture uses the UE multi-homing capability to support interworking level D [39].

2.2 Interworking Architectures

Several interworking architectures have been proposed in literature. This section describes implementations of I-WLAN, GAN, interworking of worldwide interoperability for microwave access (WiMAX) and 3rd generation partnership project (3GPP) networks, and IEEE 802.21 media independent handover (MIH) architectures.

2.2.1 Interworking-WLAN (I-WLAN)

I-WLAN [37] is commonly referred to as loose coupling interworking, and it can provide levels A and B interworking of WLANs and cellular networks, such as, 3GPP universal mobile telecommunications system (UMTS) networks, for packet services. In this scheme, cellular network extends its AAA functions to the WLAN by exchanging extensible authentication protocol (EAP) messages between these two networks. The users can then be authenticated and authorized at the WLAN based on the credentials provided by the cellular network. In the interworking level B, data is transferred between UEs and the packet data gateway (PDG) in home cellular network through secure IPsec tunnels.

2.2.2 Generic Access Networks (GAN)

GAN [39] is referred to as tight coupling interworking, and it supports interworking levels B, C and D for circuit and packet services between cellular networks and other broadband access networks, such as, WLANs. In this architecture, GAN controller (GANC) is introduced to the cellular network to connect the cellular network with the other broadband networks. Access control is done at the cellular network unlike in the I-WLAN architecture. Data is transferred between the UEs and GANC via IPsec tunnels. Interworking level-D is achieved by using multi-homing capability of the dual-radio UEs. Multi-homing capability is used for simultaneously connecting UEs to cellular network and GANC via WLAN. Therefore, in this architecture, network layer or data link layer handover optimization is not required.

2.2.3 WiMAX and 3GPP Networks Interworking

The first release of WiMAX architecture has incorporated solutions based on the I-WLAN architecture to facilitate interworking between 3GPP cellular networks and WiMAX networks to support interworking levels A and B. Interworking level C can be achieved between WiMAX and LTE/SAE networks by using MIP and PMIP technologies while interworking level D can be achieved by using multi-homing capability of the UEs. However, achieving interworking level D for single-radio UEs requires sending the handover related information (for handover optimization, resource reservation, access control, etc.) to the target network. These information can be sent to the target network using the tunnels created through the network which currently serves the UE [40]. In the interworking system consisting of WiMAX and 3GPP networks, 3GPP network extends its AAA functionalities to WiMAX access service network (ASN) gateway to authenticate and authorize 3GPP network users. In this architecture, the MIP client is located at the UE when MIP is used, while it is located at the WiMAX ASN gateway when PMIP is used. HA is located at the packet data network (PDN) gateway in the LTE/SAE network [10].

2.2.4 IEEE 802.21 Media Independent Handover (MIH)

MIH standard [41] has been proposed to provide interworking level D functionalities among IEEE 802 wireless/wired networks (e.g., IEEE 802.11 and 802.16) and cellular networks. MIH standard defines new functionalities to be added to the data link layer of the UEs and the networks. These new functions will check the resource availability in the candidate networks and reserves resources in preparation for the handover. Further, UEs will be updated with the information about the configurations of the reserved resources. However, MIH focuses only on initiation and preparation phases of a handover. It does not provide optimized mechanisms to be followed during the handover process.

2.3 Summary

This chapter describes the four main interworking levels, which are defined based on the QoS during handovers and the services that users are able to access through a visited network. The mechanisms that can be used to achieve those interworking levels are also presented. Finally, implementations of the interworking architectures that have been proposed in literature are described.

Chapter 3

System Model

The system model is described in this chapter. Section 3.1 presents an overview of the system model, while Sections 3.2 and 3.3 provides details of the cellular networks and the WLANs. A summary of the chapter is given in Section 3.4.

3.1 Network Overview

A wireless network that uses interworking, D2D communication and small cell deployment techniques to enhance network throughput, QoS and coverage performance is considered. An example of such network is shown in Fig. 3.1. There are three types of areas of interest in this network: 1) areas which are only covered by the macrocell network, 2) areas which are covered by both macrocell network and a WLAN, and 3) areas which are covered by both macrocell and hyper-dense small cell networks. In the example shown in Fig. 3.1, $\mathrm{UE}_3-\mathrm{UE}_5$, UE_9 and UE_{10} are in the first area, and these UEs communicate only through the macrocell network. The macrocell network and the WLANs are interconnected using interworking mechanisms. Thus, the UEs in the second area are able communicate through macrocell network or WLAN, or through both networks using the UE multi-homing capability. For example, UE_1 communicates only using the WLAN while UE_2 communicates simultaneously using both WLAN and the macrocell network. Using D2D communication, users in proximity can directly communicate among themselves, e.g., UE_4 and UE_5. Moreover, since UE_{11} and UE_{12} are within coverage of both macrocell network and a WLAN, they can form the D2D link using either network or

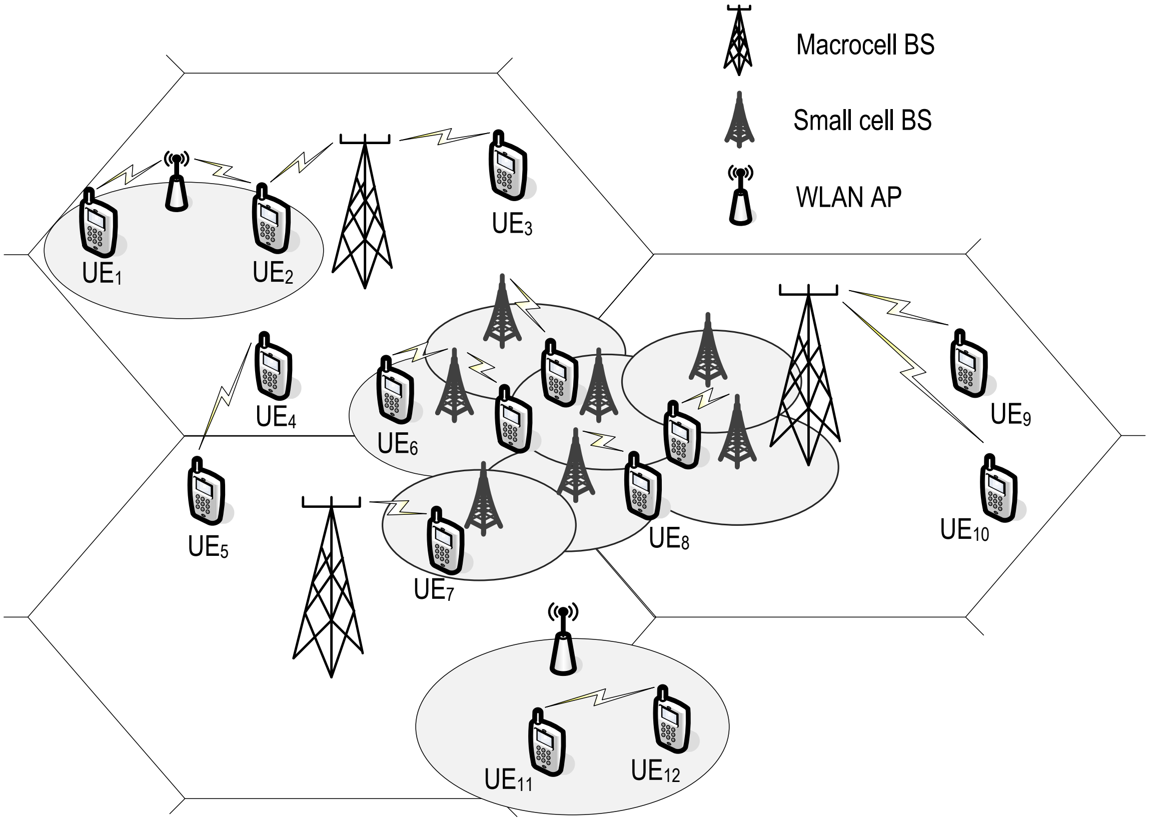


Figure 3.1: A network that uses different spectrum efficiency enhancing techniques.

both networks. Macrocell and hyper-dense small cell networks are interconnected using interworking mechanisms. Therefore, in the third area, UEs can communicate using either network or both networks. For example, UE_6 and UE_7 communicate using the hyper-dense small cell network and the macrocell network, respectively.

Macrocell and hyper-dense small cell networks are OFDMA based networks, similar to LTE and LTE-A networks [14]. WLANs consist of contention-based and contention-free polling based channel access mechanisms, as in IEEE 802.11n WLANs [42].

3.2 Cellular Networks

Macrocell and hyper-dense small cell networks consist of OFDMA based PHY layers. MAC layer of these networks centrally coordinates the allocation of the network resources. Resources to be allocated are OFDMA subcarriers and transmit power of the BSs and UEs. In the uplink resource allocation, only UEs' transmit power is considered. User allocation is also considered as a part of the resource allocation when multiple cells are considered.

3.2.1 User Throughputs via Cellular Networks

Available frequency band is divided into subcarriers with bandwidth of Δf. The maximum achievable error free data rate by the ith user over the kth OFDM subcarrier of the cellular network can be expressed by

$$R^C_{i,k}(P^C_{i,k}) = \Delta f \log_2(1 + \alpha^C_{i,k} P^C_{i,k}), \tag{3.1}$$

where $\alpha^C_{i,k}$ is the signal-to-interference plus noise ratio (SINR) of the channel between cellular BS and the ith user over the kth subcarrier with unit transmitted power; and $P^C_{i,k}$ is the transmit power level of the ith user over the kth subcarrier.

3.3 WLAN

The WLANs consist of an OFDM based PHY and a MAC that grants transmission opportunities (TXOP) for the users using two channel access mechanisms: contention-based channel access during contention period (CP) and contention-free polling based channel access during contention-free period (CFP). CP and CFP alternate over time and they repeat once every T_P. Durations of CP and CFP are denoted by T_{CP} and T_{CFP}, respectively. Operation of these two channel access mechanisms is shown in Fig. 3.2. In the contention-based channel access, users contend for the channel to obtain TXOPs, and each of these TXOPs allows the user who obtained the TXOP to send a data packet of D bits as channel capture is not allowed. Furthermore, four-way handshaking scheme with request-to-send (RTS) and clear-to-send (CTS) messages is used for improving the efficiency of the WLAN by reducing the duration of collisions, and for solving the hidden terminal problem. In the contention-free polling based channel access, users are granted TXOPs using a centralized polling mechanism, and each of these TXOPs is defined as a fixed duration (T_{CF}) which allows users to transmit. Contention-based channel access is more suitable for variable bit rate data traffic while contention-free channel access is more suitable for constant bit rate voice traffic [43]. Thus, different sets of users can be assigned to these two channel access mechanisms, based on the user requirements.

User throughputs achieved by contention-free channel access mechanism can be determined using the Shannon capacity formula, similar to throughput calculation for the cellular network users in Section 3.2.1, due to coordination of TXOP allocation by a centralized polling mechanism. However, in the contention-based channel access, users obtain TXOPs in a random manner. Moreover, data transmissions during an obtained TXOP may not be successful due to collisions of different users' data packets. Thus, in the resource allocation problem formulation, average user throughputs and average power consumptions of the users who use contention-based channel access are used.

3.3.1 User Throughputs via Contention-Free Channel Access

The maximum achievable error free data rate by the ith user using the jth contention-free TXOP can be expressed by

$$R_{i,j}^{CF}(P_{i,j}^{CF}) = B^W \log_2(1 + \alpha_i^W P_{i,j}^{CF}), \tag{3.2}$$

where α_i^W is the SINR of the channel between the WLAN access point (AP) and the ith user with unit transmitted power; $P_{i,j}^{CF}$ is the transmit power level of the ith user over the jth contention-free TXOP; and B^W represents the bandwidth of a WLAN.

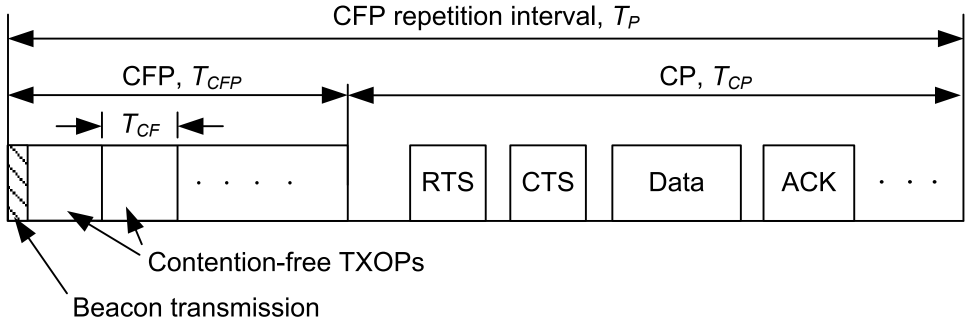


Figure 3.2: TXOP allocation in the WLAN.

As wireless channels are assumed to be fixed within their coherence time, resource allocation interval for the WLANs is selected such that it is shorter than the channel coherence time. Therefore, the SINRs (i.e., α_i^W) remain unchanged within a resource allocation interval. Furthermore, T_P is selected such that it is shorter than or equivalent to the resource allocation interval. Thus, SINR levels are the same for contention-free and contention-based channel access mechanisms.

3.3.2 User Throughputs via Contention-Based Channel Access

Average throughput achieved by the ith of the users who use contention-based channel access during a CP is given by [13, 44]

$$R_i^{CB}(\mathbf{L}) = \frac{\tau(1-\tau)^{N_W-1}D}{T_1 + N_W\tau(1-\tau)^{N_W-1}\sum_{j=1}^{N_W} L_j}, \tag{3.3}$$

where T_1 can be calculated as:

$$\begin{aligned} T_1 =& N_W\tau(1-\tau)^{N_W-1}(T_{CTS}+T_{ACK}+3T_{SIFS}) + (1-(1-\tau)^{N_W})(T_{RTS}+T_{AIFS}) \\ &+ (1-\tau)^{N_W}\sigma_0. \end{aligned}$$

In (3.3), L_i is the duration of a packet transmitted by the ith of the users who use contention-based channel access, N_W is the number of users who use contention-based channel access, and T_{SIFS}, T_{AIFS}, T_{ACK}, T_{RTS}, T_{CTS} and σ_0 are the durations of short interframe space, arbitration interframe space, acknowledgment, RTS message, CTS message and an empty slot, respectively. τ is the probabilities of a user transmits a packet in a randomly chosen time slot, and it can be found by [44].

$R_i^{CB}(\mathbf{L})$ given by (3.3) cannot be used in a resource allocation problem as it is not in terms of the transmit power levels of the users. Therefore, L_i is rewritten in terms of the throughput that is achieved by the ith user during a successful transmission, and substitute it into (3.3) [45]. Then, $R_i^{CB}(\mathbf{L})$ can be rewritten as

$$R_i^{CB}(\mathbf{P}^{CB}) = \frac{\tau(1-\tau)^{N_W-1}D}{T_1 + \frac{DN_W\tau(1-\tau)^{N_W-1}}{B^W}\sum_{j=1}^{N_w}\frac{1}{\log_2(1+P_j^{CB}\alpha_j^W)}}, \tag{3.4}$$

where P_i^{CB} is the transmit power level of the ith user during the CP over the WLAN interface, and $\mathbf{P}^{CB}$ is a vector consisting of P_i^{CB}'s. From (3.4), it can be seen that $R_i^{CB}(\mathbf{P}^{CB})$ is the same for all the users who use contention-based channel access. Furthermore, $R_i^{CB}(\mathbf{P}^{CB})$ is a concave function when N_W is fixed, and the proof of convexity of $R_i^{CB}(\mathbf{P}^{CB})$ is given in Appendix A.1.

3.3.3 Average User Transmit Power Through Contention-Based Channel Access

Average power usage of a contention-based channel access user through the WLAN interface during a CP can be determined as follows. Since $\tau(1-\tau)^{N_W-1}$ is the probability of a successful transmission during a CP, the average power usage of the ith of the contention-based channel access users is

$$P_{avg,i}^{CB}(\mathbf{P}^{CB}) = \tau(1-\tau)^{N_W-1}\frac{L_i P_i^{CB}}{T_{avg}}\frac{T_{CP}}{T_P}, \tag{3.5}$$

where T_{avg} is the average duration of channel occupancy due to an event of successful transmission, collision, or empty slot in which no user transmits [44]. After simplifying

(3.5), $P_{avg,i}^{CB}(\mathbf{P}^{CB})$ can be expressed as

$$P_{avg,i}^{CB}(\mathbf{P}^{CB}) = \begin{cases} \frac{T_{CP} P_i^{CB} R_i^{CB}(\mathbf{P}^{CB})}{T_P B^W \log_2(1+\alpha_i^W P_i^{CB})}, & \text{if } P_i^{CB} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{3.6}$$

3.4 Summary

In this chapter, first an overview of the system model was presented. This overview discusses how different spectrum efficiency enhancing techniques, such as, interworking, D2D communication and deployment of hyper-dense small cell networks, are used. Second, details of the cellular networks, including user throughput calculation, are presented. Third, details of the WLANs and the two channel access mechanisms of the WLANs are described. Calculations of the WLAN user throughputs and power consumptions are also given.

Chapter 4

Resource Allocation for Cellular/WLAN Interworking

In this chapter, we investigate uplink resource allocation for cellular/WLAN interworking system to maximize the system throughput and achieve QoS satisfaction. The cellular network is OFDMA based, similar to LTE/LTE-A networks. The WLAN operates on both contention-based and contention-free polling based channel access mechanisms, similar to IEEE 802.11n WLANs with HCF based MAC [13]. Resources of the system are subcarriers of the cellular network, TXOPs via the two channel access mechanisms of the WLAN, and UE transmit power.

As stated in Section 1.1, one of the key challenges for allocating resources for this system is the high complexity of resource allocation algorithms due to existence of multiple PHY and MAC technologies, and the resources of the cellular network and the WLAN need to be allocated at a fast and a slow time-scale, respectively [18]. The set of possible resource allocation decisions and the achievable user throughputs over a network depends on the PHY and MAC technologies of the network [11, 17, 45]. For example, in an OFDMA based network, bandwidth allocation decisions should be in multiples of a subcarrier bandwidth, while during the contention-based channel access of the WLAN, user throughputs should be calculated considering the transmission collisions which occur due to the MAC scheme. Therefore, efficient resource allocation schemes should be designed based on the underlying PHY and MAC technologies in order to make feasible resource allocation decisions while properly estimating the throughputs that users will achieve. The challenge in an interworking system is that there are diverse PHY and MAC technologies of different types of networks to be considered.

Resource allocation interval of the cellular networks is much shorter than that of the WLANs as cellular networks and WLANs are designed to support high and low mobility users with speeds upto 350kmh^{-1} and 3kmh^{-1}, respectively [13, 14]. The resource allocation intervals are calculated based on the channel coherence times. Therefore, resources of these two networks need to be allocated in a fast and a slow time-scale, respectively [11]. Existence of multiple time-scales in an interworking system poses a challenge for resource allocation, as prediction of the throughputs that will be achieved over future time slots is required to optimally allocate the resources. That is, when resources of the interworking system are allocated at a beginning of a slow time-scale time slot, it is required to predict the throughputs that will be achieved over the future fast time-scale time slots which lie within the current slow time-scale time slot.

In this chapter, to allocate resources for the cellular/WLAN interworking system, first we propose a resource allocation framework for cellular/WLAN interworking operating on two time-scales. Second, we formulate the resource allocation problem in the proposed framework as a MMDP based on the PHY and MAC technologies of the two networks, and derive a resource allocation algorithm which consists of decision policies for the MMDP. Third, to further reduce the time complexity, we propose a heuristic resource allocation algorithm.

The remainder of this chapter is organized as follows. Section 4.1 summarizes related works and Section 4.2 describes the cellular/WLAN interworking system model. Section 4.3 presents the MMDP based resource allocation. Sections 4.4 and 4.5 discuss resource allocations at upper and lower levels of the proposed framework, respectively. The heuristic resource allocation algorithm is presented in Section 4.6, while simulation results and the summary of the chapter are given in Sections 4.7 and 4.8, respectively.

4.1 Related Work

Existing resource allocation schemes can be classified into three categories: schemes using a single network interface of each UE at any given time [46, 47, 48, 49, 50, 51], schemes utilizing the multi-homing capability of UEs [11, 52, 53, 54], and schemes that are designed based on different PHY and MAC technologies [45, 12]. A load balancing scheme to improve resource utilization in cellular/WLAN interworking is presented in [48]. New voice and data calls are assigned to a network based on a set of precalculated probabilities. Assigned calls are re-distributed whenever necessary to another network by using dynamic vertical handoffs to reduce network congestion and improve QoS satisfaction. To further improve QoS satisfaction, the scheme proposed in [46] allocates voice calls preferably for the cellular network. The resource allocation scheme proposed for WiMAX/WLAN interworking in [47] assigns all streaming calls to the WiMAX network to guarantee QoS satisfaction; data calls that are served by the WiMAX network are preempted to free up bandwidth for the incoming streaming calls when required. The main advantage of these schemes in the first category is that they are easy to deploy as each network can use its own/existing resource allocation scheme to allocate resources. Further, designing an efficient resource allocation scheme is simpler for an individual network than for an interworking system.

When UEs are capable of multi-homing, restricting a UE or a certain traffic type of a user to access only one network limits the flexibility in distributing resources of the interworking system among users. Thus, the resource allocation schemes in the second category take advantage of the multi-homing capability of UEs to efficiently utilize resources of the interworking system. For computational simplicity, it is typically assumed that the WLAN uses a resource reservation protocol to avoid channel contention collisions. Hence, resources of the WLAN are modeled as frequency channels or time slots. Bandwidth allocation algorithms for UEs with different types of traffic requirements are presented in literature. In [11], each network gives more priority to satisfy its own subscribers' QoS requirements, while utility fairness among users in the interworking system is maintained in [52]. A game theoretic approach for bandwidth allocation and admission control is used in [53]. Each network allocates its bandwidth for different service areas on a long-term basis based on the statistics of call arrivals; bandwidths for each service area from different networks are then allocated to users on a short-term basis. To ensure QoS satisfaction, a new call is accepted only if its minimum data rate requirement can be satisfied. Algorithms to allocate time slots in a WLAN and subcarriers in a cellular network subject to a proportional rate constraint are presented in [54].

The third category includes the resource allocation schemes proposed in [45, 12]. These schemes are based on PHY and MAC technologies of the different networks to guarantee the feasibility of resource allocation decisions. Specifically, the effect of transmission collisions caused by the contention-based channel access in the WLAN is considered. In

[12], resource allocation and admission control schemes are proposed for an interworking system consisting of a code division multiple access (CDMA) based cellular network and an IEEE 802.11 distributed coordination function (DCF) based WLAN. Maximizing total network welfare ensures QoS satisfaction in the system. In [45], interworking of an OFDMA based femtocell network and an IEEE 802.11 DCF based WLAN is considered. Resources of both femtocell and WLAN are allocated on the same time-scale, and WLAN uses basic access scheme with two-way handshaking.

The existing resource allocation schemes allocate resources of different networks in the interworking system at the same time-scale, and do not fully utilize the QoS support in WLANs. Allocating resources of different networks at the same time-scale is not practical as different networks have different resource allocation intervals. To facilitate QoS in WLANs, recent WLAN standards offer contention-based and contention-free polling based channel access mechanisms. These two channel access mechanisms and their QoS capabilities should be considered to maximize the efficiency of the interworking system. In addition, jointly allocating transmit power levels for different network interfaces at multi-homing capable UEs is essential for an efficient resource utilization. Joint transmit power allocation is studied in [45] without taking the user QoS requirements into account.

In this chapter, we study the resource allocation for cellular/WLAN interworking to satisfy the QoS requirements of multi-homing UEs. Based on the PHY and MAC technologies of these two networks, the resources are allocated to multi-homing UEs at two time-scales: one time-scale for allocating resources of each network. We consider power allocation for multi-homing UEs, and the two channel access mechanisms of the WLANs.

4.2 Cellular/WLAN Interworking System Model

The system under consideration for this chapter focuses on first and second type areas described in Section 3.1. Such system is shown in Fig. 4.1, and it consists of a single cell of a cellular network (specifically, a single cell of a macrocell network) and a WLAN within the coverage of the cell. We focus on the resource allocation for the uplink. In the system, there are N users belonging to two groups: high-mobility users and low-mobility users. The set of all the users is denoted by $\mathcal{S}_N$. The set of low-mobility users within the WLAN coverage is denoted by $\mathcal{S}_M$, while the set of remaining users is denoted by

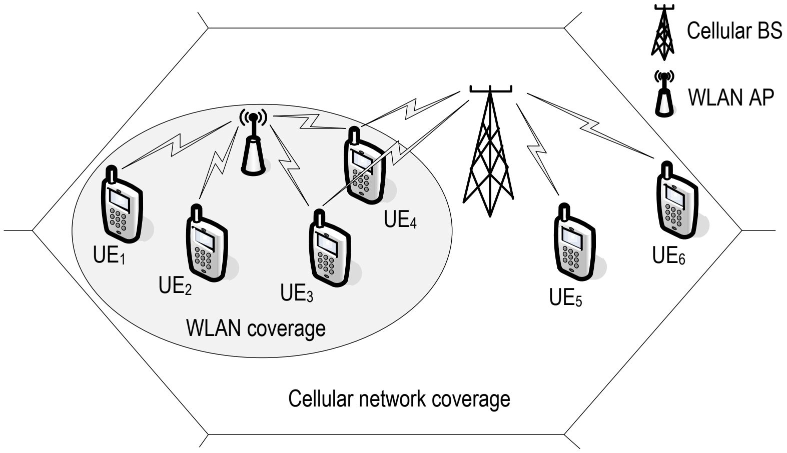


Figure 4.1: Cellular/WLAN interworking.

$\mathcal{S}_S$. For example, in Fig.4.1, UE$_1$ to UE$_4$ are in $\mathcal{S}_M$, while UE$_5$ and UE$_6$ are in $\mathcal{S}_S$. Each user has voice and data traffic requirements. All the UEs are equipped with WLAN and cellular network interfaces, and have the multi-homing capability. Users in $\mathcal{S}_M$ are allowed to simultaneously communicate over cellular network and WLAN, while users in $\mathcal{S}_S$ are only allowed to communicate over the cellular network.

The set of subcarriers available at the cellular network BS is denoted by $\mathcal{K}^C$. At any time, each subcarrier is allocated to only one user in order to avoid CCI among the users. Both voice and data traffic services are served through the cellular network. The set of available contention-free TXOPs during a CFP is denoted by $\mathcal{K}^{CF}$. To avoid CCI among the serss, each contention-free TXOP is allocated for only one user at any given time. Contention-based channel access is more suitable for variable bit rate data traffic, while contention-free channel access is more suitable for constant bit rate voice traffic [43]. To optimize resource utilization subject to QoS requirements, voice traffic is served by contention-free channel access, and data traffic is served by both channel access mechanisms. The sets of users communicate using contention-based and contention-free channel access are denoted by $\mathcal{S}^{CB}$ and $\mathcal{S}^{CF}$ respectively, where $\mathcal{S}^{CB}, \mathcal{S}^{CF} \subseteq \mathcal{S}_M$ and possibly $\mathcal{S}^{CB} \bigcap \mathcal{S}^{CF} \neq \emptyset$.

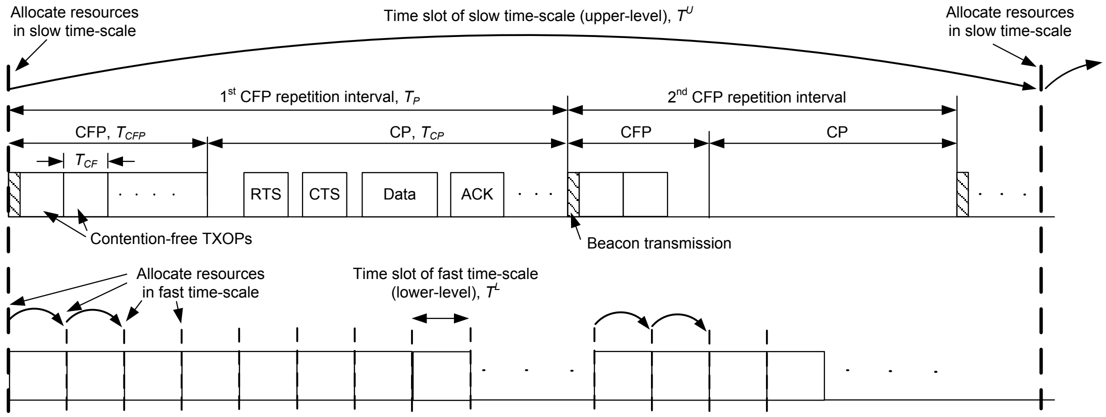

Allocate resources in slow time-scale
Time slot of slow time-scale (upper-level), T^U
Allocate resources in slow time-scale
1st CFP repetition interval, T_P
2nd CFP repetition interval
CFP, T_{CFP}
CP, T_{CP}
CFP
CP
T_{CF}
RTS
CTS
Data
ACK
Contention-free TXOPs
Beacon transmission
Allocate resources in fast time-scale
Time slot of fast time-scale (lower-level), T^L

Figure 4.2: Resource allocation at slow and fast time-scales.

4.2.1 Two Time-Scale Resource Allocation Framework

Resource allocation intervals of existing cellular networks are shorter than those of the existing WLANs, as cellular networks and WLANs are designed to support high mobility and low mobility users, respectively [13, 14]. Therefore, as shown in Fig. 4.2, resources in the cellular network are allocated at a faster time-scale than in the WLAN. The duration of a time slot in a time-scale is the resource allocation interval of the corresponding network, denoted by T^L and T^U in the fast and slow time-scales ($T^L < T^U$) for the cellular network and the WLAN, respectively. The resource allocation processes at fast and slow time-scales are referred to as lower and upper levels of the resource allocation process, respectively.

As the WLAN resource allocation interval is relatively long, to satisfy the strict delay and jitter requirements of periodically arriving constant bit rate voice traffic, several short CFPs are used within a resource allocation interval of the WLAN instead of using a long CFP [55]. For simplicity, assume $V_L (= T^U/T^L)$ is an integer and the boundaries of the first time slots in the two time-scales are aligned.

4.2.2 Symbols and Notations for the Chapter

Since there is a large number of symbols used in this chapter, following notations are used for clarity of the symbols. The lth ($l \in \{0, 1, ..., V_L - 1\}$) fast time-scale time slot within the uth slow time-scale time slot is referred to as (u, l)th time slot. Commonly used symbols are written in the form of $X^n_{i,y}$ or $X^n_{i,y}(\cdot)$, where superscript n, $n \in \{C, W, CB, CF, L, U\}$, represents the network or the level of resource allocation process. Superscripts C, W, CB and CF denote the cellular network, WLAN, contention-based channel access and contention-free channel access respectively, while L and U denote lower and upper levels of the resource allocation process respectively. The subscripts denote the user, a particular resource of network n, and a time slot. When $n \in \{W, CB\}$, only one subscript is used representing the user. Boldface letters are used for vectors and matrices, and vector $\mathbf{X}$ is represented as $\mathbf{X} = \{X_1, ..., X_{|\mathbf{X}|}\}$ with $|\mathbf{X}|$ being the number of elements in $\mathbf{X}$. The optimum value of variable X is denoted by X^*. The active (or the determined) decision policy $\mathbb{X}$ and the optimal set $\mathbf{X}$ are denoted by $\mathbb{X}^*$ and $\mathbf{X}^*$, respectively.

4.2.3 Traffic Model

The traffic generated by each user can be divided into two classes: constant bit rate voice and delay tolerant data. Every user always has at least one packet in the data traffic queue to transmit. The minimum data rates of voice and data traffic classes required by the ith user are denoted by $R_{Vmin,i}$ and $R_{Dmin,i}$, respectively. As voice traffic flows are highly susceptible to delay and jitter, voice traffic requirements are satisfied in average sense over each time slot at the slow time-scale. The data traffic requirements are satisfied in average sense over an infinite time horizon due to their delay tolerance.

4.2.4 Channel Model

Wireless channels are modeled as a finite-state Markov process to capture the channel time-correlation [56]. The channel gain is time invariant (i.e., quasi-static fading) within each coherence time (T_{coh}) interval. The different wireless channels vary independently from each other. The channel gain domain is partitioned into K_S non-overlapping states. The transition probabilities between different states of a Rayleigh fading channel can be calculated as in [56], assuming that T^U and T^L are not longer than the corresponding channel coherence times to ensure the states do not change within a time slot.

4.2.5 Subcarrier and Contention-Free TXOP allocations, and User Throughputs

Expressions that calculate user throughputs via the cellular network, and contention-free channel access and contention-based channel access of the WLAN are presented in Sections 3.2.1, 3.3.1 and 3.3.2, respectively. Further modifications to those expressions, and additional constraints to ensure that subcarriers and contention-free TXOPs are allocated without causing CCI are discussed in this section.

Define a variable $\rho^n_{i,y}$ such that $\rho^n_{i,y} = 1$ if the ith user is allocated the yth resource of network $n \in \{C, CF\}$, i.e., the yth OFDM subcarrier or TXOP; and $\rho^n_{i,y} = 0$ otherwise. Furthermore, $\rho^{CF}_{i,y} = 0, \forall y$ if $i \notin \mathcal{S}_M$. As each resource is allocated to only one user to avoid CCI,

$$\sum_{i \in \mathcal{S}_N} \rho^n_{i,y} \leq 1 \ , \ \forall y \in \mathcal{K}^n. \tag{4.1}$$

Then, from (3.1) and (3.2), the maximum achievable error free data rate by the ith user using the yth resource of network n can be expressed by

$$R^n_{i,y}(P^n_{i,y}) = \sum_{y \in \mathcal{K}^n} \rho^n_{i,y} B \log_2(1 + \alpha^n_{i,y} P^n_{i,y}), \tag{4.2}$$

where B represents the bandwidth of WLAN (B^W) or bandwidth of an OFDM subcarrier ($\Delta f = B^C / |\mathcal{K}^C|$); and B^C is the system bandwidth of the cell. Furthermore, as explained in Section (3.3.1), $\alpha^{CF}_{i,y} = \alpha^W_i, \forall y$ over each channel coherence time interval in the WLAN.

4.2.6 Power Usage of Multi-homing UEs

The operating time of a UE is governed by the energy (or average power) consumption of the uplink communications through WLAN and cellular interfaces of the UE [11]. Therefore, we limit the total average power consumption of each UE over each time slot in the slow time-scale to a predefined maximum. The average power usage through the WLAN interface for contention-based channel access (i.e., $P^{CB}_{avg,i}(\mathbf{P}^{CB})$) is given by (3.6). Then, the constraint on the total average power consumption of each UE over the uth time slot can then be expressed as

$$P^{C}_{avg,i} + P^{CB}_{avg,i}(\mathbf{P}^{CB}) + \frac{T_{CF}}{T_P} \sum_{j \in \mathcal{K}^{CF}} \rho^{CF}_{i,j} P^{CF}_{i,j} \leq P_{T,i}, \forall i \in \mathcal{S}_N, \tag{4.3}$$

where $P^{C}_{avg,i}$ is the average power usage through the cellular interface during the time slot and $P_{T,i}$ is the total average power available for the ith user.

4.3 MMDP-Based Optimal Resource Allocation

The objective of resource allocation is to maximize the total throughput of the interworking system subject to the satisfaction of QoS requirements. As discussed in Section 4.2.1, the resource allocation process consists of two (upper and lower) levels operating at slow and fast time-scales respectively, based on the channel state information. Resources of the WLAN and the cellular network are allocated at the beginnings of the uth and the (u, l)th time slots respectively, where $u = \{0, 1, 2, ...\}$ and $l = \{0, ..., V_L - 1\}$. The set of channel gains of the channels between users in $\mathcal{S}_M$ and the WLAN AP at the beginning of the uth time slot is referred to as the state of the upper-level during the uth time slot (ψ^U_u). The set of channel gains of the channels between all the users and the cellular BS at the beginning of the (u, l)th time slot is referred to as the state of the lower-level during the (u, l)th time slot ($\psi^L_{u,l}$). While the system state $\{\psi^U_u, \psi^L_{u,l}\}$ is denoted by $\psi_{u,l}$, the sets of all the possible states of upper and lower levels are denoted by Ψ^U and Ψ^L, respectively.

An overview of the resource allocation framework is shown in Fig. 4.3. The optimal resource allocation problem for cellular/WLAN interworking is formulated as an MMDP [18] for three reasons: 1) the resource allocation process operates at two time-scales as explained in Section 4.2.1, 2) state transition of each level is a Markov process due to the Markov channel model, and 3) resource allocations at multiple time slots are jointly optimized to satisfy the user QoS requirements over multiple time slots (see Section 4.2.3).

The MMDP formulation consists of upper and lower level resource allocation policies [18]. As shown in Fig. 4.3, the decisions of the upper-level are made considering the throughputs achieved through and the power consumed at the lower-level. Therefore, the

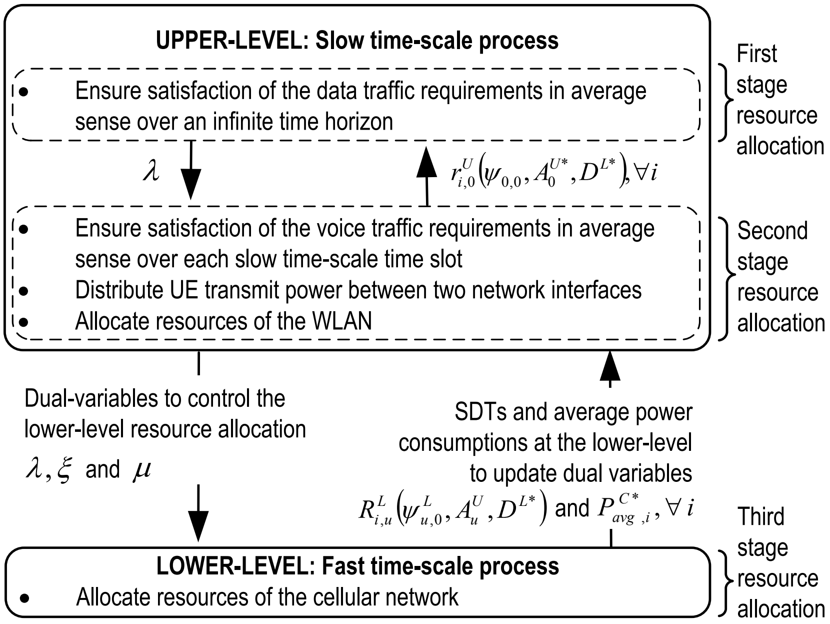

UPPER-LEVEL: Slow time-scale process
• Ensure satisfaction of the data traffic requirements in average sense over an infinite time horizon
First stage resource allocation
λ
$r_{i,0}^{U}\left(\psi_{0,0}, A_0^{U*}, D^{L*}\right), \forall i$
• Ensure satisfaction of the voice traffic requirements in average sense over each slow time-scale time slot
• Distribute UE transmit power between two network interfaces
• Allocate resources of the WLAN
Second stage resource allocation
Dual-variables to control the lower-level resource allocation
λ, ξ and μ
SDTs and average power consumptions at the lower-level to update dual variables
$R_{i,u}^{L}\left(\psi_{u,0}^{L}, A_u^{U}, D^{L*}\right)$ and $P_{avg,i}^{C*}, \forall i$
Third stage resource allocation
LOWER-LEVEL: Fast time-scale process
• Allocate resources of the cellular network

Figure 4.3: Overview of the MMDP-based two time-scale resource allocation framework.

upper-level policy ($\mathbb{D}^U$) maps system state $\psi_{u,0}$ to a set of resource allocation decisions (A_u^U) at the beginning of uth time slot, $u = \{0, 1, 2, ...\}$. The lower-level policy ($\mathbb{D}^L$) maps state $\psi_{u,l}^L$ to a set of resource allocation decisions ($A_{u,l}^L$) at the beginning of (u, l)th time slot, $l = \{0, ..., V_L - 1\}$. Decisions in A_u^U and $A_{u,l}^L$ are $\{P_i^{CB}, P_{i,j}^{CF}, \rho_{i,j}^{CF} | \forall i \in \mathcal{S}_M, j \in \mathcal{K}^{CF}\}$ and $\{P_{i,k}^C, \rho_{i,k}^C | \forall i \in \mathcal{S}_N, k \in \mathcal{K}^C\}$, respectively. For notation simplicity, we use $\mathbb{D}$ to denote the system policy $\{\mathbb{D}^U, \mathbb{D}^L\}$.

We use the summation of discounted throughputs (SDTs) [57, 58] over an infinite time horizon as a reward (objective) function. The SDT based reward function reduces the susceptibility of the determined decision policies to the unpredictable channel changes in the future by giving less importance to those decisions made (and rewards achieved) at far future. The SDTs achieved by the ith user at the upper-level over an infinite time horizon with the initial state of $\psi_{0,0}$ and at the lower-level during the uth time slot with the initial state of $\psi_{u,0}^L$ are denoted by $R_i^U(\psi_{0,0}, \mathbb{D})$ and $R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^L)$, respectively [18]. As the decision policies are stationary (to be discussed), $R_i^U(\psi_{0,0}, \mathbb{D})$ and $R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^L)$ can be interpreted as the average throughputs that are achieved by the ith user over the same periods of time at the upper and lower levels, respectively [57]. They are given by [57, 58]

$$R_i^U(\psi_{0,0}, \mathbb{D}) = \lim_{V_U \to \infty} (1-\theta) \sum_{u=0}^{V_U-1} \theta^u r_{i,u}^U(\psi_{u,0}, A_u^U, \mathbb{D}^L) \tag{4.4}$$

and

$$R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^L) = (1-\beta) \sum_{l=0}^{V_L-1} \beta^l r_{i,u,l}^L(\psi_{u,l}^L, A_u^U, A_{u,l}^L), \tag{4.5}$$

where $\theta, \beta \in (0,1)$ are discount factors; and $r_{i,u}^U(\psi_{u,0}, A_u^U, \mathbb{D}^L)$ and $r_{i,u,l}^L(\psi_{u,l}^L, A_u^U, A_{u,l}^L)$, which denote the throughputs achieved by the ith user at the upper and lower levels during the uth and (u,l)th time slots respectively, are given by [59]

$$r_{i,u}^U(\psi_{u,0}, A_u^U, \mathbb{D}^L) = \begin{cases} R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^L), \text{ if } i \in \mathcal{S}_S; \\ R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^L) + \frac{T_{CF}}{T_P} \sum_{j \in \mathcal{K}^{CF}} \rho_{i,j}^{CF} R_{i,j}^{CF}(P_{i,j}^{CF}), \text{if } i \in \mathcal{S}_M \setminus \mathcal{S}^{CB}; \\ R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^L) + \frac{T_{CF}}{T_P} \sum_{j \in \mathcal{K}^{CF}} \rho_{i,j}^{CF} R_{i,j}^{CF}(P_{i,j}^{CF}) + \frac{T_{CP}}{T_P} R_i^{CB}(\mathbf{P}^{CB}), \text{ if } i \in \mathcal{S}^{CB}; \end{cases} \tag{4.6}$$

and

$$r_{i,u,l}^L(\psi_{u,l}^L, A_u^U, A_{u,l}^L) = \sum_{k \in \mathcal{K}^C} \rho_{i,k}^C R_{i,k}^C(P_{i,k}^C). \tag{4.7}$$

The data traffic requirements of the users are served through both networks while the voice traffic requirements are served through the contention-free channel access and the cellular network. Therefore, the QoS constraints (see Section 4.2.3), which ensure data and voice traffic requirement satisfaction over an infinite time horizon and over the uth time slot ($u = \{0, 1, 2, ...\}$) respectively, can be stated as

$$R_i^U(\psi_{0,0}, \mathbb{D}) \geq R_{Vmin,i} + R_{Dmin,i} \ , \ \forall i \in \mathcal{S}_N \tag{4.8}$$

and

$$R_{i,u}^{L}(\psi_{u,0}^{L}, A_{u}^{U}, \mathbb{D}^{L}) + \frac{T_{CF}}{T_{P}} \sum_{j\in\mathcal{K}^{CF}} \rho_{i,j}^{CF} R_{i,j}^{CF}(P_{i,j}^{CF}) \geq R_{Vmin,i} \ , \ \forall i \in \mathcal{S}_{N}. \tag{4.9}$$

As the sum of discounted costs provides a good approximation for the average cost when the policies are stationary [57], $P_{avg,i}^{C}$ over the uth time slot can be calculated by

$$P_{avg,i}^{C} = (1-\beta) \sum_{l=0}^{V_{L}-1} \beta^{l} P_{tot,i,l}^{C}(\psi_{u,l}^{L}, A_{u}^{U}, A_{u,l}^{L}), \tag{4.10}$$

where $P_{tot,i,l}^{C}(\psi_{u,l}^{L}, A_{u}^{U}, A_{u,l}^{L})$ is the total power allocated by the ith user to communicate over the cellular network during the (u,l)th time slot, and is also equivalent to $\sum_{k\in\mathcal{K}^{C}} \rho_{i,k}^{C} P_{i,k}^{C}$ over the (u,l)th time slot.

The MMDP based optimal resource allocation problem can then be stated as [18, 57]

$$\begin{aligned} \mathcal{P}1: \quad & \max_{\mathbb{D}^{U}} \max_{\mathbb{D}^{L}} \quad \sum_{i\in\mathcal{S}_{N}} R_{i}^{U}(\psi_{0,0}, \mathbb{D}) \\ & \text{s.t.} \qquad\quad (4.1) \text{ for } n \in \{C, CF\}, (4.3), (4.8) \text{ and } (4.9). \end{aligned}$$

To find the optimal $\mathbb{D}^{U}$ and $\mathbb{D}^{L}$ solving problem $\mathcal{P}1$, resource allocation should be optimized over three different time intervals: 1) resource allocation over an infinite time horizon is optimized to satisfy (4.8), 2) resource allocation over each upper-level time slot is optimized to optimally use upper-level resources while satisfying (4.3) and (4.9), and 3) resource allocation over each lower-level time slot is optimized to optimally use lower-level resources. Therefore, problem $\mathcal{P}1$ is solved in three stages, where the first, second and third stages allocate resources over an infinite time horizon, for each upper-level time slot, and for each lower-level time slot, respectively. The resource allocation problem for the mth stage ($m = \{2, 3\}$) is derived by decomposing the $(m-1)$th stage problem into a set of problems, each of which allocates resources over the resource allocation interval of the mth stage, and by imposing constraints that must be satisfied within the resource allocation interval of the mth stage.

The optimality of the solution, which is obtained using the three stage approach, for problem $\mathcal{P}1$ is ensured by iterating the mth stage ($m = \{1, 2\}$) solution until it reaches the optimal while calculating the optimum $(m+1)$th stage solution for each mth stage iteration. During the iteration process, the dual variables of the mth stage are passed to the $(m+1)$th stage while the throughputs/SDTs achieved and power consumed at the $(m+1)$th stage are feedback to the mth stage. At the $(m+1)$th stage, the received dual variables are used for configuring the objective function of the resource allocation problem such that the $(m+1)$th stage assists maximizing the mth stage objective. At the mth stage, the received information is used for updating the dual variables.

Problem $\mathcal{P}1$ is a non-convex problem. Therefore, we determine the policies by relaxing problem $\mathcal{P}1$ to reduce the computational complexity. Due to the relaxations, the policies determined in this work ($\mathbb{D}^{U*}$ and $\mathbb{D}^{L*}$) are not optimal for problem $\mathcal{P}1$ in certain scenarios. Therefore, we refer to $\mathbb{D}^{U*}$ and $\mathbb{D}^{L*}$ as active (or determined) upper and lower level decision policies, respectively. Derivations of $\mathbb{D}^{U*}$, which is found by solving the first and second stage resource allocation problems, and $\mathbb{D}^{L*}$, which is found by solving the third stage resource allocation problem, are discussed in Sections 4.4 and 4.5, respectively.

Using the state transition probabilities calculated based on the channel statistics, $\mathbb{D}^{U*}$ and $\mathbb{D}^{L*}$ can be determined in advance and applied to the system based on the initial states. The applied policies select A_u^{U*} and $A_{u,l}^{L*}$ for the uth and the (u,l)th time slots respectively, based on the states of the two levels during the time slots. The policies $\mathbb{D}^{U*}$ and $\mathbb{D}^{L*}$ are required to be recalculated when the channel statistics or the number of users in the system or their QoS requirements change.

4.4 Upper-Level Resource Allocation

To determine $\mathbb{D}^{U*}$, first we maximize the total SDT at the upper-level subject to satisfaction of (4.8) over an infinite time horizon with the initial system state of $\psi_{0,0}$; this first stage problem is denoted by $\mathcal{P}2$. By further investigating problem $\mathcal{P}2$, we find out that problem $\mathcal{P}2$ is a convex optimization problem and can be solved by solving the dual problem [60]. To find the dual function, the minimum of the Lagrangian is determined by decomposing the Lagrangian into a set of terms, each of which is a negative summation of weighted throughputs of the users corresponding to one time slot. Then, A_u^{U*} for the uth time slot ($u = \{0, 1, 2, ...\}$) is determined such that it maximizes the summation of weighted throughputs corresponding to the uth time slot subject to satisfaction of (4.1) for $n = CF$, (4.3) and (4.9); this second stage problem is denoted by $\mathcal{P}3$. First and second stage problems are solved in Sections 4.4.1 and 4.4.2, respectively. In addition, the conditions which the third stage resource allocation at the lower-level should satisfy to ensure the optimality of the three stage solution for problem $\mathcal{P}1$ are derived in Section 4.4.2.

4.4.1 First Stage Resource Allocation

First stage resource allocation problem can be stated as

$$\begin{aligned}\mathcal{P}2: \quad \max_{\mathbb{D}^U} \quad & \sum_{i\in\mathcal{S}_N} R_i^U(\psi_{0,0}, \mathbb{D}^U, \mathbb{D}^{L*}) \\ \text{s.t.} \quad & \text{C1}: (4.8).\end{aligned}$$

The active policy $\mathbb{D}^{L*}$ is used in problem $\mathcal{P}2$ as for each iteration of the algorithm which solves problem $\mathcal{P}2$, $\mathbb{D}^{L*}$ is calculated by solving the third stage resource allocation problem. From (4.4)−(4.7), the objective function of problem $\mathcal{P}2$ is a concave function, and the feasible region is a convex set. Therefore, problem $\mathcal{P}2$ is a convex optimization problem, and is solved by maximizing the dual function which is obtained by minimizing the Lagrangian of problem $\mathcal{P}2$ with respect to $\mathbb{D}^U$ [60]. The Lagrangian of problem $\mathcal{P}2$ is

$$L^U(\psi_{0,0}, \boldsymbol{\lambda}, \mathbb{D}^U, \mathbb{D}^{L*}) = \sum_{i\in\mathcal{S}_N} \Big[\lambda_i\big(R_{Vmin,i} + R_{Dmin,i}\big) - \big(1+\lambda_i\big)R_i^U(\psi_{0,0}, \mathbb{D}^U, \mathbb{D}^{L*})\Big], \tag{4.11}$$

where $\lambda_i, \forall i$ are dual variables.

The iterative algorithm which solves $\mathcal{P}2$ can be summarized as follows. First, $\boldsymbol{\lambda}$ is initialized (e.g., $\boldsymbol{\lambda} \leftarrow \{0, ..., 0\}$). Second, we find $\mathbb{D}^U$ which minimizes $L^U(\psi_{0,0}, \boldsymbol{\lambda}, \mathbb{D}^U, \mathbb{D}^{L*})$ for the $\boldsymbol{\lambda}$. To update $\boldsymbol{\lambda}$ for the next iteration, $R_i^U(\psi_{0,0}, \mathbb{D}^U, \mathbb{D}^{L*}), \forall i$ are also found in this step. Third, $\boldsymbol{\lambda}$ is adjusted toward $\boldsymbol{\lambda}^*$ using the subgradient method [11, 61, 62]. The second and the third steps are repeated until $\boldsymbol{\lambda}$ reaches $\boldsymbol{\lambda}^*$. When $\boldsymbol{\lambda}$ reaches $\boldsymbol{\lambda}^*$, each λ_i satisfies the complementary slackness condition [60] and we have found $\mathbb{D}^{U*}$.

To implement the above algorithm, $\mathbb{D}^U$ and $R_i^U(\psi_{0,0}, \mathbb{D}^U, \mathbb{D}^{L*}), \forall i$ for any $\boldsymbol{\lambda}$ can be calculated as follows. From (4.11) and since $\sum_{i\in\mathcal{S}_N} \lambda_i\big(R_{Vmin,i}+R_{Dmin,i}\big)$ does not depend on $\mathbb{D}^U$, $\mathbb{D}^U$ is determined such that it maximizes $\sum_{i\in\mathcal{S}_N} \big(1+\lambda_i\big)R_i^U(\psi_{0,0}, \mathbb{D}^U, \mathbb{D}^{L*})$. When

$\sum_{i\in\mathcal{S}_N}\big(1+\lambda_i\big)R_i^U(\psi_{0,0},\mathbb{D}^U,\mathbb{D}^{L*})$ is maximized, by (4.4) and using the Bellman optimality equation [57], it is given by the following optimality equation.

$$\begin{aligned} L^U_{sup}(\psi_{0,0},\boldsymbol{\lambda}) = (1-\theta)\max_{A_0^U}\Big[\sum_{i\in\mathcal{S}_N}(1+\lambda_i)r^U_{i,0}(\psi_{0,0},A_0^U,\mathbb{D}^{L*})\Big] \\ +\theta\sum_{\psi_1^U\in\Psi^U}\sum_{\psi_{1,0}^L\in\Psi^L}\mathrm{P}^U_{\psi_0^U\psi_1^U}\mathrm{P}^L_{\psi_{0,0}^L\psi_{1,0}^L}L^U_{sup}(\psi_{1,0},\boldsymbol{\lambda})\end{aligned} \tag{4.12}$$

with

$$L^U_{sup}(\psi_{u,0},\boldsymbol{\lambda}) = \sup_{\mathbb{D}^U}\Big[\sum_{i\in\mathcal{S}_N}\big(1+\lambda_i\big)R_i^U(\psi_{u,0},\mathbb{D}^U,\mathbb{D}^{L*})\Big],$$

where $\mathrm{P}^U_{\psi_0^U\psi_1^U}$ and $\mathrm{P}^L_{\psi_{0,0}^L\psi_{1,0}^L}$ are the probabilities of the upper and lower level states change from ψ_0^U to ψ_1^U and from $\psi_{0,0}^L$ to $\psi_{1,0}^L$ at the end of the 0th time slot, respectively. Equation (4.12) is a recursive formula, and A_u^{U*} for the uth time slot is determined such that the summation of weighted throughputs given by $\sum_{i\in\mathcal{S}_N}(1+\lambda_i)r^U_{i,u}(\psi_{u,0},A_u^U,\mathbb{D}^{L*})$ is maximized [58]. Furthermore, these resource allocation problems corresponding to different time slots are independent of each other. As $\mathbb{D}^{U*}$ is a stationary policy (to be discussed in Section 4.4.2), finding A_0^{U*} for each $\psi_{0,0}\in\{\Psi^U,\Psi^L\}$ at the 0th time slot is sufficient to find $\mathbb{D}^{U*}$. Next, $R_i^U(\psi_{0,0},\mathbb{D}^U,\mathbb{D}^{L*})$ can be determined by solving the Bellman optimality equation for the ith user written using (4.4). Bellman optimality equation solving methods, such as Value Iteration algorithm and its variants, are explained in [58].

4.4.2 Second Stage Resource Allocation

We derive A_u^{U*} such that it maximizes $\sum_{i\in\mathcal{S}_N}(1+\lambda_i)r^U_{i,u}(\psi_{u,0},A_u^U,\mathbb{D}^{L*})$ at system state $\psi_{u,0}$ subject to satisfaction of (4.1) for $n=CF$, (4.3) and (4.9) during the uth time slot. This optimization problem is a non-convex problem due to the integer constraint which is imposed on $\rho^{CF}_{i,j}$ (see Section 4.2.5). Therefore, to reduce the computational complexity required to solve the problem, we relax the problem to be a convex optimization problem by relaxing the integer constraint such that $\rho^{CF}_{i,j}\in[0,1]$. To calculate power usage and throughputs over partially allocated TXOPs, we define $\bar{P}^{CF}_{i,j}=\rho^{CF}_{i,j}P^{CF}_{i,j}$ and $\bar{R}^{CF}_{i,j}(\bar{P}^{CF}_{i,j},\rho^{CF}_{i,j})=(T_{CF}/T_P)\rho^{CF}_{i,j}R^{CF}_{i,j}(\bar{P}^{CF}_{i,j}/\rho^{CF}_{i,j})$, where $\bar{R}^{CF}_{i,j}(\bar{P}^{CF}_{i,j},\rho^{CF}_{i,j})$ is a concave function [59]. For notation simplicity, we define $\bar{R}^{CB}_i(\mathbf{P}^{CB})=(T_{CP}/T_P)R^{CB}_i(\mathbf{P}^{CB})$.

Then, substituting from (4.6) to $r^U_{i,u}(\psi_{u,0}, A^U_u, \mathbb{D}^{L*})$, the relaxed problem can be stated as

$$
\begin{aligned}
\mathcal{P}3: \max_{A^U_u} \quad & \sum_{i\in\mathcal{S}_N}(1+\lambda_i)\Big[R^L_{i,u}(\psi^L_{u,0}, A^U_u, \mathbb{D}^{L*}) + \sum_{j\in\mathcal{K}^{CF}} \bar{R}^{CF}_{i,j}(\bar{P}^{CF}_{i,j}, \rho^{CF}_{i,j})\Big] \\
& + \sum_{i\in\mathcal{S}^{CB*}} (1+\lambda_i)\bar{R}^{CB}_i(\mathbf{P}^{CB}) \\
\text{s.t.} \quad & \text{C2}: \sum_{i\in\mathcal{S}_M} \rho^{CF}_{i,j} \leq 1 \ , \ \forall j \in \mathcal{K}^{CF} \\
& \text{C3}: R^L_{i,u}(\psi^L_{u,0}, A^U_u, \mathbb{D}^{L*}) + \sum_{j\in\mathcal{K}^{CF}} \bar{R}^{CF}_{i,j}(\bar{P}^{CF}_{i,j}, \rho^{CF}_{i,j}) \geq R_{Vmin,i} \ , \ \forall i \in \mathcal{S}_N \\
& \text{C4}: P^{CB}_{avg,i}(\mathbf{P}^{CB}) + P^C_{avg,i} + \frac{T_{CF}}{T_P}\sum_{j\in\mathcal{K}^{CF}} \bar{P}^{CF}_{i,j} \leq P_{T,i} \ , \ \forall i \in \mathcal{S}_N \\
& \text{C5}: 0 \leq \rho^{CF}_{i,j} \leq 1 \ , \ \forall i \in \mathcal{S}_M, j \in \mathcal{K}^{CF} \\
& \text{C6}: \bar{P}^{CF}_{i,j} \geq 0 \ , \ P^{CB}_i \geq 0 \ , \ \forall i \in \mathcal{S}_N, j \in \mathcal{K}^{CF}.
\end{aligned}
$$

Problem $\mathcal{P}3$ is a convex optimization problem. In Appendix A.2, convexity of C4, i.e., convexity of the set $\{P^C_{avg,i}, \bar{P}^{CF}_{i,j}, P^{CB}_i | \text{C4 is satisfied}, i \in \mathcal{S}_N, j \in \mathcal{K}^{CF}\}$, is proved.

Next, we illustrate the relationship between problem $\mathcal{P}3$ and the third stage resource allocation which determines $\mathbb{D}^{L*}$ for the lower-level. Then, we derive A^{U*}_u by solving problem $\mathcal{P}3$ using Karush-Kuhn-Tucker (KKT) conditions [60]. The Lagrangian for the problem $\mathcal{P}3$ can be written as

$$
\begin{aligned}
L^{U(2)}(A^U_u, \boldsymbol{\gamma}, \boldsymbol{\xi}, \boldsymbol{\mu}) = & -\sum_{i\in\mathcal{S}_N}(1+\lambda_i+\xi_i)\Big[R^L_{i,u}(\psi^L_{u,0}, A^U_u, \mathbb{D}^{L*}) + \sum_{j\in\mathcal{K}^{CF}} \bar{R}^{CF}_{i,j}(\bar{P}^{CF}_{i,j}, \rho^{CF}_{i,j})\Big] \\
+ \sum_{i\in\mathcal{S}_N}\Big[\sum_{j\in\mathcal{K}^{CF}} \gamma_j \rho^{CF}_{i,j} + \xi_i R_{Vmin,i} & + \mu_i\Big(P^{CB}_{avg,i}(\mathbf{P}^{CB}) + P^C_{avg,i} + \frac{T_{CF}}{T_P}\sum_{j\in\mathcal{K}^{CF}} \bar{P}^{CF}_{i,j} - P_{T,i}\Big)\Big] \\
& - \sum_{i\in\mathcal{S}^{CB*}}(1+\lambda_i)\bar{R}^{CB}_i(\mathbf{P}^{CB}) - \sum_{j\in\mathcal{K}^{CF}}\gamma_j,
\end{aligned}
\tag{4.13}
$$

where γ_j, ξ_i and $\mu_i, \forall i, j$ are the dual variables associated with C2, C3 and C4, respectively. As the optimal solution for problem $\mathcal{P}3$ minimizes (4.13) subject to C5 and C6, $\mathbb{D}^{L*}$ is determined such that it maximizes $\sum_{i\in\mathcal{S}_N}(1+\lambda_i+\xi_i)R^L_{i,u}(\psi^L_{u,0}, A^U_u, \mathbb{D}^L)$. Further, $\mathbb{D}^{L*}$ should satisfy the following KKT condition of $\mathcal{P}3$ to ensure the optimality of the solution, which is obtained using the three stages, for problem $\mathcal{P}1$.

$$(1+\lambda_i+\xi_i)\frac{\partial R^L_{i,u}(\psi^L_{u,0}, A^U_u, \mathbb{D}^{L*})}{\partial P^C_{avg,i}}\Big|_{P^C_{avg,i}=P^{C*}_{avg,i}} \begin{cases} = \mu_i, \text{ if } P^{C*}_{avg,i} > 0 \ ; \\ < \mu_i, \text{ otherwise;} \end{cases} \qquad \forall i \in \mathcal{S}_N. \quad (4.14)$$

Dual variables $\boldsymbol{\mu}$ and $\boldsymbol{\xi}$ couple the upper and the lower level resource allocations to optimally distribute the transmit power available at the UEs among WLAN and cellular network interfaces and to optimally utilize the resources of the two networks to satisfy the users' voice traffic requirements, respectively. Due to this coupling, once resources of the lower-level are allocated, achieved SDTs (i.e., $R^L_{i,u}(\psi^L_{u,0}, A^U_u, \mathbb{D}^{L*}), \forall i$) and the average power consumptions at the lower-level (i.e., $P^{C*}_{avg,i}, \forall i$) are feedback to the upper-level to solve problem $\mathcal{P}3$, as shown in Fig. 4.3.

Solution for problem $\mathcal{P}3$ is found as follows. First, $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$ are initialized. Second, the optimal allocations of contention-free TXOPs, UE transmit power levels during contention-free and contention-based TXOPs, $R^L_{i,u}(\psi^L_{u,0}, A^U_u, \mathbb{D}^{L*}), \forall i$ and $P^{C*}_{avg,i}, \forall i$ are calculated based on $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$. Third, $\boldsymbol{\mu}$ is updated toward $\boldsymbol{\mu}^*$ using the subgradient method [11, 61, 62]. The second and third steps are repeated until $\boldsymbol{\mu}^*$ is found. Forth, $\boldsymbol{\xi}$ is updated toward $\boldsymbol{\xi}^*$ using the subgradient method. The last three steps are repeated until $\boldsymbol{\xi}^*$ is found.

In the following, we derive the decision set $A^U_u{}^*(=\{P^{CB*}_i, \bar{P}^{CF*}_{i,j}, \rho^{CF*}_{i,j} | \forall i \in \mathcal{S}_M, j \in \mathcal{K}^{CF}\})$ by solving problem $\mathcal{P}3$, and explain how $\mathcal{S}^{CB*}$ is determined. In addition, the optimality of A^{U*}_u for the initial problem (i.e., the problem prior to the relaxation) is also discussed.

Allocations of Contention-free TXOPs and Transmit Power Levels

Based on the KKT conditions for $\mathcal{P}3$, the optimal transmit power levels of the users during contention-free TXOPs are given by

$$\bar{P}^{CF*}_{i,j} = \rho^{CF*}_{i,j}\Theta_i \ , \ \forall i \in \mathcal{S}_M, j \in \mathcal{K}^{CF}, \qquad (4.15)$$

where

$$\Theta_i = \left[\frac{B^W}{\ln(2)}\frac{(1+\lambda_i+\xi^*_i)}{\mu^*_i} - \frac{1}{\alpha^W_i}\right]^+$$

and $[x]^+ = \max\{0, x\}$. Next, the optimal contention-free TXOP allocation can be determined as follows. Let

$$\begin{aligned}\Gamma_{i,j} &= (1+\lambda_i+\xi_i^*)\frac{\partial \bar{R}_{i,j}^{CF}(\bar{P}_{i,j}^{CF},\rho_{i,j}^{CF})}{\partial \rho_{i,j}^{CF}}|_{\bar{P}_{i,j}^{CF}=\bar{P}_{i,j}^{CF*}} \\ &= \frac{(1+\lambda_i+\xi_i^*)T_{CF}B^W}{T_P}\Big[\log_2(1+\alpha_i^W\Theta_i) - \frac{1}{\ln(2)}\frac{\alpha_i^W\Theta_i}{1+\alpha_i^W\Theta_i}\Big],\ \forall i \in \mathcal{S}_M, j \in \mathcal{K}^{CF}.\end{aligned} \tag{4.16}$$

Due to the fact that $\Gamma_{i,j}$ is independent of $\rho_{i,j}^{CF}$ and from the KKT conditions, the jth TXOP is allocated to the user with the largest $\Gamma_{i,j}$ [59]. However, when there are multiple users with their $\Gamma_{i,j}$ values equal to the largest $\Gamma_{i,j}$ for the jth TXOP, the optimal solution for the problem $\mathcal{P}3$ allocates fractions of the TXOP among these users allowing them to time-share the TXOP.

Since the channel gain (or α_i^W) and Θ_i remain unchanged over the uth time slot, we can see from (4.16) that $\Gamma_{i,j}$ of the ith user is the same for all the TXOPs. Consequently, the ith user is allocated the same fraction from each TXOP or is allocated all the TXOPs. When there are N' users $\{i_1, i_2, ..., i_{N'}\}$ with their $\Gamma_{i,j}$ values equal to the largest, the optimal fractional values for $\rho_{i,j}^{CF}, i = \{i_1, i_2, ..., i_{N'}\}$ are determined based on the primal feasibility of those $\rho_{i,j}^{CF}$'s with respect to C2, C3 and C4. That is, the optimal set of $\rho_{i,j}^{CF}, i = \{i_1, i_2, ..., i_{N'}\}$ is a solution which satisfies C2 with equality and the following set of linear inequalities

$$\rho_{i,j}^{CF}|\mathcal{K}^{CF}|\frac{T_{CF}B^W\log_2(1+\alpha_i^W\Theta_i)}{T_P} \geq R_{Vmin,i} - R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^{L*})\ ,\ i = \{i_1, i_2, ..., i_{N'}\} \tag{4.17}$$

and

$$\rho_{i,j}^{CF}|\mathcal{K}^{CF}|\Theta_i\frac{T_{CF}}{T_P} \leq P_{T,i} - P_{avg,i}^{CB}(\mathbf{P}^{CB}) - P_{avg,i}^{C}\ ,\ i = \{i_1, i_2, ..., i_{N'}\}. \tag{4.18}$$

As the objective of this work is to allocate resources based on the PHY and the MAC technologies of the networks, a near optimal TXOP allocation for the initial problem is found by rounding $\rho_{i,j}^{CF}|\mathcal{K}^{CF}|$ values to the nearest integers. The rounded values indicate the number of TXOPs allocated to each user. Moreover, if $\rho_{i,j}^{CF}|\mathcal{K}^{CF}|, \forall i$ are integers, they are the optimal TXOP allocation for the initial problem.

Allocations of Users and Transmit Power Levels for Contention-based Channel Access

The second stage resource allocation problem should be formulated as a convex optimization problem to reduce the required computational capacity. However, $R_i^{CB}(\mathbf{P}^{CB})$ given by (3.4) is a non-concave function when N_W varies. Therefore, to formulate the second stage problem as a convex optimization problem, $\mathcal{S}^{CB*}$ should be determined prior to allocating the other upper-level resources. In the MMDP based resource allocation algorithm, $\mathcal{S}^{CB*}$ which achieves the highest total SDT at the upper-level is found via searching over $\mathcal{S}_M$. A low complexity method to find a near optimal $\mathcal{S}^{CB}$ is presented in Section 4.6.

From (3.4), it can be seen that $\bar{R}_i^{CB}(\mathbf{P}^{CB})$ depends not only on the ith user's transmit power level, but also on the transmit power levels of the other users in $\mathcal{S}^{CB*}$. Thus, $P_i^{CB*}, \forall i \in \mathcal{S}^{CB*}$ are correlated. Based on the KKT conditions for problem $\mathcal{P}3$, $P_i^{CB*} > 0$ only if

$$\frac{\partial \bar{R}_i^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}}\Big|_{\substack{\mathbf{P}_{-i}^{CB}=\mathbf{P}_{-i}^{CB*}\\ P_i^{CB}=0}} > \frac{\mu_i^*}{1+\lambda_i}\frac{\partial P_{avg,i}^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}}\Big|_{\substack{\mathbf{P}_{-i}^{CB}=\mathbf{P}_{-i}^{CB*}\\ P_i^{CB}=0}}\ , \tag{4.19}$$

where $\mathbf{P}_{-i}^{CB}$ is a vector which consists of the power levels of the users in $\mathcal{S}^{CB*}$ except the ith user. Since these partial derivatives are not defined when $P_i^{CB} = 0$, we rewrite (4.19) by taking the limits of the partial derivatives as $P_i^{CB} \to 0$. Then, (4.19) reduces to

$$P_i^{CB*}\begin{cases} > 0, & \text{if } \frac{B^W \alpha_i^W}{\ln(2)} > \frac{\mu_i^*}{1+\lambda_i}\ ; \\ = 0, & \text{otherwise;} \end{cases} \quad \forall i \in \mathcal{S}^{CB*}. \tag{4.20}$$

Moreover, for $P_i^{CB*} > 0$, the two sides of (4.19) become equal when the partial derivatives are evaluated at $P_i^{CB} = P_i^{CB*}$. Therefore, value of P_i^{CB*} when $P_i^{CB*} > 0$ can be found by solving (4.21) using Newton's method if $\mathbf{P}_{-i}^{CB*}$ is known [63].

$$\begin{aligned}\frac{1+\lambda_i}{\mu_i^*}\cdot\frac{N_W\alpha_i^W\bar{R}_i^{CB}(\mathbf{P}^{CB*})}{1+P_i^{CB*}\alpha_i^W} = \frac{T_{CP}}{T_P}\Big[\ln(1+P_i^{CB*}\alpha_i^W)-\frac{P_i^{CB*}\alpha_i^W}{1+P_i^{CB*}\alpha_i^W}\Big]\\ +\frac{\ln(2)N_W P_i^{CB*}\alpha_i^W\bar{R}_i^{CB}(\mathbf{P}^{CB*})}{B^W\cdot(1+P_i^{CB*}\alpha_i^W)\ln(1+P_i^{CB*}\alpha_i^W)}.\end{aligned} \tag{4.21}$$

Existence of a solution for (4.21) is proved in Appendix A.3.

As $P_i^{CB*}, \forall i \in \mathcal{S}^{CB*}$ are correlated, $\mathbf{P}^{CB*}$ is found using an iterative algorithm. In each iteration, $P_i^{CB*}, \forall i \in \mathcal{S}^{CB*}$ are calculated by (4.21) using the $\mathbf{P}^{CB*}$ calculated in the previous iteration. The algorithm terminates when the changes to $P_i^{CB*}, \forall i \in \mathcal{S}^{CB*}$ are negligible. Convergence of this iterative algorithm is proved in Appendix A.4.

From (4.15)−(4.21), it can be seen that A_u^{U*} is independent of the time slot (i.e., u). Therefore, A_u^{U*} which is determined for the uth time slot can be used at the u'th time slot ($u' = \{0, 1, 2, ...\}$) when states during the uth and the u'th time slots are equivalent (i.e., $\psi_{u,0} = \psi_{u',0}$). Thus, $\mathbb{D}^{U*}$ is a stationary policy [58]. The algorithm which determines $\mathbb{D}^{U*}$ for a given initial state $\psi_{0,0}$ is shown in Algorithm 1.

4.5 Lower-Level Resource Allocation

The objective of the lower-level resource allocation is to maximize $\sum_{i\in\mathcal{S}_N}(1+\lambda_i+\xi_i)R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^L)$ subject to (4.1) for $n = C$ and (4.14) over the uth time slot (see Section 4.4.2). To determine $\mathbb{D}^{L*}$, first we decompose the resource allocation problem over the uth time slot to a set of independent subproblems, each of which allocates resources for one lower-level time slot. Second, $A_{u,l}^{L*}$ for the (u, l)th time slot, $l = \{0, 1, ..., V_L - 1\}$, is found by solving the subproblem which corresponds to the same time slot; this third stage resource allocation problem is denoted by $\mathcal{P}4$ (see Section 4.3). Based on $A_{u,l}^{L*}$ found for the (u, l)th time slot, we show that $\mathbb{D}^{L*}$ is a stationary policy.

To decompose the resource allocation problem which maximizes $\sum_{i\in\mathcal{S}_N}(1+\lambda_i+\xi_i)R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^L)$ over the uth time slot to a set of independent subproblems, the Bellman optimality equation written for the lower-level using (4.5) with the assumption of V_L is very large is used. This assumption is reasonable as $T^L \ll T^U$. The Bellman

Algorithm 1 : Upper-Level Policy

```
input   : $\{\psi_0^U, \psi_{0,0}^L\}$, $\mathcal{S}_M$, $\mathcal{S}_S$, and $P_{T,i}, R_{Vmin,i}$ and $R_{Dmin,i}, \forall i$
output  : $A_u^{U*} = \{P_i^{CB*}, \rho_{i,j}^{CF*}, \bar{P}_{i,j}^{CF*} | \forall i \in \mathcal{S}_M, j \in \mathcal{K}^{CF}\}$ for every $\{\psi^U, \psi^L\} \in \{\Psi^U, \Psi^L\}$, and
          $\mathcal{S}^{CB*}$
while $\mathcal{S}^{CB}$ is not optimal do
    $\boldsymbol{\lambda} \leftarrow \{0, ..., 0\}$.
    while $\boldsymbol{\lambda}$ is not optimal do
        for each $\{\psi^U, \psi^L\} \in \{\Psi^U, \Psi^L\}$ do
            $\boldsymbol{\xi} \leftarrow \{0, ..., 0\}$ and $\boldsymbol{\mu} \leftarrow \{\mu_1, ..., \mu_N\}$.
            while $\boldsymbol{\xi}$ is not optimal do
                while $\boldsymbol{\mu}$ is not optimal do
                    Calculate $\bar{P}_{i,j}^{CF*}, \rho_{i,j}^{CF*}$ and $P_i^{CB*}$ by (4.15)−(4.18), (4.20) and (4.21) at state $\psi^U$.
                    Allocate resources at the Lower-Level by Algorithm 2 when the initial state is $\psi^L$.
                    Update $\mu_i, \forall i$.
                end
                Update $\xi_i, \forall i$.
            end
        end
        For each user, calculate $R_i^U(\psi_{0,0}, \mathbb{D})$ by solving the Bellman optimality equation written using (4.4).
        Update $\lambda_i, \forall i$.
    end
    Calculate the total SDT at the upper-level, $\sum_{i=1}^N R_i^U(\psi_{0,0}, \mathbb{D})$.
end
```

optimality equation is then given by [13, 14].

$$\sum_{i\in\mathcal{S}_N}(1+\lambda_i+\xi_i)R^L_{i,u}(\psi^L_{u,0},A^U_u,\mathbb{D}^{L*})=(1-\beta)\max_{A^L_{u,0}}\Big[\sum_{i\in\mathcal{S}_N}(1+\lambda_i+\xi_i)r^L_{i,u,0}(\psi^L_{u,0},A^U_u,A^L_{u,0})\Big]$$
$$+\beta\sum_{\psi^L_{u,1}\in\Psi^L}\mathrm{P}^{L(2)}_{\psi^L_{u,0}\psi^L_{u,1}}\sum_{i\in\mathcal{S}_N}(1+\lambda_i+\xi_i)R^L_{i,u}(\psi^L_{u,1},A^U_u,\mathbb{D}^{L*}),\tag{4.22}$$

where $\mathrm{P}^{L(2)}_{\psi^L_{u,0}\psi^L_{u,1}}$ is the probability of lower-level state changes from $\psi^L_{u,0}$ to $\psi^L_{u,1}$ at the end of the $(u,0)$th time slot. As the left hand side of (4.22) is maximized when $A^{L*}_{u,l}$ for the (u,l)th time slot maximizes $\sum_{i\in\mathcal{S}_N}(1+\lambda_i+\xi_i)r^L_{i,u,l}(\psi^L_{u,l},A^U_u,A^L_{u,l})$, resources of the lower-level (i.e., subcarriers and transmit power levels) are allocated for the (u,l)th time slot such that $\sum_{i\in\mathcal{S}_N}(1+\lambda_i+\xi_i)r^L_{i,u,l}(\psi^L_{u,l},A^U_u,A^L_{u,l})$ is maximized [58]. It should be noted that these resource allocation subproblems corresponding to the lower-level time slots are independent of each other.

Similar to the non-convexity caused by the integer constraint which is imposed on $\rho^{CF}_{i,j}$ (see Section 4.4.2), the integer constraint which is imposed on $\rho^C_{i,k}$ makes the subproblem corresponding to the (u,l)th time slot non-convex (see Section 4.2.5). To reduce the computational capacity required to solve the subproblem, we relax it by following the same relaxation process which is used in Section 4.4.2. That is, we let $\rho^C_{i,k}\in[0,1]$ and define $\bar{P}^C_{i,k}=\rho^C_{i,k}P^C_{i,k}$ and $\bar{R}^C_{i,k}(\bar{P}^C_{i,k},\rho^C_{i,k})=\rho^C_{i,k}R^C_{i,k}(\bar{P}^C_{i,k}/\rho^C_{i,k})$. The relaxed subproblem is considered as the problem $\mathcal{P}4$.

Since problem $\mathcal{P}4$ is solved subject to satisfaction of (4.14) over the uth time slot, (4.14) is first translated into a set of constraints, each corresponds to one lower-level time slot, by substituting (4.5), (4.7) and (4.10) into (4.14). Then, the constraint corresponding to the (u,l)th time slot is given by

$$(1+\lambda_i+\xi_i)\frac{\partial\bar{R}^C_{i,k}(\bar{P}^C_{i,k},\rho^C_{i,k})}{\partial\bar{P}^C_{i,k}}\Big|_{\bar{P}^C_{i,k}=\bar{P}^{C*}_{i,k}}\begin{cases}=\mu_i, & \text{if } \bar{P}^{C*}_{i,k}>0;\\ <\mu_i, & \text{otherwise;}\end{cases}\quad\forall i\in\mathcal{S}_N,k\in\mathcal{K}^C.\tag{4.23}$$

From (4.23),

$$\bar{P}^{C*}_{i,k}=\rho^{C*}_{i,k}\left[\frac{\Delta f}{\ln(2)}\frac{(1+\lambda_i+\xi_i)}{\mu_i}-\frac{1}{\alpha^C_{i,k}}\right]^+.\tag{4.24}$$

Next, the remaining subcarrier allocation problem for the (u,l)th time slot can be stated as

$$\begin{aligned} \mathcal{P}5:\ \max_{\boldsymbol{\rho}^C} \quad & \sum_{i\in\mathcal{S}_N}\sum_{k\in\mathcal{K}^C}(1+\lambda_i+\xi_i)\bar{R}^C_{i,k}(\bar{P}^C_{i,k},\rho^C_{i,k}) \\ \text{s.t.} \quad & \text{C7}: \sum_{i\in\mathcal{S}_N}\rho^C_{i,k}\le 1\ ,\ \forall k\in\mathcal{K}^C \\ & \text{C8}: 0\le\rho^C_{i,k}\le 1\ ,\ \forall i\in\mathcal{S}_N, k\in\mathcal{K}^C. \end{aligned}$$

Problem $\mathcal{P}5$ is a convex optimization problem. Therefore, from (4.24) and the KKT conditions for problem $\mathcal{P}5$, and using the same approach used for deriving $\rho^{CF*}_{i,j}$ in Section 4.4.2, the optimal subcarrier allocation can be expressed as [59]

$$\rho^{C*}_{i',k} = \begin{cases} 1, & \text{if } i' = \arg\max_{\forall i}\left\{\Lambda_{i,k}\right\}; \\ 0, & \text{otherwise;} \end{cases} \quad \forall i'\in\mathcal{S}_N, k\in\mathcal{K}^C, \tag{4.25}$$

where

$$\Lambda_{i,k} = (1+\lambda_i+\xi_i)\Big[\log_2(1+\alpha^C_{i,k}P^{C*}_{i,k}) - \frac{1}{\ln(2)}\frac{\alpha^C_{i,k}P^{C*}_{i,k}}{1+\alpha^C_{i,k}P^{C*}_{i,k}}\Big].$$

However, when there are multiple users with their $\Lambda_{i,k}$ values equivalent to the largest $\Lambda_{i,k}$ for the kth subcarrier, the optimal solution for the problem $\mathcal{P}5$ requires allocation of fractions, which satisfies C7 with equality, of the kth subcarrier among these users.

When such equality of $\Lambda_{i,k}$ occurs, we randomly allocate the kth subcarrier to one of the users with the largest $\Lambda_{i,k}$ due to the fact that fractional subcarrier allocations are not supported by the PHY. Random subcarrier allocation in this scenario does not significantly deviate the system throughput/QoS performance from the optimum due to two reasons: 1) subcarrier bandwidth is small as there is a large number of subcarriers; 2) probability of multiple users having equivalent $\Lambda_{i,k}$ values for more than one subcarrier or for a certain subcarrier over multiple time slots is very small, because the channel gains over different subcarriers are different and varies over time slots.

From (4.24) and (4.25), it can be seen that $A^{L*}_{u,l}$ is independent of the time slot. Thus, $A^{L*}_{u,l}$ which is determined for the (u,l)th time slot can be used at the (u,l')th time slot ($l'=\{0,...,V_L-1\}$) when states during these two time slots are equivalent (i.e., $\psi^L_{u,l}=\psi^L_{u,l'}$). Therefore, $\mathbb{D}^{L*}$ is a stationary policy [58].

As $\mathbb{D}^{L*}$ is a stationary policy, calculating $A^{L*}_{u,0}$ for each state $\psi^L_{u,0}\in\Psi^L$ at the $(u,0)$th

Algorithm 2 : Lower-Level Policy

```
input   : $\psi^L$, $\mathcal{S}_N$, $\boldsymbol{\lambda}, \boldsymbol{\xi}$ and $\boldsymbol{\mu}$
output  : $A^{L*}_{u,l} = \{\rho^{C*}_{i,k}, \bar{P}^{C*}_{i,k} | \forall i \in \mathcal{S}_N, k \in \mathcal{K}^C\}$ for each $\psi^{L(2)} \in \Psi^L$, and $R^L_{i,u}(\psi^L, A^U_u, \mathbb{D}^{L*})$ and
          $P^{C*}_{avg,i}, \forall i$
for each $\psi^{L(2)} \in \Psi^L$ do
    Calculate $\bar{P}^{C*}_{i,k}$ and $\rho^{C*}_{i,k}$ by (4.24) and (4.25) at state $\psi^{L(2)}$.
    $r^L_{i,u,l}(\psi^{L(2)}, A^U_u, A^{L*}_{u,l}) \leftarrow \sum_{k \in \mathcal{K}^C} \rho^{C*}_{i,k} R^C_{i,k}(P^{C*}_{i,k})$.
    $P^C_{tot,i,l}(\psi^{L(2)}, A^U_u, A^{L*}_{u,l}) \leftarrow \sum_{k \in \mathcal{K}^C} \bar{P}^{C*}_{i,k}$.
end
For each user, calculate $R^L_{i,u}(\psi^L, A^U_u, \mathbb{D}^{L*})$ by solving the Bellman optimality equation written
using (4.5).
Calculate $P^{C*}_{avg,i}, \forall i$ by (4.10).
```

time slot is sufficient to determine $\mathbb{D}^{L*}$. Next, $R^L_{i,u}(\psi^L_{u,0}, A^U_u, \mathbb{D}^{L*})$ and $P^{C*}_{avg,i}, \forall i$ can be found by solving (4.5) and (4.10), respectively. Equations (4.5) and (4.10) can be solved using the methods explained in [58] by writing them as Bellman optimality equations. Values of $R^L_{i,u}(\psi^L_{u,0}, A^U_u, \mathbb{D}^{L*})$ and $P^{C*}_{avg,i}, \forall i$ are then feedback to the upper-level to update $\boldsymbol{\lambda}$, $\boldsymbol{\xi}$ and $\boldsymbol{\mu}$ as shown in Algorithm 2, which determines $\mathbb{D}^{L*}$.

The MMDP based resource allocation algorithm (i.e., $\mathbb{D}^{U*}$ and $\mathbb{D}^{L*}$) efficiently allocates resources of the interworking system. However, it has a high time complexity as it requires to find A^{U*}_0 and $A^{L*}_{u,0}$ for each system state, where the total number of system states in a system model is given by $(K_S)^{N|\mathcal{K}^C|+|\mathcal{S}_M|}$. Therefore, we propose a heuristic resource allocation algorithm with low time complexity when the number of system states is significantly higher due to the large number of users and/or OFDM subcarriers.

4.6 Heuristic Resource Allocation

The heuristic algorithm consists of two steps. The first step is executed only once at the beginning, and it calculates the dual variables which correspond to data and voice traffic constraints (i.e., $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$) based on the average square channel gains (Ω's), where $\Omega = \mathbb{E}\{h^2\}$, $\mathbb{E}\{\cdot\}$ is the ensemble average operator and h is the channel gain. The second step uses the dual variables calculated in the first step, and allocates upper and lower level resources based on the instantaneous channel gains subject to total power constraints of the users (i.e., C4). Since these two steps are executed based on a single system state which consists of either Ω's or instantaneous channel gains, solving the Bellman optimality equations is not required in the heuristic algorithm for calculating $R_i^U(\psi_{0,0}, \mathbb{D})$, $R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^L)$ and $P_{avg,i}^C$. In the MMDP based algorithm, the Bellman optimality equations are solved by determining A_0^U and $A_{u,0}^L$ for each possible system state. In addition, $\mathcal{S}^{CB}$ is determined using a simple method in the heuristic algorithm (to be discussed). Due to these two reasons, the heuristic algorithm has low time complexity.

In the first step, $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$ are found by solving problems $\mathcal{P}2$, $\mathcal{P}3$ and $\mathcal{P}4$. Therefore, the solutions which we obtained for problems $\mathcal{P}3$ and $\mathcal{P}4$ in Sections 4.4.2 and 4.5 are used in this step with modifications to utilize Ω's as follows. Average throughput over a Rayleigh fading channel is given by [56, 64]

$$\mathbb{E}\{R\} = \int_0^\infty \frac{2Bh}{\Omega} \log_2(1 + \frac{h^2 p}{n}) \mathrm{e}^{-\frac{h^2}{\Omega}}\, \mathrm{d}h = \frac{B}{\ln(2)} \mathrm{e}^{\frac{n}{\Omega p}} \mathrm{E}_1\Big(\frac{n}{\Omega p}\Big), \tag{4.26}$$

where p is the transmit power level, B is the bandwidth, n is the total noise plus interference power, and $\mathrm{E}_1(x)$ is the exponential integral which is defined as [64]

$$\mathrm{E}_1(x) = \int_x^\infty \frac{\mathrm{e}^{-x}}{x} \mathrm{d}x. \tag{4.27}$$

Since $0.5\mathrm{e}^{-x}\ln(1+2x)$ provides a tight lower bound for $\mathrm{E}_1(x)$ [64], by (4.26)

$$\mathbb{E}\{R\} > \frac{B}{2} \log_2\Big(1 + \frac{2\Omega p}{n}\Big). \tag{4.28}$$

Thus, the solutions for the problems $\mathcal{P}3$ and $\mathcal{P}4$ are modified to calculate the throughput over each wireless channel by $(B/2)\log_2(1 + 2\Omega p/n)$. That is, the equations in Sections 4.4.2 and 4.5 are modified with the substitutions of $B/2$ to B, and 2Ω to h^2. The latter is also equivalent to the substitution of $2\mathbb{E}\{\alpha_{i,y}^n\}$ to $\alpha_{i,y}^n$. Moreover, as Ω's are used in this step, the number of possible system states in the MMDP reduces to one. Consequently, $R_i^U(\psi_{0,0}, \mathbb{D}) = r_{i,0}^U(\psi_{0,0}, A_0^U, \mathbb{D}^L)$, $R_{i,0}^L(\psi_{0,0}^L, A_0^U, \mathbb{D}^L) = r_{i,0,0}^L(\psi_{0,0}^L, A_0^U, A_{0,0}^L)$ and $P_{avg,i}^C = \sum_{k\in\mathcal{K}^C} \rho_{i,k}^C P_{i,k}^C$.

Furthermore, $\log_2(1+x) > 0.5\log_2(1+2x), \forall x \in \mathbb{R}^+$. Therefore, $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$ calculated in the first step will satisfy the QoS requirements with an additional margin if the throughputs are given by $B\log_2(1 + \Omega p/n)$. Therefore, when the resources are allocated

Algorithm 3 : First step of the heuristic algorithm

```
input    : Average square channel gains ($\Omega$'s), $\mathcal{S}_M$, $\mathcal{S}_S$, and $P_{T,i}, R_{Vmin,i}$ and $R_{Dmin,i}, \forall i$
output   : $\boldsymbol{\lambda}^*, \boldsymbol{\xi}^*$ and $\mathcal{S}^{CB*}$
Form $\psi_0^U$ and $\psi_{0,0}^L$ using $\sqrt{2\Omega}$ values of the channels.
$\alpha_{i,k}^C \leftarrow 2\mathbb{E}\{\alpha_{i,k}^C\}$ and $\alpha_i^W \leftarrow 2\mathbb{E}\{\alpha_i^W\}$.
Sort users in $\mathcal{S}_M$ in the descending order of their $\mathbb{E}\{\alpha_i^W\}$.
while $\mathcal{S}^{CB}$ corresponding to maximum $\sum_{i \in \mathcal{S}_N} R_i^U(\psi_{0,0})$ is not found do
    $\mathcal{S}^{CB} \leftarrow \mathcal{S}_M(|\mathcal{S}^{CB}| + 1)$.
    $\boldsymbol{\lambda} \leftarrow \{0, ..., 0\}$, $\boldsymbol{\xi} \leftarrow \{0, ..., 0\}$ and $\boldsymbol{\mu} \leftarrow \{\mu_1, ..., \mu_N\}$.
    while $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$ are not optimal do
        while $\boldsymbol{\mu}$ is not optimal do
            Calculate $\bar{P}_{i,j}^{CF*}, \rho_{i,j}^{CF*}, P_i^{CB*}, \rho_{i,k}^{C*}$ and $\bar{P}_{i,k}^{C*}$ by (4.15)$-$(4.18), (4.20), (4.21), (4.24)
            and (4.25) substituting $B^W/2$ to $B^W$ and $\Delta f/2$ to $\Delta f$.
            Update $\mu_i, \forall i$.
        end
        $R_{i,0}^L(\psi_{0,0}^L) \leftarrow \sum_{k \in \mathcal{K}^C} \bar{R}_{i,k}^C(\bar{P}_{i,k}^C, \rho_{i,k}^C)$.
        $P_{avg,i}^C \leftarrow \sum_{k \in \mathcal{K}^C} \bar{P}_{i,k}^C$.
        Calculate $r_{i,0}^U(\psi_{0,0})$ by (4.6), and $R_i^U(\psi_{0,0}) \leftarrow r_{i,0}^U(\psi_{0,0})$.
        Update $\lambda_i$ and $\xi_i, \forall i$.
    end
end
```

in the second step, QoS requirements of the users are satisfied with a higher satisfaction though the instantaneous channel gains are used in this step. First step of the heuristic algorithm is shown in Algorithm 3.

In the second step, resources of both upper and lower levels are jointly allocated at the beginning of uth time slot ($u = \{0, 1, 2, ...\}$) subject to C4 and assuming that the current lower-level state remains unchanged during the uth time slot (i.e., $\psi_{u,0}^L = \psi_{u,l}^L, \forall l \in \{1, ..., V_L - 1\}$). With this assumption, $R_{i,u}^L(\psi_{u,0}^L, A_u^U, \mathbb{D}^L) = r_{i,u,0}^L(\psi_{u,0}^L, A_u^U, A_{u,0}^L)$ and $P_{avg,i}^C = \sum_{k \in \mathcal{K}^C} \rho_{i,k}^C P_{i,k}^C$. The algorithms which solve problems $\mathcal{P}3$ and $\mathcal{P}4$ are used for allocating resources while using $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$ from the first step. Note that these algorithms need to calculate $\boldsymbol{\mu}$ only and that UE power is distributed between WLAN and cellular network interfaces at this point. At the beginnings of the remaining (u, l)th time slots

Algorithm 4 : Second step of the heuristic algorithm

```
input   : Instantaneous channel gains during $(u,l)$th time slot, $\boldsymbol{\lambda}^*, \boldsymbol{\xi}^*$ and $\mathcal{S}^{CB*}$
output  : $A^{U*} = \{P_i^{CB*}, \rho_{i,j}^{CF*}, \bar{P}_{i,j}^{CF*} | \forall i \in \mathcal{S}_M, j \in \mathcal{K}^{CF}\}$ and $A^{L*} = \{\rho_{i,k}^{C*}$ and $\bar{P}_{i,k}^{C*} | \forall i \in \mathcal{S}_N, k \in \mathcal{K}^C\}$
Form $\psi_u^U$ and $\psi_{u,l}^L$ using instantaneous channel gains during $(u,l)$th time slot.
while $\boldsymbol{\mu}$ is not optimal do
    if $l = 0$ then
        Calculate $\bar{P}_{i,j}^{CF*}, \rho_{i,j}^{CF*}, P_i^{CB*}, \rho_{i,k}^{C*}$ and $\bar{P}_{i,k}^{C*}$ by (4.15)–(4.18), (4.20) and (4.21), (4.24) and (4.25) based on $\psi_u^U$ and $\psi_{u,l}^L$.
    else
        Recalculate $\rho_{i,k}^{C*}$ and $\bar{P}_{i,k}^{C*}$ by (4.24) and (4.25) based on $\psi_{u,l}^L$.
    end
    Update $\mu_i, \forall i$
end
```

(i.e., $l = \{1, ..., V_L - 1\}$), lower-level resources of subcarriers and the amount of power dedicated for the cellular network interfaces are reallocated based on the current state $\psi_{u,l}^L$ to fully exploit the multi-user diversity in the cellular network. The second step of the heuristic algorithm is shown in Algorithm 4.

The optimal $\mathcal{S}^{CB}$ consists of only a few users with strong channel conditions due to two reasons: 1) allocation of too many users for contention-based channel access degrades the aggregated throughput of the users due to increased number of collisions [13], and 2) allocation of a user with a weaker channel degrades the throughputs of all the users as the weak user takes a longer time to transmit a packet [16]. Based on these characteristics, users for the contention-based channel access are allocated as follows. First, the users in $\mathcal{S}_M$ are sorted in the descending order of their $\mathbb{E}\{\alpha_i^W\}$. Next, the first step of the heuristic algorithm is repeated, each time adding the next user in $\mathcal{S}_M$ to $\mathcal{S}^{CB}$, until the total throughput achieved at the upper-level reaches the maximum.

4.7 Simulation Results

Wireless channels are modeled as Rayleigh fading channels, and their path loss is proportional to d^{-4}, where d denotes the distance between users and the WLAN AP or the

Table 4.1: Simulation Parameters

Parameter	Value (unit)
B^W	20MHz
D	4095 octets
$\|\mathcal{K}^C\|$	4 or 128
$\|\mathcal{K}^{CF}\|$	2 or 10
T^L	4.23ms
T^U, T_P	63.45ms
T_{ACK}	24.5μs
T_{AIFS}	34μs
T_{CF}	$31.72/\|\mathcal{K}^{CF}\|$ ms
T_{CP}	31.72ms
T_{CTS}	24.5μs
T_{RTS}	24.7μs
T_{SIFS}	16μs
Δf	$5/\|\mathcal{K}^C\|$ MHz
σ_0	9μs
Additive white Gaussian noise density	-174 dBm/Hz
Initial window size of WLAN	16
Maximum number of backoff stages in WLAN	6

cellular BS. Further, the channels over the cellular network are generated at the carrier frequency of 2.1GHz and mobile speed of 50kmh^{-1}, while those over the WLAN are generated at the carrier frequency of 2.4GHz and mobile speed of 3kmh^{-1}. Based on the coherence times of the channels, T^L and T^U are selected to be 4.23ms and 63.45ms, respectively [65, 66]. The radiuses of the WLAN and the cellular coverage areas are 50m and 1000m respectively, and users are uniformly distributed over the coverage areas. The total power available at each user is uniformly distributed between 0 and 1watt. Table 4.1 shows the remaining parameters.

First we evaluate the performance of the MMDP based resource allocation algorithm (MM) and the heuristic algorithm (HM) in a small-scale system, denoted by system-1, and compare the performance with that of a benchmark algorithm (BM1) which resembles the first category resource allocation algorithms (see Section 4.1). Algorithm BM1 allocates the resources as follows. First, it assigns users for the two networks via exhaustive search such that the total average system throughput is maximized. In this step, average users' throughputs are calculated using $(B/2)\log_2(1+2\Omega p/n)$, as in the first step of HM. Next, each network individually allocates its resources among the assigned users to maximize the network throughput. This step utilizes instantaneous channel gains and is repeated at the every time slot. It should be noted that BM1 does not allow UE multi-homing and it allocates resources at two time-scales based on the PHY and MAC technologies of the networks.

In the system-1, there are four users ($|\mathcal{S}_S| = |\mathcal{S}_M| = 2$), four subcarriers and two contention-free TXOPs. Two-state Markov channels are used [67], and the boundary between the two states of each channel is determined such that the steady state probability of each state is 0.5. For each channel, the channel is at the first state if $h < \sqrt{\Omega \ln(2)}$; otherwise, it is at the second state. When a channel is at first and second states, the channel gains are considered to be $\sqrt{\Omega(1 - \ln(2))}$ and $\sqrt{\Omega(1 + \ln(2))}$, respectively. These channel gains are determined by averaging the square of the channel gain of a continuous-envelope Rayleigh fading channel within the boundaries of the respective state. Transition probabilities of the states are calculated as in [56]. Discount factors θ and β are set to be 0.9 in MM.

Fig. 4.4 compares the throughputs achieved by MM, HM and BM1 in system-1 for different QoS requirements. Algorithm MM provides throughput improvement of at least 10.7% compared to HM, and both MM and HM provide higher throughputs than BM1 as they enable multi-homing. In BM1, each user is allowed to access one network only. When multi-homing is enabled, users achieve higher throughputs due to efficient resource utilization, which is a result of catering user QoS requirements utilizing multiple network resources and of dynamically adjusting resource allocation including UE power distribution for the two network interfaces based on the instantaneous channel gains. In addition, MM outperforms HM due to the fact that MM allocates the resources statistically considering the future state changes using an MMDP, whereas HM allocates the resources based on the current states only.

The satisfaction index (SI), which can be used for quantifying the ability of a resource allocation algorithm to satisfy the QoS requirements [62], is defined for a particular traffic class as

$$\mathrm{SI} = \mathbb{E}\left\{ 1_{R \geq R_{min}} + 1_{R_{min} > R} \cdot \frac{R}{R_{min}} \right\}, \tag{4.29}$$

where R and R_{min} are the achieved and the required throughputs, and $1_{x \geq y} = 1$ if $x \geq y$ but it is zero otherwise. All three algorithms have achieved SI for voice traffic (SI_V) of one in system-1, and the achieved SI's for data traffic (SI_D) by them in system-1 are

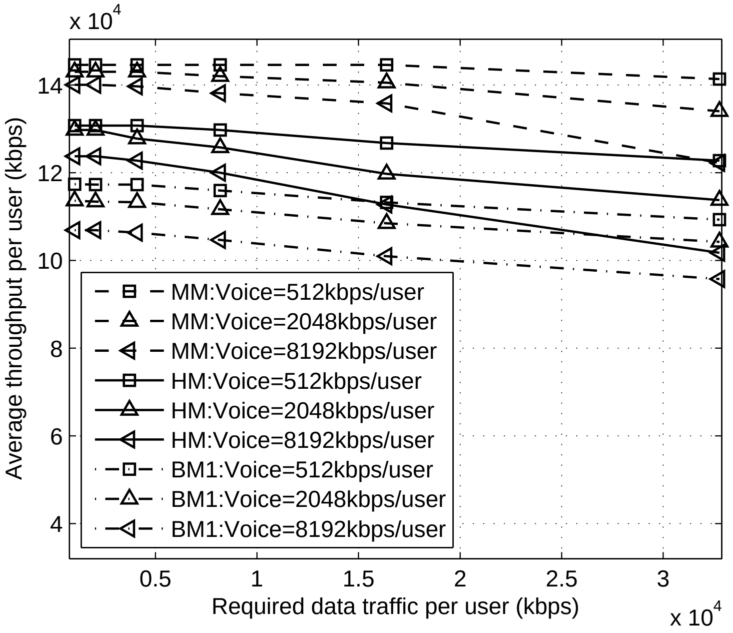


Figure 4.4: Throughputs achieved by different algorithms in system-1.

shown in Fig. 4.5. Similar to the throughput performance, MM and HM achieve higher SI_D's compared to BM1, providing users with better QoS. Difference between these SI_D's is significant at the higher data traffic requirements as multi-homing is particularly useful for catering higher user requirements via multiple networks.

Complexities of the algorithms are measured in terms of the required number of iterations in the inner most loop per user per time slot in fast time-scale, and Fig. 4.6 compares them in system-1. As MM solves an MMDP based resource allocation problem with 2^{18} system states, it requires a large number of iterations. Algorithm BM1 requires a higher number of iterations than HM as BM1 recalculates $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$ at each time slot, whereas HM calculates $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$ only once in the first step.

Next, the performance of HM is evaluated in a large-scale system, denoted by system-2, and is compared with the performance of a benchmark algorithm (BM2). Since the highest number of resource blocks per cell in a LTE system is 110, system-2 consists of 128 subcarriers, 10 contention-free TXOPs and 40 or 80 users. This system uses continuous envelope Rayleigh fading channels generated at the same carrier frequencies and mobile speeds as in system-1. The performance of MM is not evaluated in this system due its high complexity. Algorithm BM2 uses a simpler user allocation mechanism than

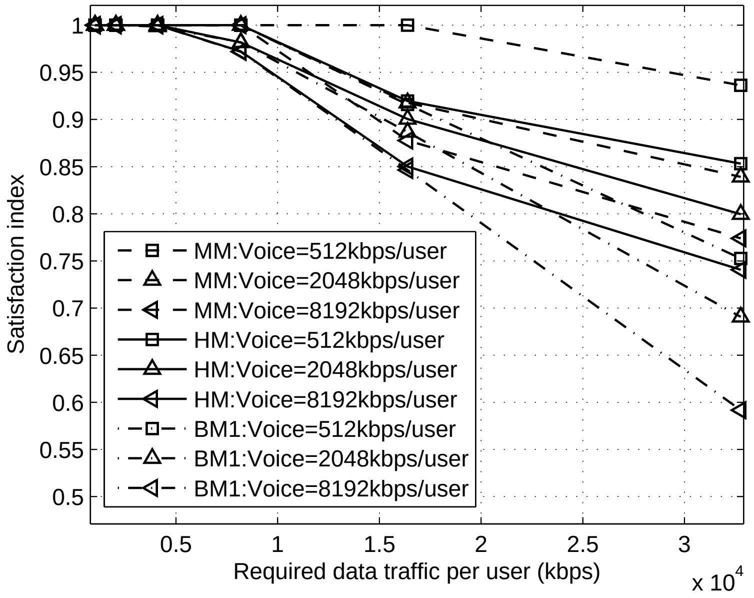


Figure 4.5: Satisfaction index achieved by different algorithms for data traffic in system-1.

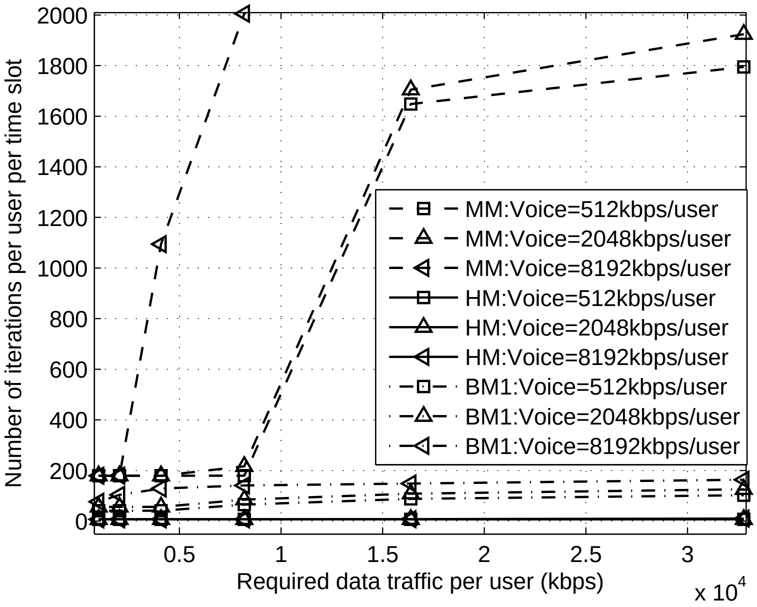


Figure 4.6: Complexities of the different algorithms in system-1.

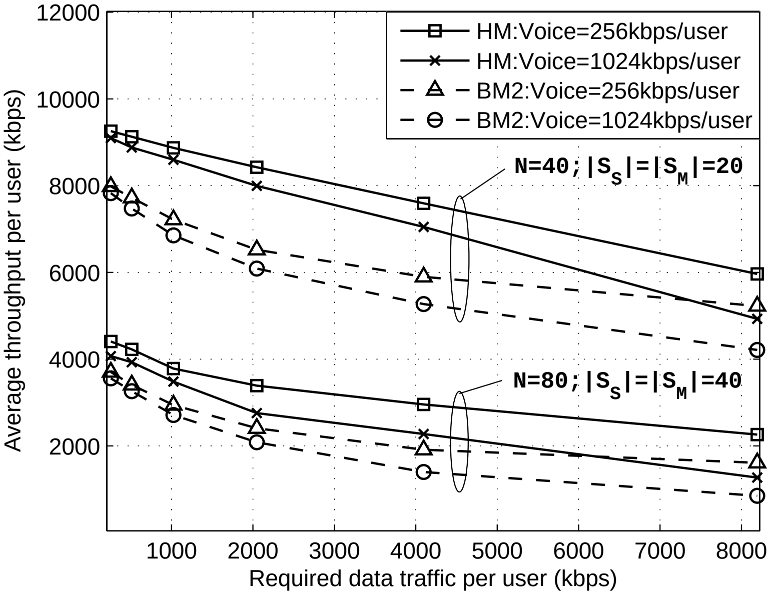


Figure 4.7: Throughputs achieved by different algorithms in system-2.

exhaustive search which is used in BM1, because exhaustive search is not feasible when there is a large number of users. It allocates users for the networks as follows. First, users in $\mathcal{S}_M$ are sorted in the descending order of their $\mathbb{E}\{\alpha_i^W\}$. Second, resources of the two networks are individually allocated $|\mathcal{S}_M|$ times while calculating users' average throughputs using $(B/2)\log_2(1+2\Omega p/n)$; at the jth resource allocation, first j users in $\mathcal{S}_M$ are allocated to the WLAN while the remaining users in $\mathcal{S}_M$ and all the users in $\mathcal{S}_S$ are allocated to the cellular network. Finally, the user allocation which resulted in the highest total average throughput is selected. Once the user allocation is completed, BM2 allocates resources of the two networks at each time slot similar to BM1.

Throughput, SI$_V$ and SI$_D$ performance of HM and BM2 in system-2 are shown in Fig. 4.7, Fig. 4.8 and Fig. 4.9, respectively. Due to the advantages of user multi-homing, HM provides better throughput and SI performance than BM2. The performance of the algorithms decreases with the number of users, because the resources are distributed among more users as each user has a certain QoS requirement. When there are 80 users in the system, HM provides at least 14.5%, 8.4% and 8.1% of improvements compared to BM2 in average throughput per user, SI$_V$ and SI$_D$, respectively.

According to the complexities of the algorithms shown in Fig. 4.10, HM converges

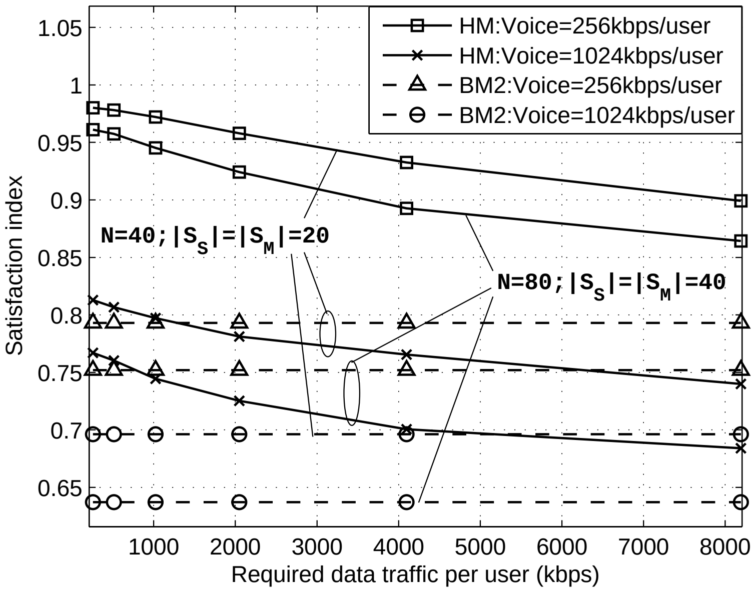


Figure 4.8: Satisfaction index achieved by different algorithms for voice traffic in system-2.

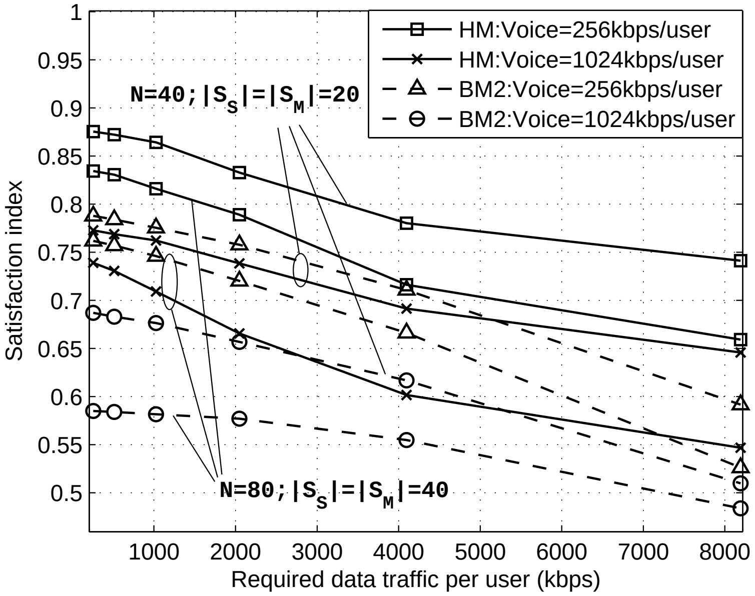


Figure 4.9: Satisfaction index achieved by different algorithms for data traffic in system-2.

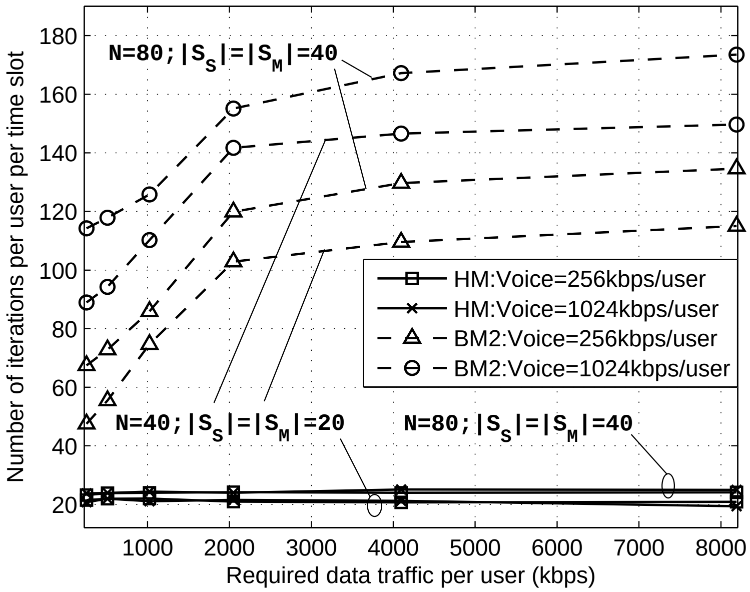


Figure 4.10: Complexities of different algorithms in system-2.

within 25 iterations per user per time slot while BM2 requires more than 47 iterations per user per time slot. The complexity of HM does not considerably vary with the QoS requirements and the number of users as HM recalculates only $\boldsymbol{\mu}$ at each time slot in the second step of the algorithm, whereas the complexity of BM2 increases with the QoS requirements and the number of users as BM2 recalculates $\boldsymbol{\lambda}, \boldsymbol{\xi}$ and $\boldsymbol{\mu}$ at each time slot.

4.8 Summary

This chapter has presented an MMDP based resource (subcarrier, TXOPs of WLAN, and power) allocation for cellular/WLAN interworking considering two time-scales; underlying PHY and MAC layer technologies of an OFDMA based cellular network and a WLAN which operates on contention-based and contention-free channel access mechanisms; and multi-homing capable users with voice and data traffic requirements. Further, by eliminating the requirement to solve Bellman optimality equations, a low time complexity heuristic resource allocation algorithm has been proposed. Simulation results have shown that the MMDP based algorithm provides at least 10.7% of throughput improvements than the heuristic algorithm, and that the heuristic algorithm provides higher throughputs and satisfaction indexes (i.e., QoS) than the benchmark algorithms (BM1 and BM2) which do not consider user multi-homing. The MMDP based algorithm has a high time complexity due to large number of states in the system model. The low time complexity heuristic algorithm converges within 25 iterations per user per time slot in practical systems, which enables it to allocate resources online based on the instantaneous channel gains.

Chapter 5

Resource Allocation for D2D Communication Underlaying Cellular/WLAN Interworking

In this chapter, uplink resource allocation for D2D communication underlaying cellular/WLAN interworking is studied. As stated in Section 1.2, there are several benefits of enabling D2D communication within an interworking system. For example, D2D communication can enhance the network throughputs when interworking cannot, by incorporating hop and reuse gains to the network [22]. In addition to the benefits discussed in Section 1.2, enabling D2D communication in a cellular/WLAN interworking system provides two more important benefits. First, high capacity D2D links can be setup between the users who are not within a WLAN coverage by pairing WLAN radios of the UEs. To pair two WLAN radios, control signals and information related to authentication are sent via cellular network. Though Wi-Fi Direct compatible WLAN radios are able to discover neighbouring devices and pair themselves, they require users to distribute authentication related information, such as a personal identification number, via another secure network and manually enter that information during the link setup phase. Therefore, the D2D link setup process, which takes advantage of UE multi-homing in the interworking system to send control signals and authentication information via the cellular network, provides users with a secure and convenient service. Second, the interference issue in the underlaying networks of D2D communication can be relaxed. In an interworking system, there are a large amount of resources with different channel conditions. Therefore, CCI between traditional and D2D links can be reduced by choosing resources with weaker interference channels between the links.

The remaining of the chapter is organized as follows. In Section 5.1, D2D communication underlaying cellular/WLAN interworking system model is described. Section 5.2 presents the technical challenges for allocating resources in the system, and discusses existing and new solutions. In Section 5.3, a resource allocation scheme is proposed to address these challenges. Section 5.4 discusses the implementation of the proposed scheme, while Section 5.5 demonstrates the achievable throughput and QoS performance enhancements. Chapter is summarized in Section 5.6.

5.1 D2D Communication Underlaying Interworking System Model

As shown in Fig. 5.1, the system model under consideration for this chapter focuses on first and second type areas described in Section 3.1. In order to discuss the implementation of the proposed resource allocation scheme in practical networks, cellular network and WLANs are assumed to be a LTE-A network and IEEE 802.11n WLANs, respectively. Enhanced NodeB (eNB) (or BS) of the LTE-A network and WLAN AP's are interconnected via LTE-A evolve packet core (EPC) network and the Internet service provider (ISP). Synchronization of APs with LTE-A network is achieved by using synchronization protocols, such as IEEE 1588-2008 and Network Time Protocol version 4 (NTPv4), over the Ethernet backhauls connected to the APs. Users can access the services connecting to one network, e.g., UE_8, or simultaneously connecting to multiple networks using the UE multi-homing capability, e.g., UE_9. In this system, network assisted (or operator controlled) D2D communication is considered. Using traditional mode, users communicate with eNB or an AP, e.g., UE_8. Using D2D mode, source and destination users in proximity directly communicate with each other, e.g., UE_4 and UE_5. These D2D links can be established using multiple networks in the interworking system, e.g., D2D link between UE_1 and UE_2 can be established over both LTE-A network and the WLAN. Furthermore, when the D2D users are not within an AP coverage, a high capacity D2D link can be setup between the users by pairing the UE WLAN radios with the assistance of LTE-A network. To pair the WLAN radios, relevant control information and authentication request/response messages are sent through the LTE-A network.

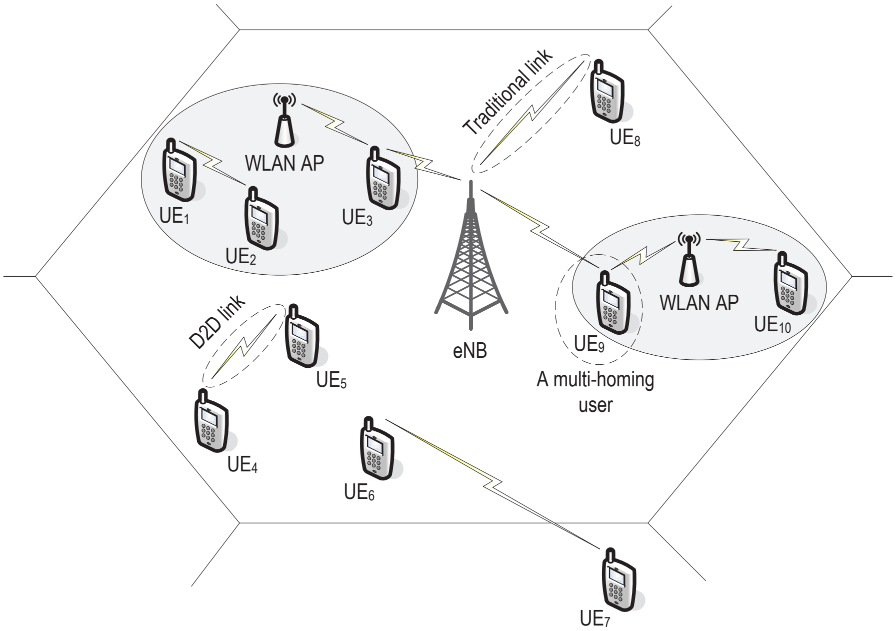


Figure 5.1: D2D communication underlaying cellular/WLAN interworking system.

5.2 Challenges for Resource Allocation and Related Work

In this section, we present three main technical challenges for allocating resources in a D2D communication underlaying cellular/WLAN interworking system: 1) allocation of resources capturing diverse radio access technologies, 2) selection of users' communication modes for multiple networks to maximize hop and reuse gains, and 3) interference management. Some related works in literature are also discussed.

5.2.1 Challenge 1: Allocation of Resources Capturing Multiple Radio Access Technologies

This challenge for allocating resources is discussed in detail in Chapter 4. To allocate resources capturing diverse PHY and MAC technologies in an interworking system, in

Chapter 4 and in [11, 12, 45], resources of the system are allocated by estimating the throughputs via the OFDMA based networks using the Shannon capacity formula and the WLANs by average user throughputs calculated considering the effect of collisions.

To allocate cellular network and WLAN resources in two different time-scales, the following approach is proposed in Chapter 4. First, the wireless channels are modeled as finite-state Markov channels [56], where the state space corresponds to the CSI of all the channels. Then, resource allocation problem at the two time-scales is formulated as a MMDP [18]. The optimal resource allocation decisions are determined by solving the MMDP problem. However, such method is highly complex due to large number of states available in practical systems [11].

5.2.2 Challenge 2: Efficient Mode Selection

The objective of enabling D2D communication in the interworking system is to provide higher data rates with enhanced QoS for the users throughout the interworking system. To achieve this objective, it is crucial to select the best communication modes which take advantage of user proximity and fully realize hop and reuse gains. In this section, we discuss the two key challenges for efficient mode selection: 1) high complexity and communication overhead, and 2) realization of hop and reuse gains.

In an interworking system, mode selection process has a high complexity as it requires estimation of a large number of channels due to availability of a large number of potential D2D and traditional links over multiple networks. It also causes a large communication overhead due to transmission of a large volume of CSI [22]. Therefore, repeating mode selection in a very fast time-scale (e.g., at every resource allocation interval in LTE-A networks with a duration of 1ms [14]) to determine the best communication modes based on the instantaneous channel conditions is not practical. To reduce the complexity and overhead, mode selection can be performed in a slower time-scale based on the channel statistics. However, the time-scale should not be too slow as D2D links may become very weak over time due to user mobility. This issue can be relaxed in the interworking system by forming D2D links for high mobility users via the cellular network only while forming D2D links for low mobility users via the cellular network and WLANs, as the cellular network coverage is much wider than a WLAN coverage.

Realization of hop gain is a challenge as the user modes which provide the highest throughputs should be selected while calculation of D2D mode throughput is complicated. Throughput of D2D mode is the sum of the D2D link throughput and the additional throughput that can be achieved utilizing the saved resources. When D2D mode is used, resources are saved as D2D links use either UL or DL resources only. The additional throughputs will be in the DL if the UL resources are used for the D2D links, and vice versa. In a time-division duplexing (TDD) system, the throughput of D2D mode can be calculated by allocating all the available resources for the D2D link while that of traditional mode can be calculated by allocating a part of available resources for the UL and the remaining resources for the DL [68]. This method provides accurate results as UL and DL share the same set of resources in a TDD system. However, in frequency-division duplexing (FDD) systems, joint allocation of UL and DL resources is necessary as UL and DL use two dedicated sets of resources on two different carrier frequencies.

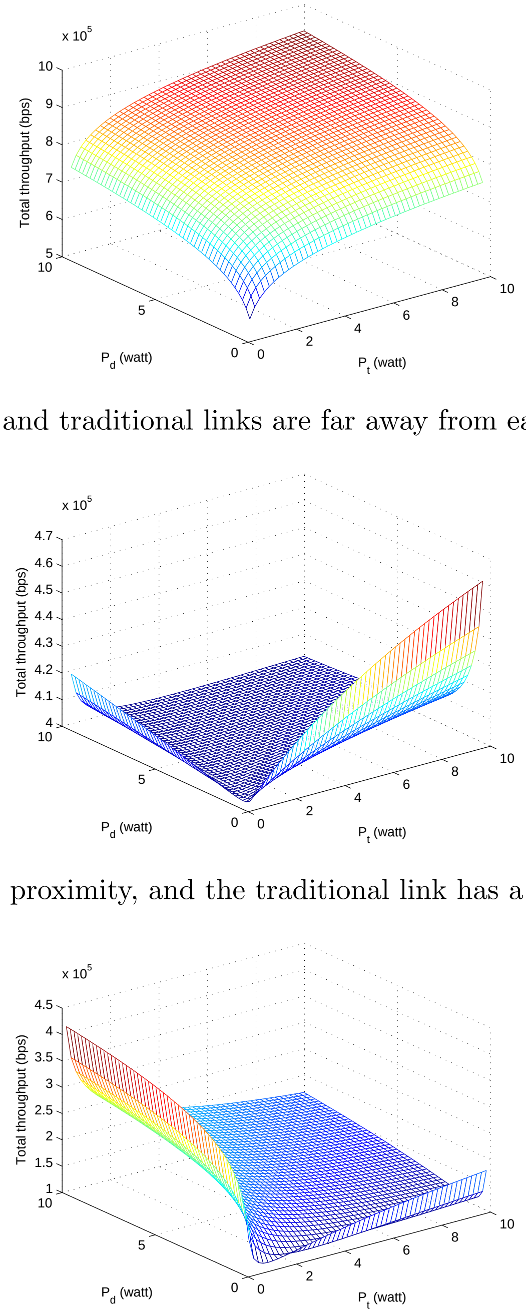
Selecting modes to realize the highest reuse gain in a network is challenging due to two main reasons: 1) calculation of optimal power levels for D2D and traditional links over a shared resource is complicated due to existence of CCI between the links, and 2) finding the optimal pair of D2D and traditional links to share a particular resource is tedious as there are a large number of different link pairs to be considered for each resource. In [68, 69, 70], power allocation to capture reuse gain is studied assuming that the number of available resources equals to the number of traditionally communicating users. Further, it is assumed that each traditionally communicating user occupies only one resource. When one D2D link reuses all the resources, the power allocation which maximizes the D2D link throughput is found in [70]. When there are multiple D2D links and each D2D link reuses only one resource, the optimum power allocation to maximize the total throughput over a resource is found in [68, 69]. Moreover, in [69], by evaluating all the possible D2D and traditional link pairs for each resource, the optimum pairing to maximize the reuse gain in the network is found by using a weighted bipartite matching algorithm. However, in large multicarrier systems (e.g., LTE-A networks), there are a large number of subcarriers or physical resource blocks (RBs) compared to the number of users. Furthermore, the set of resources allocated to one traditional link could be reused by several D2D links, where each D2D link reuses a subset of the resources allocated to the traditional link, and vice versa. In this setting, allocation of RBs in a cellular network based on a reverse iterative combinatorial auction based approach is investigated in [71, 72].

5.2.3 Challenge 3: Interference Management

Intercarrier interference (ICI) and CCI caused by D2D communication degrade the throughput performance of the D2D communication underlaying interworking system. ICI occurs in multicarrier systems, such as LTE-A network, when the signals over different subcarriers arrive at a receiver with different delays. Therefore, to maximize the throughput performance, it is essential to manage the interference. However, interference management is a challenge as it complicates the resource allocation process, by requiring to make three additional resource allocation decisions: 1) whether to allocate orthogonal or non-orthogonal resources for D2D links, 2) whether to utilize UL or DL resources for D2D links, and 3) determine which D2D and traditional link pair to reuse a resource and the transmit power levels of the link pair (discussed under *Challenge 2*). In addition, the characteristics of interworking systems, such as low transmit power levels of the multi-homing users over individual networks, should be considered in the resource allocation process to ease interference management.

To attain high system throughput, selection of orthogonal or non-orthogonal resources for each D2D link should be determined based on the achievable throughputs with each resource type, considering CCI. An orthogonal resource is utilized by only one link whereas a non-orthogonal resource is shared/reused by a D2D and a traditional link. When a D2D link is far away from a traditional link and the two D2D communicating users are in proximity, allocating non-orthogonal resources for these two links is beneficial due to limited CCI between the links [68, 69]. In this scenario, the total achievable throughput via the two links reusing a RB of the LTE-A network is shown in Fig. 5.2a, where P_t and P_d are transmit power levels of traditional and D2D link transmitters. However, as shown in Fig. 5.2b and Fig. 5.2c, when the two links are in a close range, a higher total throughput can be achieved by allocating orthogonal resources for the links; the total throughput reaches the highest when the link with higher channel gain uses the RB.

Use of UL and of DL resources for D2D links affect differently the interference management and the system complexity. When DL resources are reused for D2D links, CCI is received by the traditionally communicating users. To calculate power levels of D2D link transmitters ensuring tolerable CCI at traditionally communicating users, it is required to estimate the channels between D2D link transmitters and traditionally communicating users. Furthermore, CCI could be severe if a D2D pair and a traditionally communicating user are located at a cell edge, or at nearby cell edges [22]. On the other hand, when



(a) D2D and traditional links are far away from each other.

(b) Two links are in proximity, and the traditional link has a higher channel gain.

(c) Two links are in proximity, and the D2D link has a higher channel gain.

Figure 5.2: Throughputs achieved reusing a RB for a D2D and a traditional link when the two links are in different proximities and have different channel gains.

UL resources are reused for D2D links, CCI is received by eNB and APs. Therefore, to manage CCI, already available CSI of the channels between users and eNB/APs can be utilized. In addition to CCI, LTE-A network users will suffer from ICI when DL resources are utilized for D2D links, because the signals from eNB and D2D transmitters arrive at the users at different time instances. However, if UL resources are utilized for D2D links, eNB will suffer from ICI. As ICI is an inherent issue in conventional OFDMA based UL systems, these systems are equipped with ICI cancellation schemes to combat ICI at the eNB, but not at the users. Therefore, use of UL resources for D2D links simplifies CCI and ICI management. Further, it is beneficial to utilize UL resources for D2D links as UL resources are less utilized compared to DL resources due to asymmetric UL and DL traffic loads [70, 69].

Interworking of networks simplifies the interference management in several ways. First, CCI between a D2D and traditional link pair varies with the reusing resource as the channel conditions over different resources vary. Moreover, there are a large amount of resources available from multiple networks. Therefore, CCI in an interworking system can be reduced by selecting resources for D2D and traditional link pairs such that CCI is minimized. Second, when the D2D links are setup over multiple networks, transmit power of the D2D link transmitters is divided among multiple network interfaces, reducing CCI. Third, interworking of a cellular network and WLANs enables the use of WLAN based D2D links. As there are several WLAN frequency channels which can be utilized for these links, multiple D2D links can be setup and simultaneously operated among the users in vicinity without causing CCI. For example, IEEE 802.11n supports three non-overlapping channels in 2.4GHz frequency band [13].

5.3 Proposed Three Time-Scale Resource Allocation Scheme

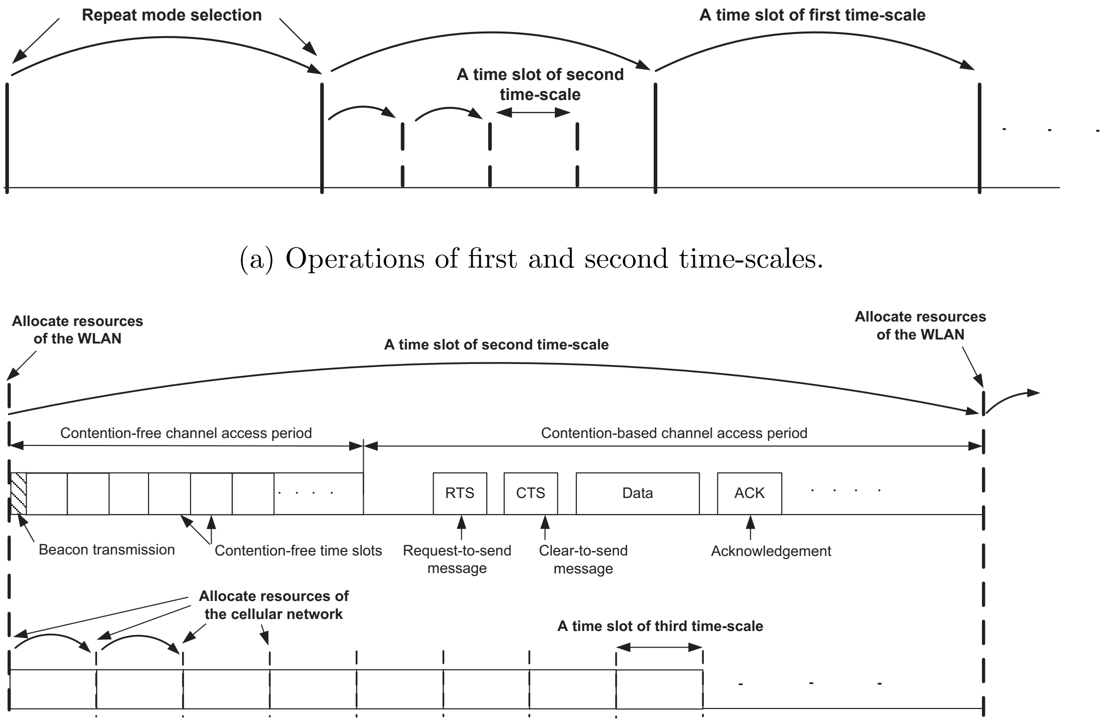
In this section, we propose a resource allocation scheme for D2D communication underlaying cellular/WLAN interworking system shown in Fig. 5.1, overcoming the various challenges mentioned earlier. The proposed scheme is designed with two objectives: 1) maximize the total system throughput subject to user QoS and total power constraints, and 2) minimize the signaling overhead and the computational complexity such that this scheme can be employed in practical systems. The total system throughput is the sum of all the D2D and traditional link throughputs achieved over both networks. Then, the resource allocation problem for this system can be stated as follows.

$$\begin{aligned}
\mathcal{P}6: \ \max \quad & \sum_{i\in\mathcal{S}_N} \left(R_i^{C(D)} + R_i^{C(T)} + R_i^{W(D)} + R_i^{W(T)}\right) \\
\text{s.t.} \quad & \text{C9}: R_i^{C(T)} + R_i^{W(T)} \geq R_{min,i}^{ND}\ ,\ \forall i \in \mathcal{S}_N \\
& \text{C10}: R_i^{C(D)} + R_i^{C(T)} + R_i^{W(D)} + R_i^{W(T)} \geq R_{min,i}^{D2D}\ ,\ \forall i \in \mathcal{S}_N \\
& \text{C11}: P_i^C + P_i^W \leq P_{T,i}\ ,\ \forall i \in \mathcal{S}_N,
\end{aligned}$$

where $R_i^{C(D)}$ and $R_i^{C(T)}$ are the throughputs achieved by the ith user via cellular network using D2D and traditional modes respectively; $R_i^{W(D)}$ and $R_i^{W(T)}$ are the throughputs achieved by the ith user via both channel access mechanisms of the WLAN using D2D and traditional modes respectively; $R_{min,i}^{ND}$ and $R_{min,i}^{D2D}$ are the minimum data rates required for the ith user's non-D2D and D2D communications respectively; and P_i^C and P_i^W are the total power used by the ith user for communications via the cellular network and the two channel access mechanisms of the WLAN respectively. Non-D2D communications are used for accessing the services such as Internet, email and voice mail. C9 and C10 are the QoS constraints, while C11 is the total power constraints. It should be noted that only C9 will be active for the ith user if the user requires only non-D2D communication. On the other hand, only C10 will be active for the user if the user requires only D2D communication. Both C9 and C10 will be active if the user requires both communications. In this work, we assume that each user requires either non-D2D or D2D communication.

An overview of the operations of the proposed resource allocation scheme is shown in Fig. 5.3. The proposed scheme operates on three different time-scales. First time-scale is the slowest while third time-scale is the fastest (i.e., a time slot in first time-scale is the longest while that in third time-scale is the shortest). Mode selection is performed in the first time-scale. Resources of the cellular network and the WLANs are jointly allocated in the second time-scale. As the cellular network has a short resource allocation interval, resources of the cellular network are reallocated in the third time-scale. The proposed scheme addresses the various challenges stated in Section 5.2 as follows.

- To address *Challenge 1*: based on the insights gain from Chapter 4, a low complex joint resource allocation for the cellular network and the WLANs, which operate on the third and the second time-scales respectively, is performed based on the


Repeat mode selection
A time slot of first time-scale
A time slot of second time-scale
Allocate resources of the WLAN
A time slot of second time-scale
Allocate resources of the WLAN
Contention-free channel access period
Contention-based channel access period
RTS
CTS
Data
ACK
Beacon transmission
Contention-free time slots
Request-to-send message
Clear-to-send message
Acknowledgement
Allocate resources of the cellular network
A time slot of third time-scale

(a) Operations of first and second time-scales.

(b) Detailed view of a time slot of second time-scale.

Figure 5.3: Proposed three time-scale resource allocation scheme.

average channel gains; and efficient and feasible resource allocation decisions are made considering the PHY and MAC technologies of the two networks.

- To address *Challenge 2*: complexity and signaling overhead are reduced by performing mode selection in the first (i.e., a slow) time-scale; hop gain is captured in the mode selection by utilizing two resources, which can be allocated to a D2D link or a traditional link with a UL and a DL, to calculate the throughput of each mode; and allocation of non-orthogonal resources is simplified by allocating resources in two steps.
- To address *Challenge 3*: non-interfering WLAN based D2D links are utilized; CCI and ICI mitigation is simplified by using UL resources for the D2D communication within the cellular network; severe CCI is avoided by preventing the allocation of non-orthogonal resources for the links in proximity; and CCI is further reduced

by enabling UE multi-homing for both D2D and traditional modes, and properly calculating the UE transmit power.

5.3.1 First Time-Scale: Mode Selection

Mode selection is performed in the first time-scale in order to reduce the involved computational complexity and the signaling overhead by less frequently (i.e., in the first time-scale) making the mode selection decisions and estimating the channels requiring for mode selection. In the mode selection process, users are allowed to use different modes for different networks as the wireless channel gains over one network differ from those over another network.

An overview of the mode selection algorithm is shown in Fig. 5.4, where CCS is a centralized control server. As shown in *Step 1*, the first step of the mode selection algorithm is to determine communication modes for the users within an AP coverage. In this scenario, users can access both networks using the UE multi-homing capability. In *Step 2*, if a user pair is not within an AP coverage, but within WLAN radio communication range, the pair is allocated a WLAN based D2D link. These links provide a high capacity without causing CCI. Using such a link, a user can utilize all the available contention-free TXOPs and achieve the throughput given by (3.2) over each TXOP. Moreover, in this scenario, $T_{CFP} = T_P$. Due to high capacity of these links, when a user pair is allocated one of these links, the user is not allocated cellular network resources for D2D communication. When the pair of users are not within the WLAN radio communication range, users are allocated cellular network resources only. *Step 2* also determines the communication modes for this scenario.

In cellular network, the mode for each user is selected based on the achievable throughput using each mode, utilizing two RBs. In traditional mode, throughput is calculated allocating one RB for UL and the other for DL; in D2D mode, throughput is calculated allocating both RBs for the D2D link to capture the hop gain. Furthermore, throughputs are calculated substituting average channel gains and unit transmit power levels into (3.1), due to two reasons: 1) instantaneous channel gains and user transmit power levels vary over the time slots of third time-scale, and 2) within each time slot of the first time-scale, there are multiple time slots of the third time-scale.

In WLANs, mode selection is performed in a similar manner, but using two contention-

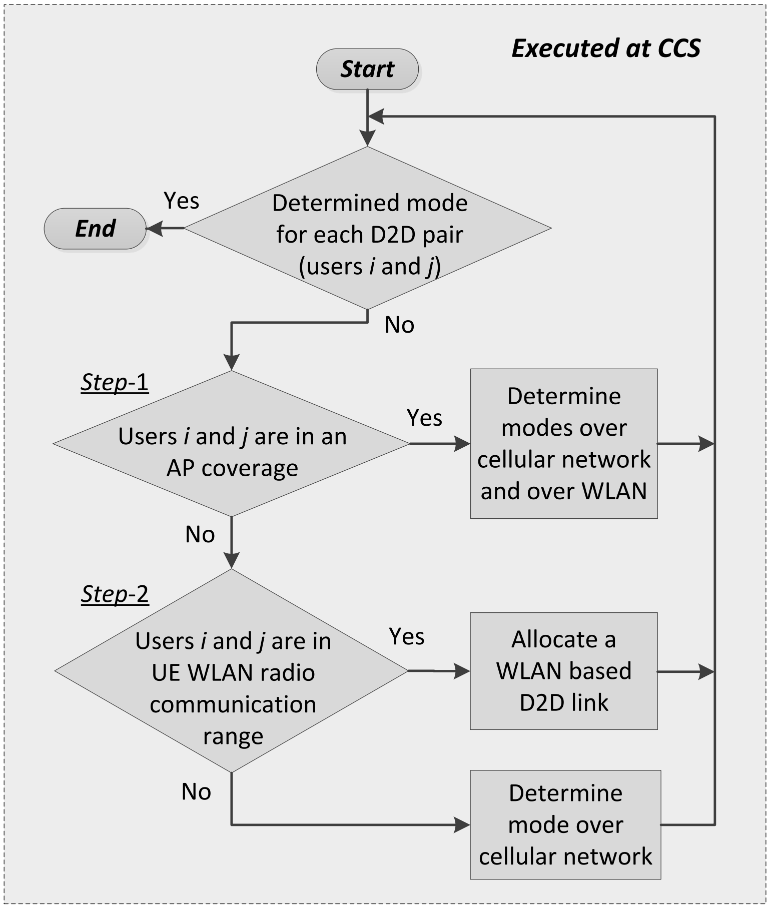

Executed at CCS
Start
Determined mode for each D2D pair (users *i* and *j*)
Yes
End
No
Step-1
Users *i* and *j* are in an AP coverage
Yes
Determine modes over cellular network and over WLAN
No
Step-2
Users *i* and *j* are in UE WLAN radio communication range
Yes
Allocate a WLAN based D2D link
No
Determine mode over cellular network

Figure 5.4: Mode selection algorithm.

free TXOPs. In each mode, throughput over a TXOP is calculated using (3.2). Users use the same mode for contention-based and contention-free channel access mechanisms.

5.3.2 Second Time-Scale: Joint Resource Allocation for Cellular Network and WLANs

Second time-scale resource allocation jointly allocates cellular network and WLAN resources, distributes power available at multi-homing UEs between their two network interfaces, and ensures QoS satisfaction. The resource allocation is executed in two steps in order to simplify the allocation of non-orthogonal resources and the calculation of transmission power levels for D2D and traditional links which share these resources. In the first step, resources are allocated for the traditional links and the D2D links which use orthogonal resources. In the second step, remaining D2D links are allocated non-orthogonal resources. $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the sets of users who are allocated resources during first and second steps respectively, where $\mathcal{S}_1 \cap \mathcal{S}_2 = \mathcal{S}_N$ and $\mathcal{S}_1 \cup \mathcal{S}_2 = \emptyset$.

Resources are allocated using instantaneous channel gains over the WLANs while using average channel gains over the cellular network as there are multiple third time-scale time slots within one time-slot of the second time-scale. Due to the same reason, average channel gains are used for mode selection (see Section 5.3.1). Using average channel gains in the joint resource allocation problem, the amounts of power which should be allocated for UE cellular network interfaces and the throughputs that should be achieved through the cellular network are determined. In the third time-scale, resources of the cellular network are allocated based on these power and throughput levels, and utilizing instantaneous channel gains.

As reusing resources for D2D and traditional links which are in proximity is inefficient, D2D links in WLANs are allocated orthogonal resources. Similarly, in cellular network, D2D links which lie within distance of d_L from the eNB are allocated orthogonal resources. Furthermore, from (3.4), it can be seen that the transmit power levels of all the users in a WLAN are correlated, because an user's throughput via contention-based channel access depends on the transmit power levels of all the users in the WLAN. Also, the users in a WLAN possibly access the cellular network, using a part of the power available in the UEs. Therefore, all the multi-homing users who access both networks should be jointly allocated resources during the first step. To facilitate that, D2D links in the cellular network, which lie at least distance of d_L from the eNB and are among the multi-homing users, are allocated orthogonal resources. Remaining D2D links in the cellular network are allocated non-orthogonal resources realizing the reuse gain in the system. An overview of the second time-scale joint resource allocation algorithm is shown in Fig. 5.5.

First Step of the Second Time-Scale Resource Allocation

First step jointly allocates resources of the cellular network and the WLANs subject to C9-C11, assuming eNB receives the worst CCI of I_c from D2D links. As the communication modes for each user have already been selected by the mode selection algorithm, we denote the throughputs achieved by the ith ($i \in \mathcal{S}_1$) user via cellular network and a WLAN using the selected modes as R_i^C and R_i^W, respectively. For example, if the D2D mode has been selected for the ith user, $R_i^C = R_i^{C(D)}$.

This step allocates resources following *Steps 3 − 6* shown in Fig. 5.5. To allocate resources, Algorithm 3 proposed in Chapter 4 is used with three modifications. First, to further reduce the time complexity of the algorithm, $\mathcal{S}^{CB}$ is assumed to be all the users within the WLAN. Second, instead of using average channel gains within the WLAN, instantaneous channel gains are used. Third, only one QoS constraint per user is considered (either C9 or C10), and it corresponds to the data traffic QoS constraint in Algorithm 3 with the dual variable $\boldsymbol{\lambda}$.

In *Step 3*, dual variables $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are updated using the subgradient method [11, 61, 62], where $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ correspond to QoS and total power constraints of the problem $\mathcal{P}6$, respectively. To update the dual variables, eNB and each AP feedback R_i^C, R_i^W, P_i^C and P_i^W, which are calculated based on the current values of the dual variables. Furthermore, when R_i^C and P_i^C are calculated for links using traditional mode, additional interference of I_c is taken into account. In *Step 4*, first APs determine $\bar{P}_{i,j}^{CF}$, $\rho_{i,j}^{CF}$ and P_i^{CB} for the ith ($i \in \mathcal{S}_1$) user by (4.15)−(4.18), (4.20), and (4.21). Next, it calculates

$$R_i^W = \frac{T_{CF}}{T_P} \sum_{j \in \mathcal{K}^{CF}} \rho_{i,j}^{CF} R_{i,j}^{CF}(\bar{P}_{i,j}^{CF}) \tag{5.1}$$

and

$$P_i^W = P_{avg,i}^{CB}(\mathbf{P}^{CB}) + \frac{T_{CF}}{T_P} \sum_{j \in \mathcal{K}^{CF}} \bar{P}_{i,j}^{CF}, \tag{5.2}$$

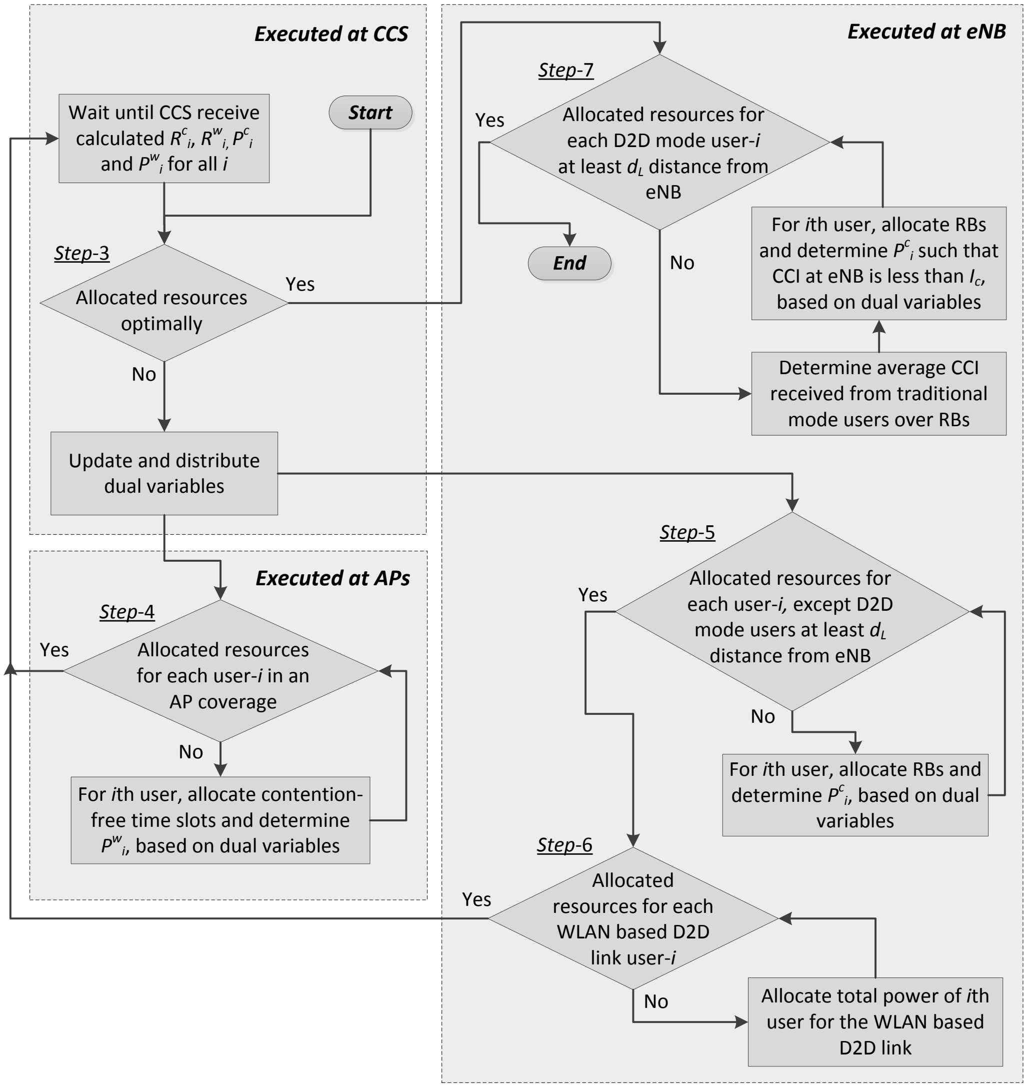

Executed at CCS
Start
Wait until CCS receive calculated R^c_i, R^w_i, P^c_i and P^w_i for all i
Step-3
Allocated resources optimally
Yes
No
Update and distribute dual variables
Executed at APs
Step-4
Allocated resources for each user-i in an AP coverage
Yes
No
For ith user, allocate contention-free time slots and determine P^w_i, based on dual variables
Executed at eNB
Step-7
Allocated resources for each D2D mode user-i at least d_L distance from eNB
Yes
End
No
For ith user, allocate RBs and determine P^c_i such that CCI at eNB is less than I_c, based on dual variables
Determine average CCI received from traditional mode users over RBs
Step-5
Allocated resources for each user-i, except D2D mode users at least d_L distance from eNB
Yes
No
For ith user, allocate RBs and determine P^c_i, based on dual variables
Step-6
Allocated resources for each WLAN based D2D link user-i
Yes
No
Allocate total power of ith user for the WLAN based D2D link

Figure 5.5: Second time-scale joint resource allocation algorithm.

where $R^{CF}_{i,j}(\bar{P}^{CF}_{i,j})$ and $P^{CB}_{avg,i}(\mathbf{P}^{CB})$ are given by (3.2) and (3.6). In *Step 5*, eNB first determines $\rho^C_{i,k}$ and $\bar{P}^C_{i,k}$ using (4.24) and (4.25). Then, it determines

$$R^C_i = \sum_{k\in\mathcal{K}^C} \rho^C_{i,k} R^C_{i,k}(\bar{P}^C_{i,k}) \tag{5.3}$$

and

$$P^C_i = \sum_{k\in\mathcal{K}^C} \bar{P}^C_{i,k}, \tag{5.4}$$

where $R^C_{i,k}(\bar{P}^C_{i,k})$ is given by (3.1). In *Step 6*, resources for the WLAN based D2D links are allocated. Calculation of throughputs of these links is explained in Section 5.3.1. In this scenario, R^W_i is given by (3.2) and $P^W_i = P_{T,i}$.

Second Step of the Second Time-Scale Resource Allocation

In the second step, cellular network resources are allocated (reused) for the D2D links of the users in $\mathcal{S}_2$ (i.e., for D2D links which use non-orthogonal resources) subject to the QoS and total power constraints. CCI received by the D2D links is taken into account, and transmit power levels of the D2D link transmitters are calculated such that they do not exceed I_c at the eNB. CCI received by a D2D link receiver can be calculated as power and RB allocations for the traditional links are already completed in the first step. The second step is performed via *Step 7* of the second time-scale joint resource allocation algorithm shown in Fig. 5.5.

In *Step 7*, cellular network resources can be allocated using the algorithm which is used for allocating resources in *Steps 3* and *5*. In *Step 7*, dual variables are updated only considering the throughputs via and power consumptions for the cellular network. First, from (4.24), eNB determines [73]

$$P^C_{i,k} = \min\left\{\frac{I_c}{(h^C_{i,k})^2}, \left[\frac{\Delta f}{\ln(2)}\frac{(1+\lambda_i)}{\mu_i} - \frac{1}{\alpha^C_{i,k}}\right]^+\right\}, \ \forall i \in \mathcal{S}_2, k \in \mathcal{K}^C, \tag{5.5}$$

where $h^C_{i,k}$ is the average channel gain between the ith user and the eNB, and $\alpha^C_{i,k}$ is calculated taking the received CCI into account. By substituting calculated $P^C_{i,k}$ into (4.25), $\rho^C_{i,k}, \forall i \in \mathcal{S}_2, k \in \mathcal{K}^C$ are calculated. Then, R^C_i and P^C_i can be determined from (5.3) and (5.4), where $\bar{P}^C_{i,k} = \rho^C_{i,k}P^C_{i,k}$.

In order to reduce the required number of channel estimations for the second step, average channel gains which can be estimated based on the distances are used for the calculation of received/caused CCI. To determine the distances, positions of the UEs can be calculated using two techniques which are supported by the LTE networks: 1) assisted global navigation satellite systems (A-GNSS) positioning, and 2) observed time difference of arrival (OTDOA) positioning. Position information can be exchanged between UEs and eNB via LTE positioning protocol (LPP).

5.3.3 Third Time-Scale: Cellular Network Resource Allocation

As cellular networks have a shorter resource allocation interval compared to WLANs, resources of the cellular network are reallocated in the third time-scale utilizing instantaneous channel gains. Further, by using a fast time-scale, multiuser diversity over the fast fading wireless channels is exploited. In this time-scale, resources are allocated following the same two-step process as in the second time-scale. In the first step, the ith multi-homing user has total power of P_i^C to communicate over the cellular network, and requires minimum rate of $R_{min} - R_i^W$ via the cellular network, where $R_{min} = R_{min,i}^{ND}$ if the ith user requires non-D2D communication, or $R_{min} = R_{min,i}^{D2D}$ otherwise. P_i^C and R_i^W are calculated in the second time-scale. Second step remains unchanged.

5.4 Implementation of the Proposed Resource Allocation Scheme

In this section, we discuss the semi-distributed implementation of the proposed resource allocation scheme in an interworking system which consists of an LTE-A network and IEEE 802.11n WLANs operating in 2.1GHz and 2.4GHz frequency bands, respectively. Semi-distributed implementation of the proposed scheme is shown in Fig. 5.6. This implementation reduces the signaling overhead and signaling delay, distributes the computational burden over the networks, and prevent a single point of failure. Different functions of the resource allocation scheme are performed at APs, eNB, and CCS which is connected to the LTE-A EPC through packet data network gateway (PDN-GW). APs and CCS communicate through WLAN access gateway (WAG), evolved packet data

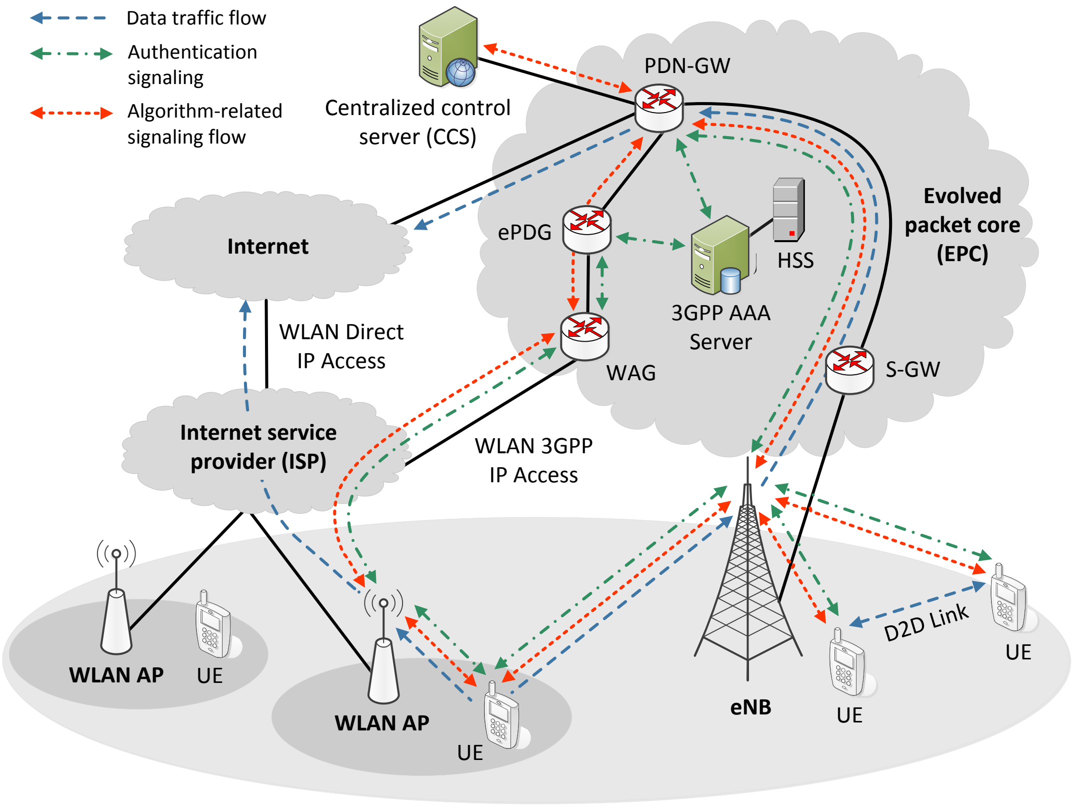


Figure 5.6: Semi-distributed implementation of the proposed resource allocation scheme.

gateway (ePDG) and PDN-GW. eNB and CCS communicate through serving gateway (S-GW) and PDN-GW.

Mode selection is performed at the CCS. To determine the user modes, average channel gains of the traditional and potential D2D links over both networks are sent to the CCS. Once the user modes are determined, the selected modes are informed to APs and eNB to setup the links.

The first step of the second time-scale resource allocation is to jointly allocate cellular network and WLAN resources. Resource allocation for each network is performed at its base station (i.e., eNB or AP), and controlled by the CCS such that resource allocation for the entire system can iteratively converge to the global optimum. Specifically, CCS broadcasts the dual variables which correspond to the total power and the QoS constraints. Then APs and eNB allocate resources based on the received dual variables, and feedback P_i^C, P_i^W, R_i^C and $R_i^W, \forall i \in \mathcal{S}_1$ to CCS. Finally, CCS updates the dual variables and broadcast them back. As shown in *Step 3* of the Fig. 5.5, this process continues until the resource allocation reaches to globally optimal.

The second step of second time-scale resource allocation and the third time-scale resource allocation are performed at the eNB as both of them allocate cellular network resources only. Further, such implementation provides a low signaling delay which is essential for third time-scale operations due to very short time slot duration.

5.5 Simulation Results

Uniformly and randomly distributed 25 high mobility and 25 low mobility users are in the system. All the users are capable of multi-homing, and $Y\%$ of them can communicate using D2D mode. Total power available at each user is 27dBm. Durations of a time slot in the first, second and third time-scales are 640ms, 64ms and 1ms, respectively. We set $d_L = 200$m and $I_c = -62$dBm in the LTE-A network. Remaining simulation parameters are described in Section 4.7.

The proposed scheme reduces the average number of channel estimations and the signaling overhead by 8.3%, 15.9% and 29.1% for $Y = 10\%$, $Y = 20\%$ and $Y = 40\%$ respectively, as the proposed scheme performs mode selection in a slower time-scale. In addition, by executing second and third time-scale resource allocations at APs and eNB, the signaling overhead is reduced by another 58.4% as a large volume of CSI are not sent to the EPC network.

Complexity of the proposed scheme is measured in terms of the number of required iterations per user for the first step of the second time-scale resource allocation, because the first step of the second time-scale resource allocation has the highest complexity as it jointly allocates cellular network and WLAN resources. Required number of iterations for this step is shown in Table 5.1. The required number of iterations per user reduces with Y, as more users are allocated WLAN based D2D links and that more D2D links are allocated non-orthogonal cellular network resources during the second step. Furthermore, it increases with R_{min}, as more iterations are required to find the dual variables which ensure the satisfaction of high QoS requirements.

Table 5.1: Average number of required iterations per user

	R_{min} = 512kbps	R_{min} = 4Mbps	R_{min} = 16Mbps
$Y = 10\%$	5.45	6.51	7.93
$Y = 20\%$	5.23	6.16	7.62
$Y = 40\%$	4.80	5.64	6.91

Throughput and QoS satisfaction performance of the proposed resource allocation scheme is compared with that of a cellular/WLAN interworking system and a conventional system. QoS satisfaction is quantified by using the satisfaction index (SI) which is defined in (4.29). In the cellular/WLAN interworking system, resources are allocated based on the second and third time-scale operations. In the conventional system, resource allocation for each network is performed individually.

According to the throughput and SI performance shown in Fig. 5.7 and Fig. 5.8, the cellular/WLAN interworking system provides higher performance than the conventional system. The proposed scheme provides further performance enhancements, and its performance increases with Y. When Y increases, more D2D links can be established, because the number of potential D2D users in the system increases with Y. As a result, the performance of the proposed scheme increases with Y. When $Y = 40\%$, the proposed scheme improves throughput by 3.4 and 10 times compared to the throughputs achieved in the cellular/WLAN interworking system and the conventional system, respectively. The reasons for such enhanced performance are that joint allocation of resources in multiple networks, exploitation of better wireless channels available between the users in proximity, realization of hop and reuse gains, utilization of WLAN based D2D links, and efficient use of orthogonal and non-orthogonal resources to manage interference. This performance comparison demonstrates the throughput and QoS improvements that can be achieved by interworking of multiple networks and enabling D2D communication within an interworking system.

5.6 Summary

In this chapter, we have studied resource allocation for the D2D communications underlaying interworking system which consists of an LTE-A network and IEEE 802.11n WLANs. A resource allocation scheme has been proposed to maximize the through-

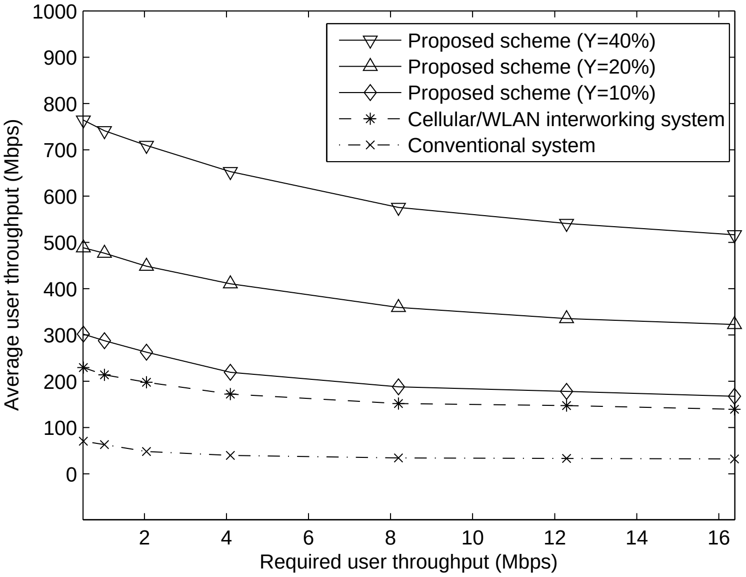


Figure 5.7: Throughput performance.

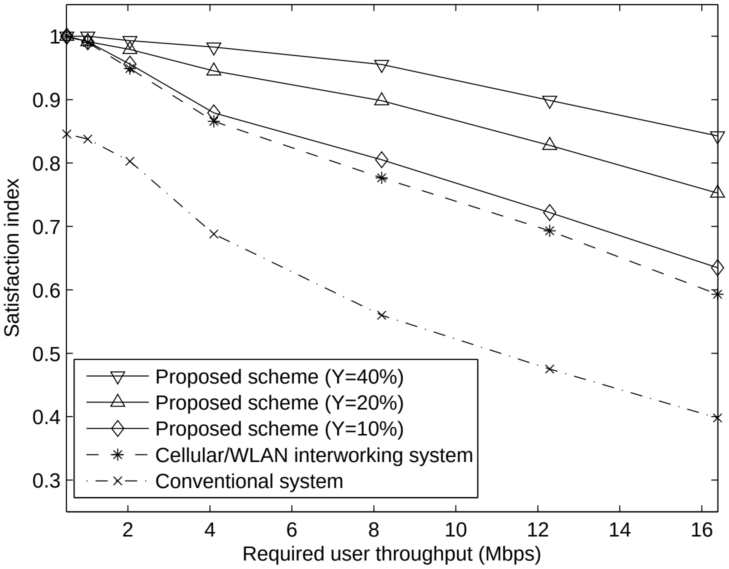


Figure 5.8: QoS satisfaction.

put of the system subject to QoS satisfaction. The proposed scheme has been designed based on the diverse PHY and MAC technologies of different networks, and to manage interference and reduce the high complexity and signaling overhead caused by the mode selection process. To further reduce the signaling overhead and delay while preventing a single point of failure, the proposed scheme has been implemented in a semi-distributed manner. Simulation results have demonstrated that the proposed scheme significantly improves the system throughput and QoS satisfaction.

Chapter 6

Resource Allocation for Interworking Macrocell and Hyper-Dense Small Cell Networks

In this chapter, uplink resource allocation for interworking macrocell and hyper-dense small cell networks is studied. As discussed in Section 1.3, interworking of macrocell and hyper-dense small cell networks provides several important benefits in addition to improving the network throughput, QoS support and coverage. However, there are several challenges for allocating resources for this system as the resources should be jointly allocated for a large number of small cells considering CCI, time-varying network loads, backhauls with limited capacities, and low-cost/low-complex small cell BSs [74].

The main challenge for jointly allocating resources for a large number of small cells is the requirement to solve a highly complex resource allocation problem consisting of a large number of variables and constraints. It should be noted that the number of the cells, which need to be jointly considered when allocating resources, is much higher in a hyper-dense small cell network than that in a macrocell network. Such complex resource allocation problem can be solved using the vastly available cloud computing resources. Further, the use of cloud computing to solve the resource allocation problem provides several advantages as compared to using a dedicated centralized server at the mobile core network [75, 76]: 1) less operational and maintenance cost; 2) ability to adapt to varying computational requirements; and 3) high reliability with spatially distributed multiple redundant servers.

With the use of cloud computing, a major challenge for designing a resource allocation scheme is the high delay when access the cloud computing facility [77]. The delay consists of three components: 1) transmission delay; 2) queuing delay; and 3) processing time. Typically, the total delay is in the order of 100s of milliseconds, and it increases with the size of the computing task. However, the resource allocation decisions should be adapted to rapidly varying wireless channel conditions within a few milliseconds.

In this chapter, we propose a joint resource allocation scheme for allocating uplink resources for interworking macrocell and hyper-dense small cell networks. Cloud computing is used for solving the resource allocation problem. The resource allocation scheme is designed considering the time-varying network loads, limited backhaul capacities and low-complex BSs of the small cell network. The remaining of the chapter is organized as follows. Section 6.1 presents the related work, and Section 6.2 describes the macrocell and hyper-dense small cell interworking system model. In Section 6.3, two time-scale resource allocation for this system is explained. Resource allocation processes at the fast and the slow time-scales are described in Sections 6.4 and 6.5, respectively. Simulation results are given in Section 6.6, while the chapter is summarized in Section 6.7.

6.1 Related Work

The existing CCI management techniques can be mainly divided into three categories: 1) interference avoidance techniques; 2) diversity combining and interference suppression techniques; and 3) interference controlling techniques.

Interference avoidance schemes include resource partitioning schemes, such as ABS allocation and FFR techniques. The FFR techniques eliminate inter cell CCI by allowing the users at a cell edge to utilize only a sub-set of the available frequency channels, while the users at cell centers utilize all the frequency channels [32, 33]. The frequency channel sub-set for cell edge users is determined such that it does not overlap with the frequency channel sub-sets assigned for the neighboring cell edge users. This technique is less efficient in hyper dense small cell networks due to small cell sizes and unplanned cell deployments, which do not allow using all the frequency channels in the cell centers and require partitioning the available frequency channels into a large number of non-overlapping sub-sets. Consequently, frequency reuse is significantly reduced. The

ABS allocation techniques schedule the transmissions within small cells during the ABS transmissions of the macro cell in order to avoid CCI [30, 31]. As ABSs are allocated by the macro cells to avoid CCI between the macro and the small cells, ABS allocation techniques do not eliminate inter cell CCI among the small cells.

Diversity combining and interference suppression techniques for uplink include joint decoding and network MIMO techniques. Joint decoding techniques decode a user's data by combining the signals received by several BSs using techniques such as selection diversity and maximal ratio combining (MRC). It improves SINR due to diversity gain. The network MIMO techniques achieve significantly higher performance compared to diversity combining techniques [34], and the performance of network MIMO techniques is investigated in [35]. Network MIMO techniques which are based on zero forcing and minimum mean squared error (MMSE) equalizers are proposed in [36, 78, 28]. To reduce the amount of control information transmitted to UEs, transmission of precoding matrix indexes is proposed in [79]. There are three key reasons which prevent employing the diversity combining and interference suppression techniques in hyper-dense small cell networks: 1) dense unplanned small cell deployments cause each cell to receive CCI not only from the neighbouring cells, but also from neighbors of the neighboring cells; 2) limited capacity small cell backhauls cannot be used for transmission of instantaneous CSI of the UEs associated with a large number of BSs in vicinity, to facilitate processing of the received signals from the users at different BSs [80, 31]; and 3) high cost of antenna arrays.

Several interference controlling schemes which are based on resource (i.e., transmit power, subcarriers, time slots, etc.) allocation are proposed in [81, 82, 80, 83]. In [81], power and subcarrier allocation to reduce the CCI caused by femto cell users to macrocell users is investigated. In [82], a distributed algorithm is proposed to optimize power and frequency resource allocation subject to inter cell CCI constraints. The work in [83] also considers stability of the UE data queues. To reduce CCI, employment of small cell BSs as relays for uplink macrocell communications is investigated in [80]. In this scheme, the transmit power levels and the amount of data routed via relays are determined by solving a non-cooperative game among UEs using a distributed learning algorithm. To efficiently manage hyper dense small cell networks, further investigation on these schemes to take high delay when access the cloud computing resources is required.

To overcome the effect of high delay when access the cloud computing resources, use of a finite state Markov chain (FSMC) to predict future wireless channel states is proposed in [84]. At each resource allocation interval, cloud resources are used for determining the network resource allocation based on the predicted channel states which are determined using the CSI received a few resource allocation intervals ago. Such solution cannot be directly implemented in hyper dense small cell networks due to three reasons: 1) number of channel states in the FSMC is very large due to existence of a large number of BSs and users in the network (discussed in Section 4.5 and in [85]); 2) transmission of CSI during each resource allocation interval consumes a large portion of the available wireless bandwidth and the backhaul capacity; and 3) determining the resource allocation decisions for such large network requires a duration that is longer than the resource allocation interval of a cellular network (e.g., 1ms). Due to the second and the third reasons, efficient resource allocation mechanisms that allocate resources once a several resource allocation intervals of a cellular network, yet achieve efficient performance, is required hyper-dense small cell networks. In this chapter, we design a resource allocation scheme which uses two time-scales and time-correlated wireless channels to overcome these three issues.

6.2 Macrocell and Hyper-Dense Small Cell Interworking System Model

The system model under consideration for this chapter focuses on first and third type areas described in Section 3.1. Uplink resources are allocated for the interworking system in the areas under consideration. Macro and small cells are deployed in a planned and an unplanned (i.e., random) manner, respectively. The interworking system is divided into C clusters. Resources of a cluster is jointly allocated, while resources of different clusters are allocated separately. As shown in Fig. 6.1, the cth cluster consists of $\mathcal{M}^{(c)}$ set of macro cells and $\mathcal{N}^{(c)}$ set of small cells. $\mathcal{U}^{(c)}$ denotes the set of users in the cth cluster, and each user connects to the cell (i.e., BS) that is closest to the user. The set of users connect to the bth cell is denoted by $\mathcal{U}_b^{(c)}$, where $b \in \mathcal{M}^{(c)} \cup \mathcal{N}^{(c)}$. Total available frequency band is divided into subcarriers with bandwidth of Δf, and $\mathcal{K}$ denotes the set of subcarriers. The set of subcarriers that are used by a cell is dynamically determined based on the cell load and the CCI among the cells. To maximize the aggregated throughput of the users in a cluster, resources of the cluster is allocated in two time-scales; a fast and a slow time-scale. Fast time-scale has a shorter period (i.e., time slot duration) compared to

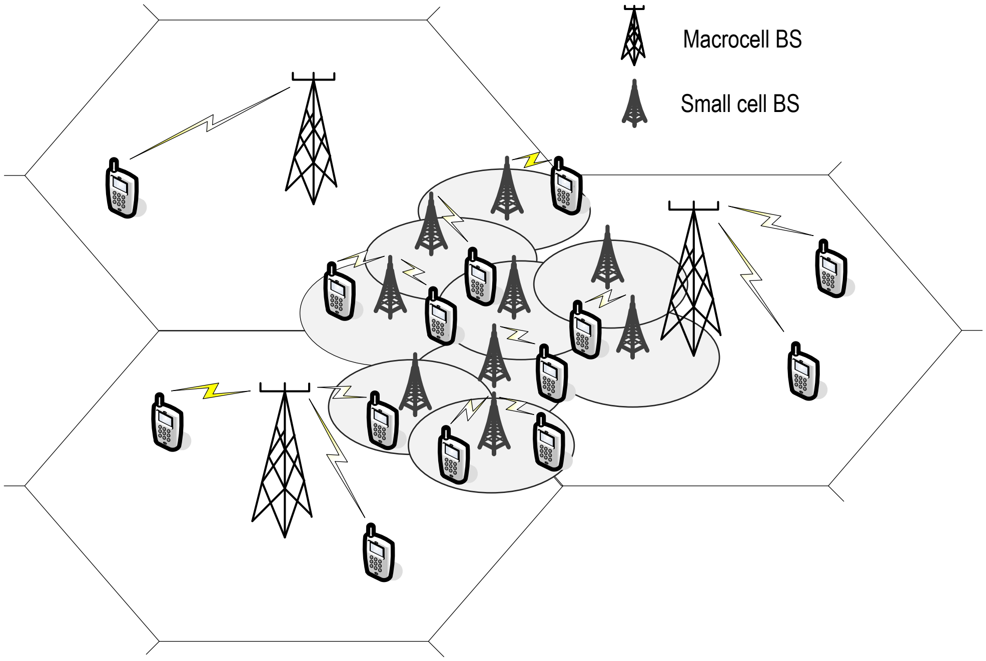


Figure 6.1: A cluster of the interworking macrocell and hyper-dense small cell networks.

that of slow time-scale. In the fast time-scale, user transmit power levels are determined based on the instantaneous CSI and the water-levels which are calculated in the slow time-scale. In the slow time-scale, in addition to the water-levels, user and subcarrier allocations are determined based on the channel statistics and using the cloud computing resources.

6.2.1 Cloud Access Model

To perform the slow time-scale resource allocation, CSI of the users are sent to the cloud at the beginning of each slow time-scale time slot. Next, the resource allocation decisions are determined using the slow time-scale resource allocation algorithm which is stored at the cloud, and the determined decisions are then distributed to the BSs.

The cloud computing resources (or cloud servers) are shared by several users/services to perform different computing tasks. Therefore, the total delay (d_{Total}) incurred when access cloud can be quantified as follows. d_{Total} is the sum of three components: 1) transmission delay (d_t); 2) queuing delay (d_q); and 3) processing time (d_p). d_t for the cth cluster is the maximum round-trip-time between a BS in the cluster and the cloud server, and is given by [86]

$$d_t = \max_{\forall b \in \mathcal{M}^{(c)} \cup \mathcal{N}^{(c)}} \{2 \times 10^{-8} l_b + 5 \times 10^{-3}\}, \tag{6.1}$$

where l_b is the distance between the bth BS ($b \in \mathcal{M}^{(c)} \cup \mathcal{N}^{(c)}$) and cloud servers.

When the computing tasks are sent to the cloud, they are queued until the required amount of computing resources are freed up. The waiting time at the queue is referred to as the queuing delay. When the cloud utilization is low, the queue can be approximated by a M/G/1 queue. Then, by assuming an exponential service (i.e., processing) time distribution [87],

$$d_q = 1/\varphi_s + \frac{(\chi_s^2 + 1/\varphi_s^2)\xi}{2(1-\tau)}, \tag{6.2}$$

where $1/\varphi_s$ and χ_s^2 are the mean and the variance of the service time distribution, respectively; ξ is the average arrival rate; and $\tau = \xi/\varphi_s$.

The processing time, which is the amount of time taken by the cloud servers to solve the resource allocation problem, is given by $\alpha_s + \nu_s$, where α_s is a non-negative constant and ν_s is an exponentially distributed random variable [87].

6.2.2 Channel Model

Wireless channels over subcarriers are modeled as time-correlated Rayleigh fading channels. The channels over different subcarriers fade independently. Complex envelop of a channel at the tth fast time-scale time slot is given by

$$\begin{aligned} \tilde{h}_t &= \rho \tilde{h}_{t-1} + \sqrt{1-\rho^2}\tilde{w}_t \ , \ t \in \{1, ..., L-1\} \\ &= \rho^t \tilde{h}_0 + \sqrt{1-\rho^{2t}}\tilde{w}, \end{aligned} \tag{6.3}$$

where ρ is the correlation coefficient; $\tilde{h}_t$ is the normalized complex channel gain at the tth fast time-scale time slot; $\tilde{w}_t$ and $\tilde{w}$ are complex Gaussian random variables, i.e., $\tilde{w}_t, \tilde{w} \sim \mathcal{CN}(0, \sigma^2)$; and σ^2 is the average power gain of the channel divided by the noise (i.e., additive white Gaussian noise) power. The normalized complex channel gain represents the channel gain divided by the square root of the noise power. The second line of (6.3) is obtained due to the fact that a sum of independent Gaussian random variables is also a Gaussian random variable. Then, $\tilde{h}_t$ can be rewritten as

$$\tilde{h}_t = (\rho^t h_{I,0} + \sqrt{1-\rho^{2t}} w_I) + j(\rho^t h_{Q,0} + \sqrt{1-\rho^{2t}} w_Q), \tag{6.4}$$

where $h_{I,0}$ and w_I are in-phase components of $\tilde{h}_0$ and $\tilde{w}$ respectively, while $h_{Q,0}$ and w_Q are quadrature-phase components of $\tilde{h}_0$ and $\tilde{w}$ respectively. Envelop of the channel gain can then be calculated by

$$|\tilde{h}_t|^2 = X_I^2 + X_Q^2, \tag{6.5}$$

where $X_I = \rho^t h_{I,0} + \sqrt{1-\rho^{2t}} w_I$ and $X_Q = \rho^t h_{Q,0} + \sqrt{1-\rho^{2t}} w_Q$. Furthermore, $X_I \sim \mathcal{N}(\rho^t h_{I,0}, (1-\rho^{2t})\sigma^2/2)$ and $X_Q \sim \mathcal{N}(\rho^t h_{Q,0}, (1-\rho^{2t})\sigma^2/2)$. Since the variances of X_I and X_Q are identical, the conditional probability $Pr(|\tilde{h}_t|^2 \big| |\tilde{h}_0|^2)$ is a noncentral Chi-square distribution with two degrees of freedom [88]. Therefore,

$$Pr(|\tilde{h}_t|^2 = y \big| |\tilde{h}_0|^2) = \frac{1}{2\bar{\sigma}^2} e^{-(s^2+y)/2\bar{\sigma}^2} \sum_{k=0}^{\infty} \frac{y^k (s/2\bar{\sigma}^2)^{2k}}{(k!)^2}, \tag{6.6}$$

where $\bar{\sigma}^2 = (1-\rho^{2t})\sigma^2/2$ and $s = \sqrt{(\rho^t h_{Q,0})^2 + (\rho^t h_{Q,0})^2} = \rho^t |\tilde{h}_0|$. This probability density function is used for calculating average throughput, average power and average CCI over a time slot of slow time-scale.

6.3 Two Time-Scale Resource Allocation

As shown in Fig. 6.2, resources of each cluster are allocated in two time-scales. The durations of fast and slow time-scale time slots are T_F and T_S, respectively. T_F is determined based on the coherence time of the wireless channels, while T_S is determined considering the cluster size, the delay when access the cloud and the control signaling overhead (e.g., CSI). T_0 denotes the difference between T_S and the time duration that is required to transmit CSI to the cloud, determine resource allocation decisions at the cloud and send the decisions back to the BSs. The number of fast time-scale time slots within one time slot of the slow time-scale is denoted by $L(= T_S/T_F)$. For simplicity we assume L is an integer and that $T_0 = L_0 T_F$, where L_0 is also an integer.

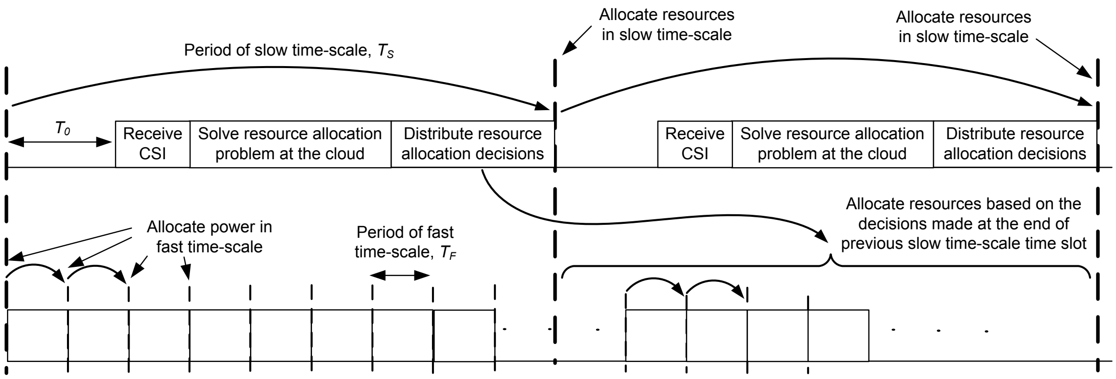

Period of slow time-scale, T_S
Allocate resources in slow time-scale
Allocate resources in slow time-scale
T_0
Receive CSI
Solve resource allocation problem at the cloud
Distribute resource allocation decisions
Receive CSI
Solve resource allocation problem at the cloud
Distribute resource allocation decisions
Allocate power in fast time-scale
Period of fast time-scale, T_F
Allocate resources based on the decisions made at the end of previous slow time-scale time slot

Figure 6.2: Two time-scale resource allocation.

Slow and fast time-scale resource allocation decisions are made at the cloud and the BSs, respectively. The resource allocation decisions which are made at the cloud during the current slow time-scale time slot are used in the cluster during the next slow time-scale time slot. In the slow time-scale, first the users are allocated to the cells. Then, by solving a joint resource allocation problem, the subcarrier allocation and the water-levels for the allocated subcarriers are determined, based on CCI in the cluster, channel statistics, and instantaneous CSI at the L_0th fast time-scale time slot within the current slow time-scale time slot (see Fig. 6.2). Channel statistics and the instantaneous CSI at the L_0th fast time-scale time slot are used for estimating average throughput, average power consumption and average CCI during the next slow time-scale time slot, taking into account the fast time-scale resource allocation. Using these averages, resource allocation decisions are made at the cloud. In the fast time-scale, transmit power levels of the users are determined based on the instantaneous CSI and the already determined water-levels, using the water filling algorithm.

6.4 Fast Time-Scale Resource Allocation

In the fast time-scale, user transmit power levels are calculated based on the instantaneous CSI using the water filling algorithm. Subcarrier allocation and the water-levels are determined in the slow time-scale. Different from the traditional water filling algorithm in which there is a fixed water-level for all the subcarriers, in this work each subcarrier

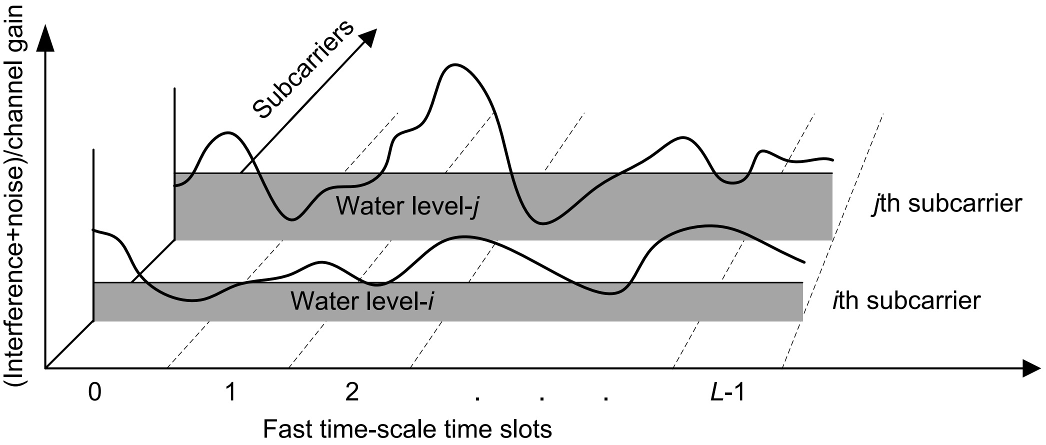


Figure 6.3: Fast time-scale resource allocation within a slot time-scale time slot.

has a different water-level and that water-level is fixed throughout a slow time-scale time slot. Water-levels of different subcarriers are different as the transmit power levels are determined considering the CCI introduced to and received from the other cells. Fig. 6.3 shows an example, where the water levels remain unchanged while the noise plus interference power divided by the channel power gain varies over the fast time-scale time slots. Whenever the water-level is higher than the noise plus interference power divided by the channel power gain, transmit power is allocated for that subcarrier.

From (3.1), throughput of the uth user during the tth fast time-scale time slot over the kth subcarrier is

$$r_{ukt} = \Delta f \log_2 \Big(1 + \frac{H_{ukt} p_{ukt}}{1 + I_{ukt}}\Big), \tag{6.7}$$

where H_{ukt} is the $|\tilde{h}_t|^2$ of the uth user over the kth subcarrier; p_{ukt} is the transmit power level of the uth user during the tth fast time-scale time slot over the kth subcarrier; and I_{ukt} is the normalized interference to the uth user's communications during the tth fast time-scale time slot over the kth subcarrier. The normalized interference is calculated dividing the CCI level by the noise power. For the resource allocation purpose, we assume an user receives the same level of interference throughout a slow time-scale time slot, and which is equivalent to the normalized average interference (I_{uk}). From the water filling algorithm, the optimal transmit power levels are then given by

$$p_{ukt}^{*} = \left[\mu_{uk} - \frac{1 + I_{uk}}{H_{ukt}}\right]^{+}, \tag{6.8}$$

where μ_{uk} is the water-level for the uth user over the kth subcarrier. To determine the transmit power levels, μ_{uk} and I_{uk} are sent to the BSs. Then, the BSs determine the transmit power levels using (6.8), based on μ_{uk}, I_{uk} and instantaneous CSI.

6.5 Slow Time-Scale Resource Allocation

As shown in Fig. 6.2, the resource allocation decisions (i.e., user allocation, subcarrier allocation and water-levels) to be used during the next slow time-scale time slot are determined within the current slow time-scale time slot. To determine these resource allocation decisions, average user throughputs, average power consumptions of the users and average CCI caused/received by the users over the next slow time-scale time slot should be determined. Furthermore, when these averages are determined, the fast time-scale resource allocation (i.e., power allocation) should also be considered, as the variations in the transmit power levels affect average throughput, power consumption and caused/received CCI. Average throughput and power consumption of the uth user during the next slow time-scale time slot over the kth subcarrier are denoted by R_{uk} and P_{uk}, respectively. Derivations of P_{uk}, R_{uk} and I_{uk} based on the channel statistics of time-correlated Rayleigh fading channels and the CSI received at the L_0th fast time-scale time slot within the current slow time-scale time slot (see Fig. 6.2) are given in Appendices B.1, B.2 and B.3, respectively. A time-correlated channel statistically models the relationship between channel gains at different time instances. Thus, using these channels and the CSI received at the L_0th fast time-scale time slot within the current slow time-scale time slot, channel gains during the next slow time-scale time slot can be estimated for calculating P_{uk}, R_{uk} and I_{uk}. Moreover, this approach to calculate P_{uk}, R_{uk} and I_{uk} eliminates the requirement to use a FSMC with a large state space.

Solving the slow time-scale resource allocation problem using the expressions that are derived in Appendices B.1, B.2 and B.3 for P_{uk}, R_{uk} and I_{uk} is highly complex due to two reasons. First, R_{uk} is a non-convex function of the water-levels. Consequently, the resource allocation problem becomes non-convex; thus, exhaustive search methods are required to solve the resource allocation problem. Second, the expressions for P_{uk}, R_{uk} and I_{uk} are highly complex. Therefore, it is not computationally feasible to calculate P_{uk}, R_{uk} and I_{uk} for each user over each subcarrier, to solve the resource allocation problem. Therefore, to gain valuable insights on the effectiveness of the proposed cloud assisted resource allocation framework for interworking macrocell and hyper-dense small cell networks, in this chapter we solve the slow time-scale resource allocation problem with the following simplifications. We calculate P_{uk}, R_{uk} and I_{uk} by assuming that the channels are not time-correlated (i.e., $\rho = 0$) and that the transmit power levels remain unchanged within a slow time-scale time slot. Based on the insights gain by solving the resource allocation problem with these simplifications, in the future works we will further investigate how to design resource allocation schemes considering the time-correlated channels (i.e., $\rho \neq 0$) and the fast time-scale power allocation, in order to further enhance the network throughput performance.

6.5.1 User Allocation

In the first step of the slow time-scale resource allocation, each user is allocated to the BS to which the user has the strongest channel gain. Such user allocation policy is optimal as the proposed resource allocation scheme dynamically adjusts the amount of bandwidth (or subcarriers) available at each BS based on the load at the BS. In contrast to using static bandwidth allocations for the BSs, dynamically adjusting the bandwidths available at the BSs reduces the over-the-air link bottlenecking and efficiently caters for the network load that significantly varies and moves across the network with time.

6.5.2 Subcarrier Allocation and Water-Level Calculation

The subcarrier allocation and the water-levels are determined assuming that the channels are not time-correlated and that the transmit power levels remain unchanged within a slow time-scale time slot. Therefore, from (6.6), (B.10) and (B.11), average throughput of the uth user ($u \in \mathcal{U}^{(c)}$) over the kth subcarrier is given by

$$\begin{aligned} R_{uk} &= \frac{1}{L}\sum_{t=0}^{L-1}\int_0^\infty \Delta f \log_2\Big(1+\frac{H_{ukt}P_{uk}}{1+I_{uk}}\Big)\frac{1}{\sigma_{uk}^2}\exp\Big(\frac{-H_{ukt}}{\sigma_{uk}^2}\Big)\mathrm{d}H_{ukt} \\ &= \frac{1}{L}\sum_{t=0}^{L-1}\frac{\Delta f}{\ln(2)}\exp\Big(\frac{1+I_{uk}}{\sigma_{uk}^2 P_{uk}}\Big)\mathrm{E}_1\Big(\frac{1+I_{uk}}{\sigma_{uk}^2 P_{uk}}\Big), \end{aligned} \tag{6.9}$$

where σ_{uk}^2 is the average of the normalized power gain of the channel over the kth subcarrier between the uth user and the BS to which the uth user is connected to; and $\mathrm{E}_1(\theta)$ is the exponential integral which is given by (4.27). Normalized power gain is the power gain of the channel divided by the noise power.

A tight lower bound for $\mathrm{E}_1(\theta)$ is given by [89]

$$\mathrm{E}_1(\theta) > 0.5\exp(-\theta)\ln(1+2/\theta). \tag{6.10}$$

Thus,

$$\begin{aligned} R_{uk} &\approx \frac{1}{L}\sum_{t=0}^{L-1}\frac{\Delta f}{2\ln(2)}\ln\Big(1+\frac{2\sigma_{uk}^2 P_{uk}}{1+I_{uk}}\Big) \\ &= \frac{\Delta f}{2}\log_2\Big(1+\frac{2\sigma_{uk}^2 P_{uk}}{1+I_{uk}}\Big). \end{aligned} \tag{6.11}$$

Next, from (B.16) and (B.17),

$$\begin{aligned} I_{uk} &= \frac{1}{L}\sum_{t=0}^{L-1}\sum_{v\in\mathcal{U}^{(c)}\backslash u}\int_0^\infty H_{vkt}^{(u)}P_{vk}\frac{1}{(\sigma_{vk}^{(u)})^2}\exp\Big(\frac{-H_{vkt}^{(u)}}{(\sigma_{vk}^{(u)})^2}\Big)\mathrm{d}H_{vkt}^{(u)} \\ &= \sum_{v\in\mathcal{U}^{(c)}\backslash u}(\sigma_{vk}^{(u)})^2 P_{vk}, \end{aligned} \tag{6.12}$$

where $H_{vkt}^{(u)}$ is the normalized power gain of the channel over the kth subcarrier between the vth user and the BS to which the uth user is connected to, during the tth fast time-scale time slot within the next slow time-scale time slot; and $(\sigma_{vk}^{(u)})^2$ is the average of the normalized power gain of the same channel.

The resource allocation problem, which determines subcarrier allocation and water-levels such that the aggregated user throughputs of the users in the cth cluster is maximized, can be stated as follows.

$$\begin{aligned}
\mathcal{P}7 : \max_{\boldsymbol{P}} \quad & \sum_{u\in\mathcal{U}}\sum_{k\in\mathcal{K}} R_{uk} \\
\text{s.t.} \quad & \text{C12 :(6.11)} \\
& \text{C13 :(6.12)} \\
& \text{C14 :} \sum_{k\in\mathcal{K}} P_{uk} \leq P_{T,u} \ , \ \forall u \in \mathcal{U}^{(c)} \\
& \text{C15 :} P_{uk} \geq 0 \ , \ \forall u \in \mathcal{U}^{(c)}, k \in \mathcal{K},
\end{aligned}$$

where $\boldsymbol{P}$ is a vector consisting of all the transmit power variables. When $P_{uk}^{*} > 0$, the k subcarrier is allocated to the uth user. If $P_{vk}^{*} > 0, \forall v \in \mathcal{V}$, where $\mathcal{V} \subseteq \mathcal{U}^{(c)}$, then all the users in $\mathcal{V}$ simultaneously transmit over the k subcarrier. From (6.11) and (6.12), the water-levels can be determined by

$$\mu_{uk} = P_{uk}^{*} + \frac{1+I_{uk}}{\sigma_{uk}^{2}}, \ \forall u \in \mathcal{U}^{(c)}, k \in \mathcal{K}. \tag{6.13}$$

Problem $\mathcal{P}7$ is a non-convex optimization problem due to non-convexity of the objective function. To reduce the required computational capacity to solve $\mathcal{P}7$, we reformulate $\mathcal{P}7$ decomposing the objective function and introducing auxiliary variables as follows.

$$\begin{aligned}
\mathcal{P}8 : \min_{\boldsymbol{P},\boldsymbol{x}} \quad & \sum_{u\in\mathcal{U}^{(c)}}\sum_{k\in\mathcal{K}} \frac{\Delta f}{2}\log_2\Bigg(1+\sum_{v\in\mathcal{U}^{(c)}\setminus u}(\sigma_{vk}^{(u)})^2 P_{vk}\Bigg) - x_{uk} \\
\text{s.t.} \quad & \text{C14, C15} \\
& \text{C16 :} \frac{\Delta f}{2}\log_2\Bigg(1+2\sigma_{uk}^{2}P_{uk}+\sum_{v\in\mathcal{U}^{(c)}\setminus u}(\sigma_{vk}^{(u)})^2 P_{vk}\Bigg) \geq x_{uk} \ , \ \forall u \in \mathcal{U}^{(c)}, k \in \mathcal{K},
\end{aligned}$$

where $\boldsymbol{x}$ is a vector consisting of all x_{uk} variables. $\mathcal{P}8$ is a concave minimization problem, and it can be optimally solved using Algorithm 5 [90], where δ is the error tolerance.

6.5.3 Implementation of Algorithm 5

Implementation of the Algorithm 5 is discussed in this section.

Algorithm 5 : Subcarrier Allocation and Water-Level Calculation

```
Let $\boldsymbol{O}(\boldsymbol{P}, \boldsymbol{x})$ denote the objective function
Let $\boldsymbol{F}$ denotes the feasible set of $\{\boldsymbol{P}, \boldsymbol{x}\}$ which satisfies constraints C14−C16
Find a feasible solution $\{\boldsymbol{P}_f, \boldsymbol{x}_f\}$ which satisfies C14−C16. E.g., $\{0, ..., 0, 0, ..., 0\}$
Find a linear polyhedron $\boldsymbol{F}^0$ which encloses $\boldsymbol{F}$
Find the vertices $(\boldsymbol{V}^0)$ of $\boldsymbol{F}^0$
$s \leftarrow 1$
while Optimal is not found do
    Choose the vertex $\{\boldsymbol{P}_m, \boldsymbol{x}_m\}$ which minimizes $\boldsymbol{O}(\boldsymbol{P}, \boldsymbol{x})$, where $\{\boldsymbol{P}_m, \boldsymbol{x}_m\} \in \boldsymbol{V}^{s-1}$
    Find the smallest $\lambda$ $(0 \leq \lambda \leq 1)$ such that $\{\lambda \boldsymbol{P}_f + (1-\lambda)\boldsymbol{P}_m, \lambda \boldsymbol{x}_f + (1-\lambda)\boldsymbol{x}_m\} \in \boldsymbol{F}$
    $\{\boldsymbol{P}_s, \boldsymbol{x}_s\} \leftarrow \{\lambda \boldsymbol{P}_f + (1-\lambda)\boldsymbol{P}_m, \lambda \boldsymbol{x}_f + (1-\lambda)\boldsymbol{x}_m\}$
    if $\boldsymbol{O}(\boldsymbol{P}_m, \boldsymbol{x}_m) - \boldsymbol{O}(\boldsymbol{P}_s, \boldsymbol{x}_s) \leq \delta$ then
        $\{\boldsymbol{P}_s, \boldsymbol{x}_s\}$ is optimal
    else
        Find an active constraint at $\{\boldsymbol{P}_s, \boldsymbol{x}_s\}$, that is $C^s(\boldsymbol{P}_s, \boldsymbol{x}_s) = 0$, where $C^s(\boldsymbol{P}, \boldsymbol{x}) \in \{\text{C14}, \text{C15}, \text{C16}\}$
        Generate a new constraint $[\nabla C^s(\boldsymbol{P}_s, \boldsymbol{x}_s)](\{\boldsymbol{P}, \boldsymbol{x}\} - \{\boldsymbol{P}_s, \boldsymbol{x}_s\}) \leq 0$
        Add the new constraint to the constraint set $\boldsymbol{F}^{s-1}$, and find the new vertices due to this constraint. Form $\boldsymbol{V}^s$ by adding the new vertices to $\boldsymbol{V}^{s-1}$
        $s \leftarrow s + 1$
    end if
end while
```

Steps 4 and 5:

First, determine a tight upper bound θ for $\sum_{u\in\mathcal{U}^{(c)}}\sum_{k\in\mathcal{K}}(P_{uk}+x_{uk})$ subject to C14$-$C16. Since $x \geq \ln(1+x)$, $\theta = \beta\Delta f/\ln(4) + \sum_{u\in\mathcal{U}^{(c)}} P_{T,u}$, where β can be found by solving the following linear programming problem.

$$\begin{aligned}\mathcal{P}9: \beta = \ \max_{\boldsymbol{P}} \quad & \sum_{u\in\mathcal{U}^{(c)}}\sum_{k\in\mathcal{K}}\Big(2\sigma_{uk}^2 P_{uk} + \sum_{v\in\mathcal{U}^{(c)}\setminus u}(\sigma_{vk}^{(u)})^2 P_{vk}\Big)\\ \text{s.t.} \quad & \text{C14, C15.}\end{aligned}$$

It should be noted that the objective function of $\mathcal{P}9$ is derived from C16, using the relationship of $x \geq \ln(1+x)$. Next, $\mathcal{P}9$ is rewritten by rearranging the terms in its objective function as follows.

$$\begin{aligned}\mathcal{P}10: \beta = \ \max_{\boldsymbol{P}} \quad & \sum_{u\in\mathcal{U}^{(c)}}\sum_{k\in\mathcal{K}} P_{uk}\Big(2\sigma_{uk}^2 + \sum_{v\in\mathcal{U}^{(c)}\setminus u}(\sigma_{uk}^{(v)})^2\Big)\\ \text{s.t.} \quad & \text{C14, C15.}\end{aligned}$$

Then, the optimal value of $P_{uk}, \forall u, k$ for $\mathcal{P}10$ (also for $\mathcal{P}9$) is given by

$$P_{uk} = \begin{cases} P_{T,u}, & \text{if } k = \arg\max\limits_{\forall k\in\mathcal{K}}\big\{2\sigma_{uk}^2 + \sum\limits_{v\in\mathcal{U}^{(c)}\setminus u}(\sigma_{uk}^{(v)})^2\big\};\\ 0, & \text{otherwise.}\end{cases} \tag{6.14}$$

Therefore,

$$\beta = \sum_{u\in\mathcal{U}^{(c)}} P_{T,u}\max_{\forall k\in\mathcal{K}}\big\{2\sigma_{uk}^2 + \sum_{v\in\mathcal{U}^{(c)}\setminus u}(\sigma_{uk}^{(v)})^2\big\}. \tag{6.15}$$

Then, the initial linear polyhedron $\boldsymbol{F}^0$ is given by $\{\boldsymbol{P},\boldsymbol{x}|\sum_{u\in\mathcal{U}(c)}\sum_{k\in\mathcal{K}} P_{uk}+x_{uk} \leq \theta\}$.

The initial set of vertices $\boldsymbol{V}^0$ consists of $2|\mathcal{K}||\mathcal{U}^{(c)}|$ vertices with each vertex being represented by $|\mathcal{K}||\mathcal{U}^{(c)}|$ of P_{uk} and $|\mathcal{K}||\mathcal{U}^{(c)}|$ of x_{uk} variables, where $|\mathcal{Y}|$ denotes the number of elements in $\mathcal{Y}$. The ith vertex's ith variable equals to θ while all the other variables of the vertex equal to zero. It should be noted that the only vertices which produce better objective function values than the initial feasible point are required to be stored and considered in the algorithm.

Step 9:

The smallest value of λ can be found by using a bisection algorithm. In the bisection algorithm, λ is decreased when all the constraints are over satisfied, and λ is increased when there are constraints that are not satisfied.

Steps 14 and 15:

Vertex $\{\boldsymbol{P}_m, \boldsymbol{x}_m\}$ is eliminated from $\boldsymbol{V}^s$ using an additional constraint. First, select an active constraint $C^s(\boldsymbol{P}, \boldsymbol{x})$ at $\{\boldsymbol{P}_s, \boldsymbol{x}_s\}$, i.e., $C^s(\boldsymbol{P}_s, \boldsymbol{x}_s) = 0$. If $C^s(\boldsymbol{P}, \boldsymbol{x}) \in$ C14 and corresponds to the uth user, the new constraint is

$$\sum_{k \in \mathcal{K}} P_{uk} \leq \sum_{k \in \mathcal{K}} P_{uk}^{(s)}, \tag{6.16}$$

where $P_{uk}^{(s)} \in \boldsymbol{P}_s$. Otherwise, if $C^s(\boldsymbol{P}, \boldsymbol{x}) \in$ C16 and corresponds to the uth user and the kth subcarrier, the new constraint is

$$x_{uk} - \frac{\Delta f}{\ln(4) 2^{(2x_{uk}^{(s)}/\Delta f)}} \Big(2\sigma_{uk}^2 P_{uk} + \sum_{v \in \mathcal{U}^{(c)} \setminus u} (\sigma_{vk}^{(u)})^2 P_{vk} \Big) \leq x_{uk}^{(s)} + \frac{\Delta f}{\ln(4)} \Big(1 - \frac{1}{2^{(2x_{uk}^{(s)}/\Delta f)}} \Big), \tag{6.17}$$

where $x_{uk}^{(s)} \in \boldsymbol{x}_s$.

Step 16:

First, form a simplex tableau using the coefficients of the inequalities in $\boldsymbol{F}^{s-1}$ and using the non-zero variables of $\{\boldsymbol{P}_m, \boldsymbol{x}_m\}$ as the basic variables. Second, add the new constraint to the simplex tableau as a new row. Finally, find all the new vertices, which are to be added to $\boldsymbol{V}^s$, by performing dual pivot operations on all the non-basic variables.

6.6 Simulation Results

A cluster of three macrocell BSs and 15 small cell BSs is considered. Small cell BSs are uniformly and randomly distributed within 100m × 100m area at the center of the cluster.

Table 6.1: Simulation Parameters

Parameter	Value (unit)
Single sided power spectra density of noise, N_0	-174dBm/Hz
Path-loss exponent, η	4
Error tolerance in Algorithm 5, δ	0.5%
Transmission delay, d_t	7ms
Queuing delay, d_q	400ms
Processing time, d_p	180ms
l_b for d_t calculation	200km
ξ for d_q calculation	5s^{-1}
$1/\varphi_s$ for d_q calculation	1/10s^{-1}
χ_s^2 for d_q calculation	0.05
α_s for d_p calculation	100ms
ν_s for d_p calculation	80ms

Macrocell BSs are located 300m from the center of the cluster. A similar network setup is shown in Fig. 6.1. Radiuses of a macrocell and a small cell are 1000m and 30m, respectively. Total bandwidth of 20MHz is divided into 128 subcarriers. Each subcarrier can be simultaneously used by multiple BSs. $|\mathcal{U}^{(c)}|$ number of users are uniformly distributed over the 100m × 100m area. Wireless channels between users and BSs are modeled as Rayleigh faded channels with path loss being proportional to $x^{-\eta}$, where η is the path loss exponent and x is the distance between the user and the BS. The total delay when access the cloud (i.e., $d_t + d_q + d_p$) is 594ms [87]. Thus, the minimum duration of T_S is 594ms. Duration of T_F is set to 1ms, similar to that in the LTE standard. Remaining parameters are shown in Table 6.1.

Performance of the proposed resource allocation scheme is compared with that of a benchmark scheme. In the benchmark scheme, first, FFR is used for determining the subcarrier subsets that users can utilize without causing CCI. Second, an optimal resource allocation algorithm is used within each cell to jointly allocate the subcarriers and determine the transmit power levels. Use of FFR is beneficial as it can be implemented in a hyper-dense small cell network without causing a significant signaling overhead in the backhauls or requiring small cell BSs to perform highly complex computational tasks. Reduction of the signaling overhead and the computational burden for the small cell BSs are two of the key objectives of this work. In FFR, users lie within 15m from a small cell BS are able to access all 128 subcarriers while users lie beyond are able to access only a subset of the 128 subcarriers, where the number of subcarriers in a subset equals to 128 divided by the number of overlapping BS coverages. Users who are outside small cell coverages are allocated to the macrocell BSs. Similar to the operation of FFR in small cells, in macrocells the users lie within 150m from a macrocell BS are able to access all the subcarriers while the remaining users are able to access only a subset of the subcarriers. Once FFR determines subcarrier subsets for different areas of each cell, subcarriers and transmit power levels are optimally allocated within each cell. For this purpose, Algorithms 1 and 2 proposed in Chapter 4 are used with the following modifications. Upper and lower level time slot durations are set to be equal, while the WLAN resources and the QoS constraints are ignored. Also, if FFR allows the uth user to access only the subcarriers in the subcarrier subset $\mathcal{K}_u$, then $\rho_{u,k}^{C*} = 0, \forall k \notin \mathcal{K}_u$ in Algorithm 2.

Throughput performance of the proposed and the benchmark schemes is compared in Fig. 6.4. The proposed scheme is able to achieve better throughput performance due to two reasons: 1) it jointly allocates subcarriers and transmit power considering CCI in the entire cluster, and 2) it increases frequency reuse in the network by allowing all the users to access all the subcarriers. In the benchmark, frequency reuse is reduced as the cell-edge users are able to access only a subset of the available subcarriers. Furthermore, when T_S increases from 594ms to 1188ms, the per user average throughput achieved by the proposed scheme reduces, because resource allocation at the cloud is performed only once per slow time-scale time slot and that the effectiveness of the resource allocation decisions decreases with T_S due to changes in the network with time (e.g., changes in channel gains). It should be noted that the water-filling algorithm, which runs at the BSs during each fast time-scale time slot, determines the optimal transmit power level for each subcarrier based on the instantaneous CSI. In addition to that, when $|\mathcal{U}^{(c)}|$ increases, the total network throughput increases due to increase in multiuser diversity. This is reflected by the increase in per user average throughput when T_S=594ms. However, per user average throughput does not increase with $|\mathcal{U}^{(c)}|$ when T_S=1188ms, because the total network throughput is increased by a smaller factor (compared to the increment when T_S=594ms) while it is divided by a larger number (i.e., $|\mathcal{U}^{(c)}|$=60) when the per user average throughput is calculated.

Fig. 6.5 demonstrates complexities of the proposed and the benchmark schemes. Complexity of the proposed scheme is measured in terms of the number of required iterations for the *while* loop in step 7 of Algorithm 5. Complexity of the benchmark is measured in terms of the number of required iterations for the inner most loop. Complex-

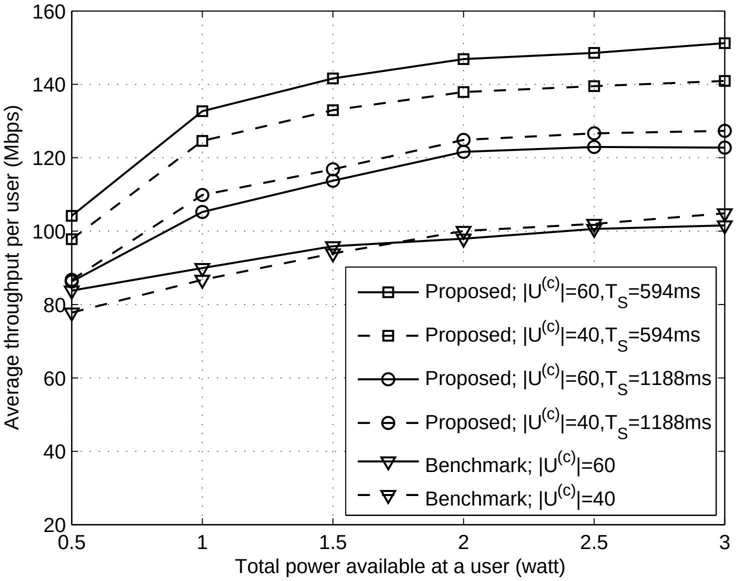


Figure 6.4: Throughput performance of the proposed and the benchmark schemes.

ity of the proposed scheme is higher than that of the benchmark as the proposed scheme jointly allocates subcarriers for the entire cluster, whereas the benchmark jointly allocates subcarriers for a cell. However, Algorithm 5 is executed at the cloud. Thus, in the proposed scheme, small cell BSs are only required to calculate the user transmit power levels using (6.8). In the proposed scheme, the number of required iterations slightly increases with $P_{T,u}$, as the feasible region enlarges with $P_{T,u}$. Consequently, more vertices on the boundary of the feasible region should be determined and checked for optimality. Thus, the number of required iterations increases with $P_{T,u}$. Moreover, it also increases with $|\mathcal{U}^{(c)}|$, due to the increase in the number of variables that need to be determined.

Signaling overhead in the small cell backhauls is reduced due to use of the two time-scale resource allocation approach. In the proposed scheme, small cell BSs are required to send CSI to the cloud only once per slow time-scale time slot. Thus, the signaling overhead is reduced by L times, as compared to the benchmark scheme or other MIMO schemes which exchange instantaneous CSI among BSs (or send to a centralized server) during each fast time-scale time slot. In this simulation setup, $L = T_S/T_F$=594 and 1188. Furthermore, if the network uses TDD, then CSI will also not be exchanged between BSs and users over the fast time-scale. That is due to users are able to estimate CSI as the

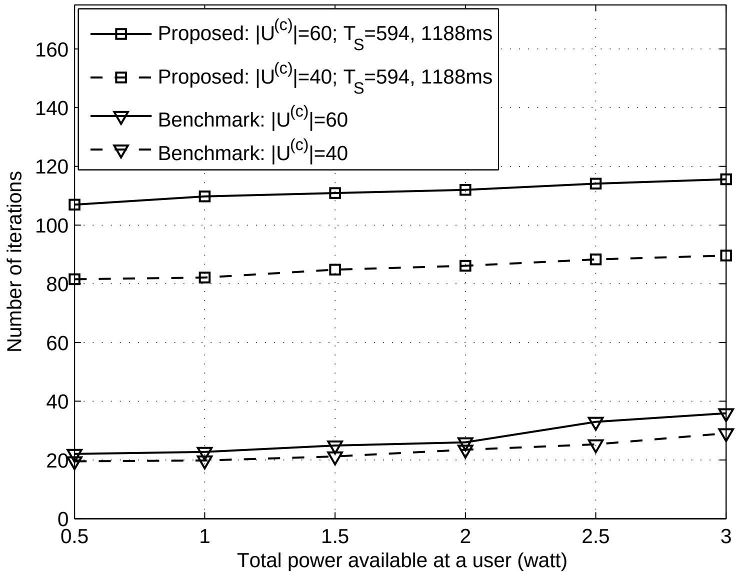


Figure 6.5: Complexities of the proposed and the benchmark schemes.

wireless channels are reciprocal.

6.7 Summary

In this chapter, we have presented a resource allocation scheme for interworking macrocell and hyper-dense small cell networks. The proposed scheme operates on two time-scales and utilizes cloud computing to determine user, subcarrier and transmit power (or water level) allocations for the network taking CCI into account. Resource allocation decisions are made at the cloud in the slow time-scale. In order to adapt these decisions to fast varying wireless channel conditions, they are further optimized at the BSs using water filling algorithm in the fast time-scale. In the proposed scheme, cloud computing is utilized in order to reduce the computational burden for low-cost small cell BSs, while two time-scales are used in order to overcome the performance degradation due to high delay when access the cloud and to reduce the signaling overhead in the backhauls. Simulation results demonstrate that the proposed scheme provides better throughput performance than the benchmark scheme, and its performance increases as the duration of a slow time-scale time slot decreases.

Chapter 7

Conclusions and Future Works

7.1 Conclusions

In this thesis, we have investigated resource allocation for wireless networks which employ interworking, D2D communication and small cell deployment techniques to enhance the network throughput, QoS support and coverage. In Chapter 4, resource allocation for interworking cellular networks and WLANs was studied. The key challenges for allocating resources for this interworking system are the resources have to be allocated capturing multiple PHY and MAC technologies and at two different time-scales. To allocate resources overcoming these challenges, a MMDP based and a low time-complex heuristic schemes which operate on two time-scales have been proposed. Through simulations, we have demonstrated the achievable throughput and QoS improvements using the proposed schemes, compared to that of benchmark schemes which do not allow user multi-homing. Thus, to achieve optimal throughput and QoS performance of an interworking system, resources of the networks in the interworking system should be jointly allocated, transmit power level through each network interface available at UEs should be considered in the same joint resource allocation problem, and user multi-homing should be enabled.

Furthermore, designing of these two resource allocation schemes demonstrates step-by-step approach to consider multiple PHY and MAC technologies in a resource allocation scheme designing process. Also, a novel mechanism to allocate transmit power for WLAN users who access the wireless channels using a DCF based MAC has been presented. This power allocation mechanism is particularly useful for allocating power for multi-homing users in future interworking heterogeneous networks. In addition, we have demonstrated the concept of allocating resources in two different time-scales. Such approach is also useful in the scenarios where an algorithm takes longer time duration to make decisions than the duration within which the decisions are needed. In this scenario, the algorithm can be divided into two sub-algorithms; one sub-algorithm makes long-term decisions in a slow time-scale, and other optimizes those decisions to current conditions of the system in a fast time-scale. Application of this approach is demonstrated in Chapter 6.

In Chapter 5, D2D communication is integrated with the cellular/WLAN interworking system to further improve the network throughput and QoS performance. The additional challenges for allocating resources for this system are the mode selection and interference management. To overcome these challenges, a resource allocation scheme which operates on three time-scales has been proposed. Using a slow time-scale for mode selection, signaling overhead caused by and required channel estimations for mode selection have been reduced. Also, using a two-step resource allocation process, allocation of non-orthogonal resources has been simplified. Simulation results have demonstrated the performance enhancements that can be achieved using the proposed scheme. Therefore, throughput and QoS performance of an interworking system can be further improved by enabling D2D communication within the system.

Another benefit of integrating D2D communication and interworking is that, high capacity non-interfering WLAN based D2D links can be setup among the users in proximity by sending the control and authentication information via the cellular network. To setup these links, WLAN coverage is not required. In addition to these links providing high capacity, they can also be used for reducing the cost of service using unlicensed frequency bands.

In Chapter 6, resource allocation for interworking macrocell and hyper-dense small cell networks was studied. The main challenge for allocating resources for this system is that resources of a large number of cells have to be jointly allocated in order to manage CCI in the system. Hence, the resource allocation schemes for this system become highly complex. Due to high complexity, a resource allocation scheme which operates using cloud computing has been proposed. The effect of high delay, which is caused by transmission delay, queuing delay and processing time at the cloud, on the resource allocation decisions is overcome by using a two time-scale resource allocation approach. Simulation results have shown the throughput improvements that can be achieved using the proposed scheme. Therefore, to achieve high throughput performance, resources of this interworking system (or of any large-scale network) can be jointly and centrally allocated using cloud computing. The negative effect of high delay, when access cloud servers, on the resource allocation decisions can be overcome using a two time-scale resource allocation approach.

7.2 Future Works

To further enhance the efficiency of resource allocation for interworking macrocell and hyper-dense small cell networks, in future works we investigate allocating resources considering QoS and backhaul capacity constraints, efficient clustering mechanisms, and allocating resources considering time-correlated wireless channels.

The main purpose of designing the resource allocation scheme proposed in Chapter 6 was to gain valuable insights on the effectiveness of the proposed cloud assisted resource allocation for interworking macrocell and hyper-dense small cell networks. Thus, for simplicity, QoS and backhaul capacity constraints have not been considered. However, these constraints are essential to guarantee satisfaction of the user QoS requirements and to make sure that the limited capacity small cell backhauls are not bottlenecked. Therefore, in future works, we will further investigate allocating resources considering these constraints.

Furthermore, we will investigate clustering methods for interworking macrocell and hyper-dense small cell networks. In Chapter 6, resources are allocated for a cluster, assuming that the network has already been divided into clusters. The performance of the network increases with the cluster size. However, increasing the cluster size increases the delay caused by cloud computing, due to increased processing time at the cloud. Hence, performance of the network reduces. Therefore, by analyzing the effect of the cluster size on the network performance, we will determine the optimal cluster size for the network. Then, efficient mechanisms will be designed to divide the network into clusters of the optimal size.

In Chapters 4, 5 and 6, we have demonstrated the usefulness of allocating resources in multiple time-scales. Initially, in Chapter 4, we proposed a MMDP based scheme. As this scheme is highly complex due to existence of a large state space, in Chapter 6 we developed a framework in which the slow time-scale resource allocation decisions are made using the average throughputs, power consumptions and interference determined using time-correlated wireless channels and considering fast time-scale power allocation. In future works, we will design resource allocation schemes using this framework to further enhance the efficiency of the interworking macrocell and hyper-dense small cell networks.

Appendices

Appendix A

A.1 Proof of Convexity of $R_i^{CB}(\mathbf{P}^{CB})$

Let

$$f(\mathbf{x}) = \frac{-1}{k_0 + \sum_{i=1}^{N} \frac{k_i}{x_i}} \ , \tag{A.1}$$

where $k_i \in \mathbb{R}^+, \forall i \in \{0, ..., N\}$; $\mathbf{x} = [x_1, ..., x_N]$; and $x_i \in \mathbb{R}^+, \forall i \in \{1, ..., N\}$. Now consider

$$f(\mathbf{z}) - f(\mathbf{x}) - \nabla f(\mathbf{x})[\mathbf{z} - \mathbf{x}]^T = \Big(\frac{1}{k_0 + \sum_{i=1}^{N} \frac{k_i}{z_i}}\Big)\Big(\frac{1}{k_0 + \sum_{i=1}^{N} \frac{k_i}{x_i}}\Big)^2 g(\mathbf{x}) \ , \tag{A.2}$$

where

$$g(\mathbf{x}) = \Big(k_0 + \sum_{i=1}^{N} \frac{k_i}{z_i}\Big)\Big(k_0 + \sum_{i=1}^{N} \frac{k_i z_i}{x_i^2}\Big) - \Big(k_0 + \sum_{i=1}^{N} \frac{k_i}{x_i}\Big)^2 \tag{A.3}$$

and $\mathbf{z} = [z_1, ..., z_N]$ with $z_i \in \mathbb{R}^+, \forall i \in \{1, ..., N\}$. Furthermore, $g(\mathbf{x})$ can be rewritten as

$$\begin{aligned} g(\mathbf{x}) = \ & k_0 \sum_{i=1}^{N} k_i \Big(\frac{1}{\sqrt{z_i}} - \frac{\sqrt{z_i}}{x_i}\Big)^2 + \sum_{i=2}^{N} \sum_{j=1}^{\lfloor i/2 \rfloor} k_{i-j+1} k_j \Bigg(\sqrt{\frac{z_j}{z_{i-j+1} x_j^2}} - \sqrt{\frac{z_{i-j+1}}{z_j x_{i-j+1}^2}}\Bigg)^2 \\ & + \sum_{j=2}^{N-1} \sum_{i=1}^{\lfloor j/2 \rfloor} \Bigg[k_{N-i+1} k_{N-j+i} \times \Bigg(\sqrt{\frac{z_{N-j+i}}{z_{N-i+1} x_{N-j+i}^2}} - \sqrt{\frac{z_{N-i+1}}{z_{N-j+i} x_{N-i+1}^2}}\Bigg)^2\Bigg]. \end{aligned} \tag{A.4}$$

Since $k_i \in \mathbb{R}^+, \forall i \in \{0, ..., N\}$, $g(\mathbf{x}) \geq 0$. Thus, by (A.2), $f(\mathbf{x})$ is a convex function as it satisfies the first order condition [60]. Moreover, $f(\mathbf{x})$ is a non-increasing function as

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \frac{-k_i}{x_i^2}\left(\frac{1}{k_0 + \sum_{i=1}^{N} \frac{k_i}{x_i}}\right)^2 \leq 0, \forall i \in \{1, ..., N\}. \tag{A.5}$$

Let $x_i = \log_2(1 + y_i), \forall i \in \{1, ..., N\}$. Since, $x_i = \log_2(1 + y_i)$ is a concave function with respect to $y_i \in \mathbb{R}^+$, and $f(\mathbf{x})$ is non-increasing in each of its argument and it is a convex function, by vector composition theory [60],

$$f(\mathbf{y}) = \frac{-1}{k_0 + \sum_{i=1}^{N} \frac{k_i}{\log_2(1+y_i)}} \tag{A.6}$$

is a convex function. Thus, $R_i^{CB}(\mathbf{P}^{CB})$ is a concave function.

A.2 Proof of Convexity of C4

From (3.6), the derivative of $P_{avg,i}^{CB}(\mathbf{P}^{CB})$ when $P_i^{CB} > 0$ is

$$\begin{aligned}\frac{\partial P_{avg,i}^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}} = & \frac{T_{CP}}{T_P B^W \cdot \left(\log_2(1 + P_i^{CB}\alpha_i^W)\right)^2} \times \Bigg[R_i^{CB}(\mathbf{P}^{CB})\big(\log_2(1 + P_i^{CB}\alpha_i^W) \\ & - \frac{P_i^{CB}\alpha_i^W}{\ln(2)(1 + P_i^{CB}\alpha_i^W)}\big) + \frac{\partial R_i^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}} P_i^{CB} \log_2(1 + P_i^{CB}\alpha_i^W)\Bigg].\end{aligned} \tag{A.7}$$

Next, we define

$$g(x) = \log_2(1 + x) - \frac{x}{(1 + x)\ln(2)}\ ,\ x \geq 0. \tag{A.8}$$

We can conclude that $g(x) \geq 0$ since $g(0) = 0$ and $\frac{dg(x)}{dx} \geq 0, \forall x \in \mathbb{R}^+$. Moreover, $R_i^{CB}(\mathbf{P}^{CB})$ is a positive non-decreasing concave function. Therefore, by (A.7), $\frac{\partial P_{avg,i}^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}} \geq 0$. Thus, $P_{avg,i}^{CB}(\mathbf{P}^{CB})$ is a non-decreasing function of P_i^{CB}. Therefore, we can show that $P_{avg,i}^{CB}(\mathbf{P}^{CB}) \leq P_{T,i}$ is a convex set [60], and hence C4 is a convex set. That is, $\{P_{avg,i}^{C}, \bar{P}_{i,j}^{CF}, P_i^{CB} | \text{C4 is satisfied}, i \in \mathcal{S}_N, j \in \mathcal{K}^{CF}\}$ is a convex set. Since $P_{avg,i}^{C}$ is a linear combination of $\bar{P}_{i,k}^{C}, \forall k$ (see (4.10)), $\{\bar{P}_{i,k}^{C}, \bar{P}_{i,j}^{CF}, P_i^{CB} | \text{C4 is satisfied}, i \in \mathcal{S}_N, j \in \mathcal{K}^{CF}, k \in \mathcal{K}^{C}\}$ is also a convex set.

A.3 Proof of Existence of a Solution for (4.21)

By differentiating (3.6) and then removing the non-negative term, we have

$$\frac{\partial \bar{R}_i^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}} \begin{cases} = \frac{B^W \alpha_i^W}{\ln(2)} \cdot \frac{\partial P_{avg,i}^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}}, \text{ at } P_i^{CB} = 0 \ ; \\ < \frac{B^W \log_2(1+\alpha_i^W P_i^{CB})}{P_i^{CB}} \cdot \frac{\partial P_{avg,i}^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}}, \text{ for } P_i^{CB} > 0. \end{cases} \tag{A.9}$$

Partial derivatives of $\bar{R}_i^{CB}(\mathbf{P}^{CB})$ and $P_{avg,i}^{CB}(\mathbf{P}^{CB})$ with respect to P_i^{CB} are positive and monotonically decreasing functions of P_i^{CB}, because $\bar{R}_i^{CB}(\mathbf{P}^{CB})$ and $P_{avg,i}^{CB}(\mathbf{P}^{CB})$ are concave increasing functions with respect to P_i^{CB}. In addition to that, the value of $(B^W \log_2(1+\alpha_i^W P_i^{CB}))/P_i^{CB}$ decreases from $B^W \alpha_i^W/\ln(2)$ to 0 as P_i^{CB} goes from 0 to ∞. Therefore, there exist a $p_i, (p_i > 0)$ such that at each $P_i^{CB} > p_i$,

$$\frac{\partial \bar{R}_i^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}} < \frac{\mu_i^*}{1+\lambda_i} \cdot \frac{\partial P_{avg,i}^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}}. \tag{A.10}$$

Thus, from (4.19), (4.20) and (A.10), there exist a P_i^{CB*} which is the solution for (4.21), and $P_i^{CB*} \in (0, p_i)$.

A.4 Proof of Convergence of the Iterative Algorithm which Calculates $\mathbf{P}^{CB*}$

At the nth iteration, P_i^{CB*} is calculated using (4.21) such that when $P_i^{CB} = P_i^{CB*}$, the left hand and the right hand side terms of (A.10) are equal. Since

$$\frac{\left(\frac{\partial^2 \bar{R}_i^{CB}(\mathbf{P}^{CB})}{\partial P_j^{CB} \partial P_i^{CB}}\right)}{\left(\frac{\partial \bar{R}_i^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}}\right)} > \frac{\left(\frac{\partial^2 P_{avg,i}^{CB}(\mathbf{P}^{CB})}{\partial P_j^{CB} \partial P_i^{CB}}\right)}{\left(\frac{\partial P_{avg,i}^{CB}(\mathbf{P}^{CB})}{\partial P_i^{CB}}\right)} \ , \forall P_j^{CB} > 0, i \neq j, \tag{A.11}$$

if P_j^{CB*} is increased at the nth iteration, the left hand side of (A.10) becomes larger than the right hand side of (A.10). Thus, P_i^{CB*} increases at the $(n+1)$th iteration. Therefore, by induction, $P_i^{CB*}, \forall i$ increase in each iteration. However, each P_i^{CB*} is upper bounded by p_i, where $(B^W \log_2(1+\alpha_i^W p_i))/p_i = \mu_i^*/(1+\lambda_i)$. Therefore, the algorithm converges. Furthermore, it should be noted that each P_i^{CB*} converges to a value which is smaller than p_i, because if $P_i^{CB*} = p_i$, it violates (A.9).

Appendix B

B.1 Calculation of Average User Transmit Power

Average transmit power of the uth user during the next slow time-scale time slot over the kth subcarrier is

$$\begin{aligned} P_{uk} &= \mathbb{E}\Big\{\frac{1}{L}\sum_{t=0}^{L-1} p_{ukt}(H_{ukt})\Big|H_{uk(L_0-L)}\Big\} \\ &= \frac{1}{L}\sum_{t=0}^{L-1} \mathbb{E}\big\{p_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\big\}, \end{aligned} \tag{B.1}$$

where $H_{uk(L_0-L)}$ is the CSI transmitted to the cloud at the L_0th fast time-scale time slot within the current slow time-scale time slot. Furthermore, $Pr(H_{ukt}|H_{uk(L_0-L)})$ is given by (6.6) with $\bar{\sigma}^2 = (1-\rho^{2(t+L-L_0)})\sigma_{uk}^2/2$ and $s = \rho^{(t+L-L_0)}\sqrt{H_{uk(L_0-L)}}$, where σ_{uk}^2 is the average power gain of the channel over the kth subcarrier between the uth user and the BS to which the uth user is connected to, divided by the noise power. From (6.6), (6.7) and (6.8),

$$\begin{aligned}\mathbb{E}\big\{p_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\big\} &= \int\limits_{0}^{\infty}\Big[\mu_{uk}-\frac{1+I_{uk}}{y}\Big]^{+}Pr(H_{ukt}=y|H_{uk(L_0-L)})dy \\ &= \int\limits_{(1+I_{uk})/\mu_{uk}}^{\infty}\Big(\mu_{uk}-\frac{1+I_{uk}}{y}\Big)Pr(H_{ukt}=y|H_{uk(L_0-L)})dy \\ &= \int\limits_{(1+I_{uk})/\mu_{uk}}^{\infty}\frac{1}{2\bar{\sigma}^2}\Big(\mu_{uk}-\frac{1+I_{uk}}{y}\Big)e^{-(s^2+y)/2\bar{\sigma}^2}\sum_{k=0}^{\infty}\frac{y^k(s/2\bar{\sigma}^2)^{2k}}{(k!)^2}dy.\end{aligned} \tag{B.2}$$

By substituting $x^2=y/\bar{\sigma}^2$, (B.2) can be written as

$$\mathbb{E}\big\{p_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\big\} = \int\limits_{b}^{\infty}\Big(\mu_{uk}x-\frac{1+I_{uk}}{\bar{\sigma}^2x}\Big)e^{-(x^2+a^2)/2}I_0(ax)dx, \tag{B.3}$$

where $a=s/\bar{\sigma}$, $b=\sqrt{(1+I_{uk})/(\bar{\sigma}^2\mu_{uk})}$, and $I_v(\theta)$ is the vth order modified Bessel function of first kind which is defined as [88]

$$I_v(\theta)=\sum_{k=0}^{\infty}\frac{(\theta/2)^{2k+v}}{k!(k+v)!}\ ,\ \forall v\in\mathbb{Z}^+. \tag{B.4}$$

Furthermore, the 1st order generalized Marcum Q-function is given by [88]

$$\begin{aligned}Q_1(\alpha,\beta) &= \int\limits_{\beta}^{\infty}xe^{-(x^2+\alpha^2)/2}I_0\big(\alpha x\big)dx \\ &= e^{-(\alpha^2+\beta^2)/2}\sum_{k=0}^{\infty}\big(\frac{\alpha}{\beta}\big)^kI_k\big(\alpha\beta\big).\end{aligned} \tag{B.5}$$

By (B.3), (B.4) and (B.5),

$$\mathbb{E}\big\{p_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\big\} = \mu_{uk}Q_1(a,b)-\frac{1+I_{uk}}{\bar{\sigma}^2}\int\limits_{b}^{\infty}\frac{1}{x}e^{-(x^2+a^2)/2}I_0(ax)dx. \tag{B.6}$$

Next, we integrate the 2nd term on the right hand side of (B.6) as follows.

$$\begin{aligned}\int_b^\infty \frac{1}{x} e^{-(x^2+a^2)/2} I_0(ax)dx &= \int_b^\infty \frac{1}{x} e^{-(x^2+a^2)/2}dx + \int_b^\infty \frac{1}{x} e^{-(x^2+a^2)/2} \sum_{k=1}^\infty \frac{(ax/2)^{2k}}{(k!)^2}dx \\ &= 0.5e^{-a^2/2}\mathrm{E}_1(b^2/2) - \int_b^\infty \frac{d\big(e^{-(x^2+a^2)/2}\big)}{dx} \sum_{k=1}^\infty \frac{a^2}{4}\frac{(ax/2)^{2k-2}}{(k!)^2}dx \\ &= 0.5e^{-a^2/2}\mathrm{E}_1(b^2/2) + e^{-(a^2+b^2)/2}\sum_{k=0}^\infty \frac{a^2}{4}\frac{(ab/2)^{2k}}{((k+1)!)^2} \\ &\quad + \int_b^\infty e^{-(x^2+a^2)/2}\sum_{k=2}^\infty \frac{(k-1)a^3}{4}\frac{(ax/2)^{2k-3}}{(k!)^2}dx,\end{aligned} \tag{B.7}$$

where $\mathrm{E}_1(\theta)$ is the exponential integral which is given by (4.27), [89]. It should be noted that there are tight closed form approximations for $\mathrm{E}_1(\theta)$ [85].

By further integrating the remaining integral in (B.7), using integration by parts,

$$\int_b^\infty \frac{1}{x} e^{-(x^2+a^2)/2} I_0(ax)dx = 0.5e^{-a^2/2}\mathrm{E}_1(b^2/2) + e^{-(a^2+b^2)/2}\sum_{i=0}^\infty\sum_{k=0}^\infty \frac{a^{2i+2}}{2^{i+2}}\frac{(k+i)!(ab/2)^{2k}}{k!((k+i+1)!)^2}. \tag{B.8}$$

Therefore, by substituting (B.8) into (B.6),

$$\begin{aligned}&\mathbb{E}\big\{p_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\big\} \\ &= \mu_{uk}Q_1(a,b) - \frac{1+I_{uk}}{\bar{\sigma}^2}\Bigg[0.5e^{-a^2/2}\mathrm{E}_1(b^2/2) + e^{-(a^2+b^2)/2}\sum_{i=0}^\infty\sum_{k=0}^\infty \frac{a^{2i+2}}{2^{i+2}}\frac{(k+i)!(ab/2)^{2k}}{k!((k+i+1)!)^2}\Bigg] \\ &= -\frac{1+I_{uk}}{2\bar{\sigma}^2}e^{-a^2/2}\mathrm{E}_1(b^2/2) \\ &\quad + e^{-(a^2+b^2)/2}\sum_{i=0}^\infty\sum_{k=0}^\infty \frac{a^{2i}}{2^i k!}\Big(\frac{ab}{2}\Big)^{2k}\Bigg[\frac{\mu_{uk}}{(k+i)!} - \frac{(k+i)!(1+I_{uk})a^2}{((k+i+1)!)^2 4\bar{\sigma}^2}\Bigg] \\ &= -\frac{1+I_{uk}}{2\bar{\sigma}^2}e^{-a^2/2}\mathrm{E}_1(b^2/2) + e^{-(a^2+b^2)/2}\sum_{i=0}^\infty\sum_{k=0}^\infty \frac{(a^2/2)^i(ab/2)^{2k}}{k!(k+i)!}\Bigg[\mu_{uk} - \frac{(1+I_{uk})a^2}{(k+i+1)^2 4\bar{\sigma}^2}\Bigg].\end{aligned} \tag{B.9}$$

Next, $\mathbb{E}\{p_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\}$ for $t \in \{0,...,L-1\}$ can be calculated by (B.9), and substituted into (B.1) to calculate the average power consumption over the next time slot of the slow time-scale.

B.2 Calculation of Average User Throughputs

Average throughput of the uth user during the next slow time-scale time slot over the kth subcarrier is

$$\begin{aligned} R_{uk} &= \mathbb{E}\Big\{\frac{1}{L}\sum_{t=0}^{L-1} r_{ukt}(H_{ukt})\Big|H_{uk(L_0-L)}\Big\} \\ &= \frac{1}{L}\sum_{t=0}^{L-1}\mathbb{E}\big\{r_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\big\}, \end{aligned} \tag{B.10}$$

where $H_{uk(L_0-L)}$ is the CSI transmitted to the cloud at the L_0th fast time-scale time slot within the current slow time-scale time slot. Furthermore, $Pr(H_{ukt}|H_{uk(L_0-L)})$ is given by (6.6) with $\bar{\sigma}^2 = (1-\rho^{2(t+L-L_0)})\sigma_{uk}^2/2$ and $s = \rho^{(t+L-L_0)}\sqrt{H_{uk(L_0-L)}}$. From (6.6), (6.7) and (6.8),

$$\begin{aligned} \mathbb{E}\big\{r_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\big\} &= \int\limits_0^\infty \Delta f \log_2\Big(1+\frac{y\big[\mu_{uk}-\frac{1+I_{uk}}{y}\big]^+}{1+I_{uk}}\Big)Pr(H_{ukt}=y|H_{uk(L_0-L)})dy \\ &= \int\limits_{(1+I_{uk})/\mu_{uk}}^\infty \Delta f \log_2\Big(\frac{y\mu_{uk}}{1+I_{uk}}\Big)Pr(H_{ukt}=y|H_{uk(L_0-L)})dy \\ &= \int\limits_{(1+I_{uk})/\mu_{uk}}^\infty \frac{\Delta f}{2\bar{\sigma}^2\ln(2)}\ln\Big(\frac{y\mu_{uk}}{1+I_{uk}}\Big)e^{-(s^2+y)/2\bar{\sigma}^2}\sum_{k=0}^\infty\frac{y^k(s/2\bar{\sigma}^2)^{2k}}{(k!)^2}dy. \end{aligned} \tag{B.11}$$

By substituting $x^2 = y/\bar{\sigma}^2$, (B.11) can be written as

$$\mathbb{E}\big\{r_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\big\} = \frac{2\Delta f}{\ln(2)}\int\limits_b^\infty x\ln\Big(\frac{x}{b}\Big)e^{-(x^2+a^2)/2}I_0(ax)dx, \tag{B.12}$$

where $a = s/\bar{\sigma}$, $b = \sqrt{(1+I_{uk})/(\bar{\sigma}^2\mu_{uk})}$, and $I_0(\theta)$ is given by (B.4).

Also, from (B.5),

$$\int xe^{-(x^2+\alpha^2)/2} I_0\big(\alpha x\big)dx = -Q_1(\alpha, x). \tag{B.13}$$

From (B.12) and (B.13),

$$\begin{aligned}
\mathbb{E}\big\{r_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\big\} &= \frac{2\Delta f}{\ln(2)} \int_b^\infty \ln\left(\frac{x}{b}\right) \frac{d(-Q_1(a,x))}{dx} dx \\
&= \frac{2\Delta f}{\ln(2)} \int_b^\infty \frac{1}{x} Q_1(a,x) dx \\
&= \frac{2\Delta f}{\ln(2)} \int_b^\infty \frac{1}{x} e^{-(x^2+a^2)/2} \sum_{k=0}^{\infty} \left(\frac{a}{x}\right)^k I_k(ax) dx \qquad \text{(B.14)} \\
&= \frac{2\Delta f}{\ln(2)} \sum_{k=0}^{\infty} \frac{a^{2k}}{k!2^k} \int_b^\infty \frac{1}{x} e^{-(x^2+a^2)/2} dx \\
&\quad + \frac{2\Delta f}{\ln(2)} \sum_{k=0}^{\infty} \frac{a^{2k}}{2^k} \int_b^\infty \frac{1}{x} e^{-(x^2+a^2)/2} \sum_{l=1}^{\infty} \frac{(ax/2)^{2l}}{l!(l+k)!} dx.
\end{aligned}$$

By following an integration process which is similar to integration of the 2nd term on the right hand side of (B.6), (B.14) can be simplified as follows.

$$\begin{aligned}
&\mathbb{E}\big\{r_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\big\} \\
&= \frac{\Delta f}{\ln(2)} \Bigg[\sum_{k=0}^{\infty} \frac{a^{2k}}{k!2^k} e^{-a^2/2} \mathrm{E}_1(b^2/2) \\
&\quad + e^{-(a^2+b^2)/2} \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \frac{a^{2(i+k+1)}}{2^{i+k+1}} \frac{(i+l)!(ab/2)^{2l}}{l!(i+l+1)!(i+k+l+1)!} \Bigg] \\
&= \frac{\Delta f}{\ln(2)} \Bigg[\mathrm{E}_1(b^2/2) + e^{-(a^2+b^2)/2} \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \frac{a^{2(i+k+1)}}{2^{i+k+1}} \frac{(i+l)!(ab/2)^{2l}}{l!(i+l+1)!(i+k+l+1)!} \Bigg] \\
&= \frac{\Delta f}{\ln(2)} \Bigg[\mathrm{E}_1(b^2/2) + e^{-(a^2+b^2)/2} \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \frac{(a^2/2)^{(i+k+1)}(ab/2)^{2l}}{l!(i+k+l+1)!(i+l+1)} \Bigg].
\end{aligned} \tag{B.15}$$

By calculating $\mathbb{E}\{r_{ukt}(H_{ukt})\big|H_{uk(L_0-L)}\}$ for $t \in \{0, ..., L-1\}$ using (B.15), the average throughput over the next time slot of the slow time-scale can be calculated using (B.10).

B.3 Calculation of Normalized Average Interference

Normalized average interference to the uth user's communications over the kth subcarrier during the next slow time-scale time slot is

$$\begin{aligned} I_{uk} &= \mathbb{E}\Big\{\frac{1}{L}\sum_{t=0}^{L-1} I_{ukt}\Big\} \\ &= \frac{1}{L}\sum_{t=0}^{L-1}\sum_{v\in\mathcal{U}^{(c)}\setminus u} \mathbb{E}\big\{I_{ukt}^{(v)}\big\}, \end{aligned} \tag{B.16}$$

where $I_{ukt}^{(v)}$ is the normalized interference introduced by the vth user to the uth user over the kth subcarrier during the tth time slot. We assume that the cloud has the knowledge of $H_{vk(L_0-L)}^{(u)}$, which is the normalized power gain of the channel between the vth user and the BS to which the uth user is connected to, during the L_0th fast time-scale time slot within the current slow time-scale time slot. The normalized power gain is calculated by dividing the power gain of the channel by the noise power. Furthermore, $H_{vk(L_0-L)}^{(u)}$ can be transmitted to the cloud as similar to the transmission of $H_{vk(L_0-L)}$, where $H_{vk(L_0-L)}$ is the CSI transmitted by the vth user to the cloud at the L_0th fast time-scale time slot within the current slow time-scale time slot. Then, $\mathbb{E}\{I_{ukt}^{(v)}\}$ can be estimated by considering the time-correlation of the channels as follows.

$$\begin{aligned} &\mathbb{E}\big\{I_{ukt}^{(v)}\big|H_{vk(L_0-L)}^{(u)}, H_{vk(L_0-L)}\big\} \\ &= \int_0^\infty\int_0^\infty x\Big[\mu_{vk} - \frac{1+I_{vk}}{y}\Big]^+ Pr(H_{vkt}^{(u)} = x|H_{vk(L_0-L)}^{(u)})Pr(H_{vkt} = y|H_{vk(L_0-L)})dxdy \\ &= \int_0^\infty\int_{(1+I_{vk})/\mu_{vk}}^\infty x\Big(\mu_{vk} - \frac{1+I_{vk}}{y}\Big)Pr(H_{vkt}^{(u)} = x|H_{vk(L_0-L)}^{(u)})Pr(H_{vkt} = y|H_{vk(L_0-L)})dydx. \end{aligned} \tag{B.17}$$

Since different wireless channels fade independently,

$$\begin{aligned}&\mathbb{E}\big\{I_{ukt}^{(v)}\big|H_{vk(L_0-L)}^{(u)}, H_{vk(L_0-L)}\big\}\\&=\int\limits_{0}^{\infty} xPr(H_{vkt}^{(u)}=x|H_{vk(L_0-L)}^{(u)})dx\int\limits_{(1+I_{vk})/\mu_{vk}}^{\infty}\Big(\mu_{vk}-\frac{1+I_{vk}}{y}\Big)Pr(H_{vkt}=y|H_{vk(L_0-L)})dy.\end{aligned} \tag{B.18}$$

By substituting (6.6) and (B.2) into (B.18),

$$\begin{aligned}\mathbb{E}\big\{I_{ukt}^{(v)}\big|&H_{vk(L_0-L)}^{(u)}, H_{vk(L_0-L)}\big\}\\&=\mathbb{E}\big\{H_{vkt}^{(u)}\big|H_{vk(L_0-L)}^{(u)}\big\}\mathbb{E}\big\{p_{vkt}(H_{vkt})\big|H_{vk(L_0-L)}\big\}\\&=(2\bar{\sigma}^2+s^2)\mathbb{E}\big\{p_{vkt}(H_{vkt})\big|H_{vk(L_0-L)}\big\}\\&=\big[(1-\rho^{2(t+L-L_0)})(\sigma_{vk}^{(u)})^2+\rho^{2(t+L-L_0)}H_{vk(L_0-L)}^{(u)}\big]\mathbb{E}\big\{p_{vkt}(H_{vkt})\big|H_{vk(L_0-L)}\big\},\end{aligned} \tag{B.19}$$

where $(\sigma_{vk}^{(u)})^2$ is the average power gain of the channel over the kth subcarrier between the vth user and the BS to which the uth user is connected to, divided by the noise power. $\mathbb{E}\{p_{vkt}(H_{vkt})\big|H_{vk(L_0-L)}\}$ is given by (B.9). Next, I_{uk} can be calculated by substituting (B.19) into (B.16).

References

[1] “Cisco visual networking index: Global mobile data traffic forecast update, 2012–2017,” [Online], Feb. 2013, available: http://www.cisco.com.

[2] “The 1000x data challenge,” [Online], June 2014, available: http://www.qualcomm.com/solutions/wireless-networks/technologies/1000x-data.

[3] “Millimeter wave propagation: spectrum management implications,” Federal Communications Commission, Tech. Rep. Bulletin Number 70, July 1997.

[4] J. Qiao, X. Shen, J. Mark, Q. Shen, Y. He, and L. Lei, “Enabling device-to-device communications in millimeter-wave 5g cellular networks,” *IEEE Communications Magazine*, vol. 53, no. 1, pp. 209–215, Jan. 2015.

[5] L. Lei, Y. Kuang, X. Shen, C. Lin, and Z. Zhong, “Resource control in network assisted device-to-device communications: solutions and challenges,” *IEEE Communications Magazine*, vol. 52, no. 6, pp. 108–117, June 2014.

[6] Y. Zhou and W. Zhuang, “Throughput analysis of cooperative communication in wireless ad hoc networks with frequency reuse,” *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 205–218, Jan. 2015.

[7] H. T. Cheng and W. Zhuang, “Joint power-frequency-time resource allocation in clustered wireless mesh networks,” *IEEE Network*, vol. 22, no. 1, pp. 45–51, Jan. 2008.

[8] D. Tse and P. Viswanath, *Fundamentals of wireless communication.* Cambridge University Press, 2005.

[9] I. Hwang, B. Song, and S. Soliman, “A holistic view on hyper-dense heterogeneous and small cell networks,” *IEEE Communications Magazine*, vol. 51, no. 6, pp. 20–27, June 2013.

[10] R. Ferrus, O. Sallent, and R. Agusti, “Interworking in heterogeneous wireless networks: Comprehensive framework and future trends,” *IEEE Wireless Communications*, vol. 17, no. 2, pp. 22 –31, Apr. 2010.

[11] M. Ismail and W. Zhuang, “A distributed multi-service resource allocation algorithm in heterogeneous wireless access medium,” *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 2, pp. 425–432, 2012.

[12] X. Pei, T. Jiang, D. Qu, G. Zhu, and J. Liu, “Radio-resource management and access-control mechanism based on a novel economic model in heterogeneous wireless networks,” *IEEE Transactions on Vehicular Technology*, vol. 59, no. 6, pp. 3047–3056, 2010.

[13] “Ieee standard for information technology–telecommunications and information exchange between systems local and metropolitan area networks–specific requirements part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications,” *IEEE P802.11-REVmb/D12*, pp. 1–2910, 2012.

[14] 3GPP, “Lte; evolved universal terrestrial radio access (e-utra) and evolved universal terrestrial radio access network (e-utran); overall description; stage 2,” Tech. Rep. TS 36.300 V11.6.0, 2013.

[15] H. Liang and W. Zhuang, “Cooperative data dissemination via roadside wlans,” *IEEE Communications Magazine*, vol. 50, no. 4, pp. 68–74, Apr. 2012.

[16] P. Liu, Z. Tao, S. Narayanan, T. Korakis, and S. Panwar, “Coopmac: A cooperative mac for wireless lans,” *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 2, pp. 340–354, 2007.

[17] M. Neely, E. Modiano, and C. ping Li, “Fairness and optimal stochastic control for heterogeneous networks,” *IEEE/ACM Transactions on Networking*, vol. 16, no. 2, pp. 396–409, 2008.

[18] H. S. Chang, P. Fard, S. Marcus, and M. Shayman, "Multitime scale markov decision processes," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 976–987, 2003.

[19] M. Ismail, A. Gamage, W. Zhuang, X. Shen, E. Serpedin, and K. Qaraqe, "Uplink decentralized joint bandwidth and power allocation for energy-efficient operation in a heterogeneous wireless medium," *IEEE Transactions on Communications*, vol. 63, no. 4, pp. 1483–1495, Apr. 2015.

[20] K. Doppler, M. Rinne, C. Wijting, C. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to lte-advanced networks," *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42–49, 2009.

[21] C.-H. Yu, K. Doppler, C. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2752–2763, 2011.

[22] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, and Z. Turanyi, "Design aspects of network assisted device-to-device communications," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 170–177, 2012.

[23] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in lte-advanced networks," *IEEE Wireless Communications*, vol. 19, no. 3, pp. 96–104, 2012.

[24] P. Li, S. Guo, T. Miyazaki, and W. Zhuang, "Fine-grained resource allocation for cooperative device-to-device communication in cellular networks," *IEEE Wireless Communications*, vol. 21, no. 5, pp. 35–40, Oct. 2014.

[25] C.-H. Yu, O. Tirkkonen, K. Doppler, and C. Ribeiro, "Power optimization of device-to-device communication underlaying cellular communication," in *IEEE International Conference on Communications (ICC)*, 2009, pp. 1–5.

[26] S. Hakola, T. Chen, J. Lehtomaki, and T. Koskela, "Device-to-device communication in cellular network - performance analysis of optimum and practical communication mode selection," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2010, pp. 1–6.

[27] J. Xu, J. Wang, Y. Zhu, Y. Yang, X. Zheng, S. Wang, L. Liu, K. Horneman, and Y. Teng, “Cooperative distributed optimization for the hyper-dense small cell deployment,” *IEEE Communications Magazine*, vol. 52, no. 5, pp. 61–67, May 2014.

[28] S.-H. Lu, L.-C. Wang, T.-T. Chiang, and C.-H. Chou, “Cooperative hierarchical cellular systems in lte-a networks,” *IEEE Systems Journal*, vol. 9, no. 3, pp. 766–774, Sep. 2015.

[29] Y. Liang, A. Goldsmith, G. Foschini, R. Valenzuela, and D. Chizhik, “Evolution of base stations in cellular networks: Denser deployment versus coordination,” in *IEEE International Conference on Communications (ICC)*, May 2008, pp. 4128–4132.

[30] M. Ding, D. Lopez-Perez, R. Xue, A. Vasilakos, and W. Chen, “Small cell dynamic tdd transmissions in heterogeneous networks,” in *IEEE International Conference on Communications (ICC)*, June 2014, pp. 4881–4887.

[31] S. Deb, P. Monogioudis, J. Miernik, and J. Seymour, “Algorithms for enhanced inter-cell interference coordination (eicic) in lte hetnets,” *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 137–150, Feb. 2014.

[32] A. Stolyar and H. Viswanathan, “Self-organizing dynamic fractional frequency reuse in ofdma systems,” in *IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2008, pp. 13–18.

[33] A. Mahmud and K. Hamdi, “A unified framework for the analysis of fractional frequency reuse techniques,” *IEEE Transactions on Communications*, vol. 62, no. 10, pp. 3692–3705, Oct. 2014.

[34] K. Balachandran, J. Kang, K. Karakayali, and K. Rege, “Network-centric cooperation schemes for uplink interference management in cellular networks,” *Bell Labs Technical Journal*, vol. 18, no. 2, pp. 23–36, Sept. 2013.

[35] H. Huang, M. Trivellato, A. Hottinen, M. Shafi, P. Smith, and R. Valenzuela, “Increasing downlink cellular throughput with limited network mimo coordination,” *IEEE Transactions on Wireless Communications*, vol. 8, no. 6, pp. 2983–2989, June 2009.

[36] G. Foschini, K. Karakayali, and R. Valenzuela, "Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency," *IEE Proceedings-Communications*, vol. 153, no. 4, pp. 548–555, Aug. 2006.

[37] 3GPP, "3gpp system to wireless local area network (wlan) interworking; system description (release 7)," Tech. Rep. TS 23.234 v. 7.7.0, June 2008.

[38] K.-S. Kong, W. Lee, Y.-H. Han, M.-K. Shin, and H. You, "Mobility management for all-ip mobile networks: mobile ipv6 vs. proxy mobile ipv6," *IEEE Wireless Communications*, vol. 15, no. 2, pp. 36–45, Apr. 2008.

[39] 3GPP, "Radio access network; generic access network; stage 2 (release 7)," Tech. Rep. TS 43.318 v. 8.3.0, Aug. 2008.

[40] ETSI, "Improved network controlled mobility between e-utran and 3gpp2/mobile wimax radio technologies," Tech. Rep. 36.938 v.9.0.0, Feb. 2010.

[41] "Ieee standard for local and metropolitan area networks: Media independent handover services," Tech. Rep. IEEE Std 802.21a-2012, 2012.

[42] "Ieee standard for information technology–telecommunications and information exchange between systems local and metropolitan area networks–specific requirements part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications," *IEEE P802.11-REVmb/D12, November 2011 (Revision of IEEE Std 802.11-2007, as amended by IEEEs 802.11k-2008, 802.11r-2008, 802.11y-2008, 802.11w-2009, 802.11n-2009, 802.11p-2010, 802.11z-2010, 802.11v-2011, 802.11u-2011, and 802.11s-2011)*, pp. 1–2910, Mar. 2012.

[43] J. Zhu and A. Fapojuwo, "A new call admission control method for providing desired throughput and delay performance in ieee802.11e wireless lans," *IEEE Transactions on Wireless Communications*, vol. 6, no. 2, pp. 701–709, 2007.

[44] G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.

[45] A. Tharaperiya Gamage and S. Shen, "Uplink resource allocation for interworking of WLAN and OFDMA-Based femtocell systems," in *IEEE International Conference on Communications (ICC)*, Budapest, Hungary, 2013, pp. 4664–4668.

[46] W. Song and W. Zhuang, "Multi-service load sharing for resource management in the cellular/wlan integrated network," *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 725–735, 2009.

[47] J. Xu, Y. Jiang, and A. Perkis, "Multi-service load balancing in a heterogeneous network," in *Wireless Telecommunications Symposium (WTS)*, Apr. 2011, pp. 1 –6.

[48] W. Song, W. Zhuang, and Y. Cheng, "Load balancing for cellular/wlan integrated networks," *IEEE Network*, vol. 21, no. 1, pp. 27–33, 2007.

[49] N. Nasser, A. Hasswa, and H. Hassanein, "Handoffs in fourth generation heterogeneous networks," *IEEE Communications Magazine*, vol. 44, no. 10, pp. 96–103, Oct. 2006.

[50] A.-E. Taha, H. Hassanein, and H. Mouftah, "On robust allocation policies in wireless heterogeneous networks," in *First International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks*, Oct. 2004, pp. 198–205.

[51] A. Hasswa, N. Nasser, and H. Hassanein, "Generic vertical handoff decision function for heterogeneous wireless," in *Second IFIP International Conference on Wireless and Optical Communications Networks*, Mar. 2005, pp. 239–243.

[52] C. Luo, H. Ji, and Y. Li, "Utility-based multi-service bandwidth allocation in the 4g heterogeneous wireless access networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2009, pp. 1–5.

[53] D. Niyato and E. Hossain, "A noncooperative game-theoretic framework for radio resource management in 4g heterogeneous wireless access networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 3, pp. 332–345, 2008.

[54] P. Xue, P. Gong, J. H. Park, D. Park, and D. K. Kim, "Radio resource management with proportional rate constraint in the heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1066–1075, 2012.

[55] P. Wang, H. Jiang, and W. Zhuang, "Capacity improvement and analysis for voice/data traffic over wlans," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1530–1541, 2007.

[56] H.-S. Wang and N. Moayeri, “Finite-state markov channel-a useful model for radio communication channels,” *IEEE Transactions on Vehicular Technology*, vol. 44, no. 1, pp. 163–171, 1995.

[57] E. Altman, *Constrained Markov Decision Processes.* Chapman & Hall/CRC, 1999.

[58] M. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*, 1994, vol. New York: J. Wiley Sons.

[59] C. Y. Wong, R. Cheng, K. Lataief, and R. Murch, “Multiuser ofdm with adaptive subcarrier, bit, and power allocation,” *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 10, pp. 1747–1758, 1999.

[60] S. P. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge University Press, 2004.

[61] D. P. Bertsekas, *Non-linear programming.* Athena Scientific, 2003.

[62] M. S. Alam, J. W. Mark, and X. S. Shen, “Relay selection and resource allocation for multi-user cooperative ofdma networks,” *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2193–2205, May 2013.

[63] K. C. T., *Solving Nonlinear Equations with Newton’s Method.* Society for Industrial and Applied Mathematics, 2003.

[64] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Function with Formulas, Graphs, and Mathematical Tables*, ser. Applied Mathematics Series 55. NBS, 1964.

[65] “WINNER II Interim Channel Models,” EC FP6, Tech. Rep. D1.1.1 v1.0, Dec. 2006. [Online]. Available: http://www.ist-winner.org/deliverables.html

[66] T. Rappaport, *Wireless Communications*, 2nd ed. Prentice Hall, 2002.

[67] H. Shen, L. Cai, and X. Shen, “Performance analysis of tfrc over wireless link with truncated link-level arq,” *IEEE Transactions on Wireless Communications*, vol. 5, no. 6, pp. 1479–1487, 2006.

[68] C.-H. Yu, K. Doppler, C. Ribeiro, and O. Tirkkonen, “Resource sharing optimization for device-to-device communication underlaying cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2752–2763, Aug. 2011.

[69] D. Feng, L. Lu, Y. Yuan-Wu, G. Li, G. Feng, and S. Li, “Device-to-device communications underlaying cellular networks,” *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3541–3551, Aug. 2013.

[70] J. Wang, D. Zhu, C. Zhao, J. Li, and M. Lei, “Resource sharing of underlaying device-to-device and uplink cellular communications,” *IEEE Communications Letters*, vol. 17, no. 6, pp. 1148–1151, June 2013.

[71] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, X. Cheng, and B. Jiao, “Efficiency resource allocation for device-to-device underlay communication systems: A reverse iterative combinatorial auction based approach,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 348–358, Sep. 2013.

[72] C. Xu, L. Song, and Z. Han, *Resource management for device-to-device underlay communication.* Springer, 2014.

[73] A. Gamage, N. Rajatheva, and M. Codreanu, “Resource allocation for ofdma-based relay assisted two-tier femtocell networks,” in *International Symposium on Wireless Communication Systems (ISWCS)*, 2011, pp. 834–838.

[74] C. Ran, S. Wang, and C. Wang, “Balancing backhaul load in heterogeneous cloud radio access networks,” *IEEE Wireless Communications*, vol. 22, no. 3, pp. 42–48, June 2015.

[75] N. Zhang, N. Cheng, A. Gamage, K. Zhang, J. Mark, and X. Shen, “Cloud assisted hetnets toward 5g wireless networks,” *IEEE Communications Magazine*, vol. 53, no. 6, pp. 59–65, June 2015.

[76] M. Marotta, N. Kaminski, I. Gomez-Miguelez, L. Zambenedetti Granville, J. Rochol, L. Dasilva, and C. Both, “Resource sharing in heterogeneous cloud radio access networks,” *IEEE Wireless Communications*, vol. 22, no. 3, pp. 74–82, June 2015.

[77] Q. Shen, X. Liang, X. Shen, X. Lin, and H. Luo, “Exploiting geo-distributed clouds for a e-health monitoring system with minimum service delay and privacy preser-

vation," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 430–439, Mar. 2014.

[78] B. Guler and A. Yener, "Uplink interference management for coexisting mimo femtocell and macrocell networks: An interference alignment approach," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2246–2257, Apr. 2014.

[79] H.-H. Lee, K.-H. Park, Y.-C. Ko, and M.-S. Alouini, "Codebook-based interference alignment for uplink mimo interference channels," *Journal of Communications and Networks*, vol. 16, no. 1, pp. 18–25, Feb. 2014.

[80] S. Samarakoon, M. Bennis, W. Saad, and M. Latva-Aho, "Backhaul-aware interference management in the uplink of wireless small cell networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 11, pp. 5813–5825, Nov. 2013.

[81] N. Sharma, D. Badheka, and A. Anpalagan, "Multiobjective subchannel and power allocation in interference-limited two-tier ofdma femtocell networks," *IEEE Systems Journal*, vol. PP, no. 99, pp. 1–12, 2014.

[82] E. Biton, A. Cohen, G. Reina, and O. Gurewitz, "Distributed inter-cell interference mitigation via joint scheduling and power control under noise rise constraints," *IEEE Transactions on Wireless Communications*, vol. 13, no. 6, pp. 3464–3477, June 2014.

[83] X. Xiang, C. Lin, X. Chen, and X. Shen, "Toward optimal admission control and resource allocation for lte-a femtocell uplink," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 7, pp. 3247–3261, July 2015.

[84] Y. Cai, F. Yu, and S. Bu, "Cloud radio access networks (c-ran) in mobile cloud computing systems," in *IEEE Computer Communications Conference Workshops (INFOCOM WKSHPS)*, Apr. 2014, pp. 369–374.

[85] A. Gamage, H. Liang, and X. Shen, "Two time-scale cross-layer scheduling for cellular/wlan interworking," *IEEE Transactions on Communications*, vol. 62, no. 8, pp. 2773–2789, Aug. 2014.

[86] A. Qureshi, "Power-demand routing in massive geo-distributed systems," Ph.D. dissertation, Massachusetts Institute of Technology, 2010.

[87] G. Liang and U. Kozat, "Fast cloud: Pushing the envelope on delay performance of cloud storage with coding," *IEEE/ACM Transactions on Networking*, vol. 22, no. 6, pp. 2012–2025, Dec. 2014.

[88] J. Proakis, *Digital Communications.* McGraw-Hill, 2001, vol. 4th ed.

[89] I. Gradshteyn and I. Ryzhik, *Tables of integrals, series and products.* Academic Press, 2007, vol. 7th ed.

[90] H. Hoffman, "A method for globally minimizing concave functions over convex sets," *Mathematical Programming*, vol. 20, no. 1, pp. 22–32, 1981.