Radio Resource Management in a Heterogeneous Wireless Access Medium

by

Muhammad Ismail Muhammad

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, 2013

© Muhammad Ismail Muhammad 2013

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

In recent years, there has been a rapid evolution and deployment of wireless networks. In populated areas, high-rate data access is enabled anywhere and anytime with the pervasive wireless infrastructure such as the fourth-generation (4G) cellular systems, IEEE 802.11-based wireless local area networks (WLANs), and IEEE 802.16-based wireless metropolitan area networks (WMANs). In such a heterogeneous wireless access medium, multi-radio devices become a trend for users to conveniently explore various services offered by different wireless systems. This thesis presents radio resource management mechanisms, for bandwidth allocation, call admission control (CAC), and mobile terminal (MT) energy management, that can efficiently exploit the available resources in the heterogeneous wireless medium and enhance the user perceived quality-of-service (QoS).

Almost all existing studies on heterogeneous networking are limited to the traditional centralized infrastructure, which is inflexible in dealing with practical scenarios, especially when different networks are operated by different service providers. In addition, in most current wireless networks, mobile users are simply viewed as service recipients in network operation, with passive transceivers completely or partially under the control of base stations or access points. In this thesis, we present efficient decentralized bandwidth allocation and CAC mechanisms that can support single-network and multi-homing calls. The decentralized architecture gives an active role to the MT in the resource management operation. Specifically, an MT with single-network call can select the best wireless network available at its location, while an MT with multi-homing call can determine a required bandwidth share from each network to satisfy its total required bandwidth. The proposed mechanisms rely on cooperative networking and offer a desirable flexibility between performance measures (in terms of the allocated bandwidth per call and the call blocking probability), and between the performance and the implementation complexity.

With the increasing gap between the MT demand for energy and the offered battery capacity, service degradation is expected if the MT cannot efficiently manage its energy consumption. Specifically, for an uplink multi-homing video transmission, the existing studies do not guarantee that the MT available energy can support the entire call, given the battery energy limitation. In addition, the energy management mechanism should take account of video packet characteristics, in terms of packet distortion impact, delay deadline, and precedence constraint, and employ the available resources in the heterogeneous wireless medium. In this thesis, we present MT energy management mechanisms that can support a target call duration, with a video quality subject to the MT battery energy limitation. In addition, we present a statistical guarantee framework that can support a consistent video quality for the target call duration with minimum power consumption.

Acknowledgements

First, I would like to express the deepest appreciation to my supervisor, Professor Weihua Zhuang. I am deeply grateful to all the time and effort she gave me during my Ph.D. program. Being an outstanding supervisor, she helped me to lead a successful and a quite enjoyable Ph.D. study. I look up to Professor Zhuang and consider her to be my role model. During my Ph.D. study, I tried my best to learn from her not just how to be a good researcher and successful Professor but also how to be a better person. I learned from her the importance of balancing the effort I put in my work with living and enjoying my life. Also, I learned from her to accept people criticisms in a positive way and always to use others' comments to improve myself. Thanks to Professor Zhuang's guidance and constant support, I have significantly improved my curriculum vitae by the end of my Ph.D. program. I believe that, in life, it is important to have a mentor, and my mentor is Professor Zhuang.

Furthermore, I would like to thank my thesis advisory committee members Professor Kumaraswamy Ponnambalam, Professor Guang Gong, Professor Patrick Mitran, and my external examiner Professor Rong Zheng for their valuable comments and suggestions which helped to improve the quality of my thesis presentation.

I would like to offer my special thanks to Professor Xuemin Shen for the great insights he shared with us in the broadband communication research (BBCR) group meetings. I learned from him that it is not sufficient to publish your research work but it is more important to achieve high impact with your research work. He always encourages us to do a solid research work and build a strong curriculum vitae to land a good career in the future. I will always remember the great moralities I learned from him.

I am in debt to Canada and its people for their great hospitality and friendly manners. They made it very easy for me to adapt to the new life in Canada and feel like at home.

Also, I would like to thank all my friends for the great time we had together. I will not mention their names as I am afraid to forget any. I shall always remember you and value your friendship as you have been there for me when I needed you most.

No words on earth can express my love and gratitude to my dear parents and lovely sister whose prayers helped me through every step in my life.

To my dear parents, Ismail and Wafaa, lovely sister Dina,
and great mentor Weihua Zhuang

"People fail not because they lack the skill but because they lack the spirit,"

M. Ismail

Contents

List of Tables

List of Figures

List of Abbreviations

ABC	Always best connection
AP	Access point
ATM	Asynchronous transfer mode
BS	Base station
CAC	Call admission control
CBR	Constant bit rate
CDMA	Code division multiple access
CDF	Cumulative distribution function
CCDF	Complement cumulative distribution function
CPRA	Constant price resource allocation
CORA	Centralized optimal resource allocation
CPA	Cutting plane algorithm
CA	Content aware
CSI	Channel state information
DORA	Decentralized optimal resource allocation
DSRA	Decentralized sub-optimal resource allocation
DAG	Directed acyclic graph

EIA	Energy independent approach
EDFA	Earliest deadline first approach
EPA	Equal power allocation
EEF	Equal energy framework
FPS	Frame per second
GoP	Group of picture
GA	Greedy approach
IP	Internet protocol
ID	Identification
KKT	Karush-Kuhn-Tucker
KP	Knapsack problem
MT	Mobile terminal
MINLP	Mixed integer non-linear program
MIP	Mixed integer program
MKP	Multiple knapsack problem
NUM	Network utility maximization
NP	Non-deterministic polynomial
PBRA	Prediction based resource allocation
PDF	Probability density function
PMF	Probability mass function
PC-KP	Precedence-constrained knapsack problem
PC-MKP	Precedence-constrained multiple knapsack problem
QoS	Quality of service
RSS	Received signal strength
SGF	Statistical guarantee framework

SVC	Scalable video coding
SNR	Signal-to-noise ratio
TEF	Total energy framework
VBR	Variable bit rate
WLAN	Wireless local area network
WMAN	Wireless metropolitan area network

List of Symbols

A	Network assignment vector for MTs with single-network service
a_m	Network assignment for MT m
B	Bandwidth allocation matrix from each network n BS/AP s to each MT m
B_m	Required bandwidth by MT m with CBR service
$B_m^{\min}$	Minimum required bandwidth by MT m with VBR service
$B_m^{\max}$	Maximum required bandwidth by MT m with VBR service
b_{nms}	Allocated bandwidth from network n to MT m through BS/AP s
b_u	Uplink allocated bandwidth to MT radio interface u
C_n	Transmission capacity for each BS/AP of network n
C_{lk}^n	Maximum number of network n subscribers with service class l that can be served in service area k
C_{lvk}	Maximum number of calls for service type v service class l in service area k for a given network subscribers
c_u	Unused transmission capacity for radio interface u
d_f	Delay deadline of a packet in frame f
E	MT available energy at the beginning of the call

E_c	Energy budget per time slot for ECA video transmission
E_t	MT available energy at time slot t
E_{ct}	Energy budget per time slot, starting from time slot t
$\mathcal{F}$	Set of video frames available for transmission in a given time slot
$F_{T_c^{lv}}(t)$	CDF of the duration of a video call of MT with service class l and service type v
$F_{T_r^k}(t)$	CDF of the residence time for a MT in service area k
$F_{T_h^{lvk}}(t)$	CDF of the channel holding time for an MT with service type v and class l in service area k
F	Total number of video frames available for transmission in a given time slot
$F_Q(q)$	CDF of the video quality that can be achieved
$f_{T_c^{lv}}(t)$	PDF of the duration of a video call of MT with service class l and service type v
$f_{T_r^k}(t)$	PDF of the residence time for a MT in service area k
$f_{T_h^{lvk}}(t)$	PDF of the channel holding time for an MT with service type v and class l in service area k
$f_{G_f}(g_f)$	PMF of number of packets in video frame f
$f_{T_{lk}}(t)$	PDF of call inter-arrival time for service class l in service area k
$f_\delta(t)$	PDF of the time duration between two successive execution of the J iteration signalling exchange for the DORA mechanism
f_{lks}^{o+1}	Maximum number of MTs with single-network calls and service class l in service area k which can be supported by BS/AP s during period T_{o+1}
fb_{lks}	flag bit to indicate if a new incoming call from a given network subscribers with single-network service and service class l in service area k can be

	admitted to BS/AP s
$f_Q(q)$	PMF of the video quality that can be achieved
$f_{\Upsilon_u}(\gamma_u)$	Received SNR PDF for radio interface u
G_f	Number of packets in video frame f
$G_f^{\max}$	Maximum allowed number of packets in video frame f
$\mathcal{G}_u$	Set of packets scheduled to radio interface u in a given time slot
$\mathcal{G}$	Set of packets scheduled for transmission in a given time slot
$\bar{\mathcal{G}}$	Set of unscheduled packets in a given time slot
$H(\cdot)$	Dual function
i_f	Distortion impact value for video packet of frame f
J	Total number of iterations required for the DORA mechanism to converge to the optimal solution
j	Iteration index
$\mathcal{K}$	Set of service areas in the geographical region
K	Total number of service areas in the geographical region
$\mathcal{L}_v$	Set of service classes for service type v
L_v	Total number of service classes for service type v
$L_{nn'}$	Number of assigned subscribers of network n, with single-network service, to network n'
$L(\cdot)$	Langrangian function
l_f	Video packet length (in bits) for frame f
$\mathcal{M}$	Set of MTs in the geographical region
$\mathcal{M}_k$	Set of MTs in service area k
$\mathcal{M}_{ns}$	Set of MTs the coverage area of network n BS/AP s
$\mathcal{M}_{ns1}$	Set of MTs in coverage area of network n BS/AP s whose home network

	is network n
$\mathcal{M}_{ns2}$	Set of MTs in coverage area of network n BS/AP s whose home network is not network n
$\mathcal{M}_{vk}$	Set of MTs with service type v in service area k
$\mathcal{M}_{r1}$	Set of MTs with CBR service
$\mathcal{M}_{r1}$	Set of MTs with VBR service
$\vec{\mathcal{M}}_{lkn}^{o+1}$	Vector of predicted number of network n subscribers with service class l in service area k that will be present in the service area at period T_{o+1}
M_{lvk}	Number of existing calls of service type v and service class l in service area k for a given network subscribers
M_{nkr}	Number of network n subscribers with service r in service area k, $r = 1$ for CBR and $r = 2$ for VBR
M_{lk}^{n}	Number of calls of service class l in service area k for network n subscribers
$\widehat{M}_{lk}^{n}$	Target number of calls of service class l in service area k for network n subscribers, in the CPRA mechanism
$\widetilde{M}_{lk}^{n}(T_{o+1})$	Maximum predicted number of network n subscribers with service class l in service area k that will be present in the service area at period T_{o+1}
$\mathcal{N}$	Set of available wireless networks in the geographical region
$\mathcal{N}_k$	Set of available wireless networks in service area k
N	Total number of wireless networks in the geographical region
$O(\cdot)$	The O-notation that reflects the computational complexity of an algorithm
$\mathbf{Pr}(\cdot)$	The probability of occurance of a given event
P_{u,v_u}	Probability that the received SNR from radio interface u exceeds

threshold Γ_{u,v_u}

p_u Instantaneous transmission power allocated by the MT to radio interface u for video transmission

$\bar{p}_u$ Average transmission power allocated by the MT to radio interface u for video transmission

p_u^{ϑ} Single point of power allocated by the MT to radio interface u in the domain of a logarithmic function

$\mathcal{Q}$ Set of different data rates and packet encoding combinations that result in the same video quality q

q_l Video quality lower bound

q_t Resulting video quality that can be achieved at time slot t

$\mathcal{R}$ Set of data rates that can be supported on each radio interface

R_u Current used transmission capacity for radio interface u

r_u Achieved data rate for radio interface u

$r(z_f)$ Required minimum data rate to transmit video packet z_f

r_{u,v_u} Data rate that can be supported at radio interface u from a discrete set $\mathcal{V}$

r Total required data rate to satisfy a video quality lower bound q_l

$\mathcal{S}_n$ Set of BSs/APs of network n in the geographical region

$\mathcal{S}_{nk}$ Set of BSs/APs of network n in service area k

$\mathcal{S}_k$ Set of BSs/APs from all networks in service area k

S_n Total number of BSs/APs of network n in the geographical region

$\mathcal{T}$ Set of time slots

$\vec{\mathcal{T}}^{o}_{lkn}$ A time vector of arrival events for calls of network n subscribers with service class l in service area k

T_c^{lv}	Duration of a video call of MT with service class l and service type v
T_h^{lvk}	Channel holding time for a MT with service class l and service type v in service area k
$\bar{T}_c^{lv}$	Mean duration of a video call of MT with service class l and service type v
T_r^k	User residence time in service area k
t_o	The beginning of period T_o
$\mathcal{U}$	Set of radio interfaces used by the MT for video transmission
U_{ns}	Total utility of network n BS/AP s
U	Total utility in the geographical region
$u_{nms}(b_{nms})$	Utility of network n allocating bandwidth b_{nms} to MT m through BS/AP s
v	Service type, $v = 1$ for single-network call, $v = 2$ for multi-homing call
w_{nms}	Binary network assignment variable for MT m with single-network service to network n BS/AP s
x_{zu}^f	Video packet scheduling decision for packet z from frame f to radio interface u
y_{zf}	Index of the radio interface where packet z_f is currently assigned to
$\mathcal{Z}_z^f$	Set of ancestors of packet z from frame f
υ_{lvk}	Poisson process parameter for service type v with service class l in service area k
ς_{lv}	Mice-elephant parameter in the heavy-tailed PDF
τ	Prediction duration in PBRA mechanism
$\widetilde{\tau}$	Video transmission time slot duration
$\Delta D_{f+1,f}$	Delay deadline difference between two consecutive frames

γ_u The received SNR at the BS/AP communicating with the MT radio interface u

ω_{nms} Priority parameter set by network n BS/AP s on its resources for MT m

λ_{ns} Langrangian multiplier for network n BS/AP s capacity constraint

ν_m Langrangian multiplier for MT m, with CBR service, QoS constraint

$\nu_m^{(1)}$ Lagrangian multiplier corresponding to the maximum required bandwidth constraint of MT m with single-network VBR service

$\nu_m^{(2)}$ Lagrangian multiplier corresponding to the minimum required bandwidth constraint of MT m with single-network VBR service

$\mu_m^{(1)}$ Lagrangian multiplier corresponding to the maximum required bandwidth constraint of MT m with multi-homing VBR service

$\mu_m^{(2)}$ Lagrangian multiplier corresponding to the minimum required bandwidth constraint of MT m with multi-homing VBR service

φ Lagrangian multiplier corresponding to the MT energy budget constraint

ψ Small tolerance

Ψ_{u,v_u} Probability that data rate r_{u,v_u} is used at MT radio interface u

σ Total time duration of the signalling exchange in the DORA mechanism

α Sufficiently small fixed step size

ϵ_{lk}^n Upper bound on call blocking probability for network n subscribers with service class n in service area k

υ_{lk}^n Arrival rate of new and handoff calls from network n subscribers with service class l in service area k

κ_{lkn} Mean of M_{lk}^n Poisson process

ϕ_τ^{lkn} Probability that a call of network n subscribers with service class l which is in service area k at time t_π^o is still present in the same service area at

	time $t_{\pi}^{o} + \tau$
θ_{τ}^{lkn}	Probability that a call of network n subscribers with service class l that arrives in service area k during $(t_{\pi}^{o}, t_{\pi}^{o} + \tau]$ is still present in the same service area at time $t_{\pi}^{o} + \tau$
$\rho(\varkappa_1, \varkappa_2)$	Binomial random variable with parameters $\varkappa_1$ and $\varkappa_2$
$\xi(\varkappa)$	Poisson random variable with mean $\varkappa$
χ_1	Average number of call arrivals over a period
χ_2	Average number of call departures over a period
δ	Time duration between two successive execution of the J iteration signalling exchange for the DORA mechanism
Ω_u	Channel power gain between the MT and BS/AP communicating with radio interface u
$\bar{\Omega}_u$	Average channel power gain between the MT and BS/AP communicating with radio interface u
η_0	One-sided noise power spectral density
η_u	Background noise power, in watt, for radio interface u
Θ	Set of all points in the domain of a logarithmic function
$\widetilde{\Theta}$	Subset of points in the domain of a logarithmic function
ε	Upper bound on video quality violation probability
ε_s	Success probability for delivering a video call with duration T_c and quality $\geq q_l$
Γ_{u,v_u}	Threshold on received SNR from MT radio interface u to support data rate r_{u,v_u}

Chapter 1

Introduction

The past decade has witnessed an increasing demand for wireless communication services, which extended beyond telephony services to include video streaming and data applications. This results in a rapid evolution and deployment of wireless networks. These wireless networks have different service capabilities in terms of bandwidth, latency, coverage area, and cost. With overlapped coverage from these networks, the wireless communication medium has become a heterogeneous environment. Radio resource management mechanisms play a vital role in such a networking environment to efficiently utilize the available resources and satisfy the user required service quality. This chapter introduces the heterogeneous wireless access medium, discusses cooperative radio resource management, and presents the thesis motivations and contributions.

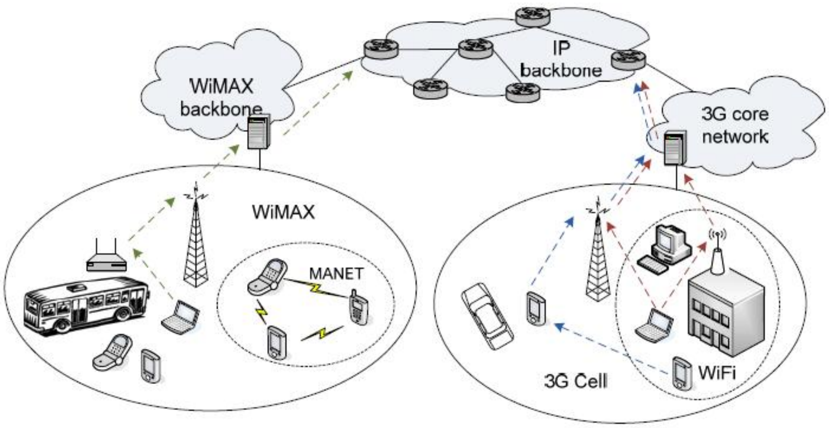


Figure 1.1: An illustration of the heterogeneous wireless communication network architecture.

1.1 The Heterogeneous Wireless Communication Network Architecture

Currently, there exist different wireless networks that offer a variety of access options. The wireless access networks include the cellular networks, the IEEE 802.11 wireless local area networks (WLANs), and the IEEE 802.16 wireless metropolitan area networks (WMANs). These networks have complementary service capabilities. For example, the IEEE 802.11 WLANs can support high data rate services in hot spots, whereas the cellular networks and the IEEE 802.16 WMANs can offer broadband wireless access over long distances.

The basic components of the heterogeneous wireless communications network architecture are mobile terminals (MTs), base stations (BSs) / access points (APs), and a core Internet protocol (IP) based network [1], as shown in Figure 1.1. Currently, mobile users are viewed as service recipients in the network operation, with passive transceivers which operate under the control of BSs or APs. It is envisioned that the future MTs will be more powerful and take a more active role in network operation and service delivery. Also, MTs are currently equipped with multiple radio interfaces for network access. Two service types can be recognized in this environment, namely single-network and multi-homing services [2]. In single-network services, an MT can connect to the best wireless network available at its location to get service. Multi-homing techniques maintain multiple simultaneous associations of an MT with different radio access networks. Hence, in a multi-homing call, the MT can obtain its service by simultaneously connecting to all available wireless networks and aggregating the offered resources from these networks. Fixed network components, such as BSs and APs, provide a variety of services to MTs, which include access to the Internet, mobility management, and resource management. Finally, the core network serves as the backbone network with Internet connectivity and packet data services.

1.2 Cooperative Networking in a Heterogeneous Wireless Access Medium

Despite the fierce competition in the wireless service market, the aforementioned wireless networks will coexist due to their complementary service capabilities. In this heterogeneous wireless access medium with overlapped coverage from different networks, cooperative networking will lead to better service quality to mobile users and enhanced performance for the networks [3].

As for mobile users, cooperative networking solutions for heterogeneous wireless networks can result in two major advantages. The first advantage is that mobile users can enjoy an always best connection. This means that a mobile user can always be connected to the best wireless access network available at his/her location. Traditionally, an

MT can keep its connection active when it moves from one attachment point to another through handoff management [4]. Hence, mobile users can enjoy an always connected experience. This is enabled by horizontal handoff, which represents a handoff within the same wireless access network, as in the handoff between two APs in a WLAN or between two BSs in a cellular network. However, in the presence of various wireless access networks with overlapped coverage, the user experience is now shifted from always connected to always best connected (ABC). The ABC experience is mainly supported by vertical handoffs among different networks. A vertical handoff represents a handoff between different wireless access networks, as in the handoff between a BS of a cellular network and an AP of a WLAN. Unlike horizontal handoffs, vertical handoffs can be initiated for convenience rather than connectivity reasons [5]. Hence, vertical handoffs can be based on service cost, coverage, transmission rate, information security, and user preference. Through cooperative networking, the inter-network vertical handoffs can be provided in a seamless and fast manner. This can support a reliable end-to-end connection at the transport layer, which preserves service continuity and minimizes disruption. The second advantage of cooperative networking for mobile users is that users can enjoy applications with high required data rates, e.g. video streaming and data applications, through aggregating the offered bandwidth from different networks. This is enabled by the multi-homing capabilities of MTs, where users can receive their required bandwidth through different networks and use multiple threads at the application layer. Recently, video streaming has gained an increasing popularity among mobile services. It has been reported that 65% of all mobile data traffic, by the end of 2015, will be due to mobile video traffic [6]. Multi-homing video transmission can benefit the achieved video quality in many aspects [7, 8]. Firstly, sending video packets over multiple networks increases the amount of aggregate bandwidth available to the application and hence increases the quality of the delivered service. Secondly, sending video packets over multiple networks can reduce the correlation between consecutive packet losses due to transmission errors or networks' congestion. Finally, video packet transmission over multiple networks allows for better mobility support which significantly reduces the probability of an outage when communication is lost with the current serving network due to user mobility out of its coverage area. In this context, cooperation is required among different networks so as to coordinate their allocated bandwidth to the MT such that the total bandwidth allocation from multiple networks satisfies the user total required bandwidth.

In addition, service providers can benefit from cooperative networking to enhance network performance in many ways. For instance, multiple heterogeneous networks can cooperate to provide a multi-hop backhaul connection in a relay manner. This results in an increase in these networks' coverage area at a reduced cost as compared to deploying more BSs for coverage extension. Also, load balancing among different networks can be supported through cooperative networking which helps in avoiding call traffic overload situations. Moreover, cooperative networking can achieve energy saving for green radio communications. Networks with overlapped coverage area can alternately switch their BSs on and off according to spatial and temporal fluctuations in call traffic load, which reduces their energy consumption and provides an acceptable quality-of-service (QoS) performance for the users [9].

1.3 Motivations and Research Contributions

This research is to develop a radio resource management framework in a heterogeneous wireless access medium. Specifically, we focus on bandwidth allocation, call admission control (CAC), and MT energy management. The bandwidth allocation and CAC mechanisms aim to satisfy the QoS requirements of mobile users and achieve efficient utilization of the available resources from different networks [10]. The MT energy management aims to support a high quality sustainable multi-homing video transmission, over a target call duration, in the heterogeneous wireless access medium. The research motivations and contributions are discussed in the following.

1.3.1 Bandwidth Allocation and CAC

Almost all the existing research works in literature on radio resource allocation in a heterogeneous wireless access medium focus on supporting either a single-network or a multi-homing service. However, it is envisioned that both service types will coexist in the future networks [11, 12]. This is because not all MTs are equipped with multi-homing capabilities, and not all services require high resource allocation that calls for a multi-homing support. As a result, some MTs will have to utilize a single-network service. Moreover, even for an MT with a multi-homing capability, the MT utilization of the multi-homing service should depend on its residual energy. Hence, when no sufficient energy is available at the MT, the MT should switch from a multi-homing service to a single-network one where the radio interface of the best available wireless network is kept active while all other interfaces are switched off to save energy. This motivates the requirement to develop a radio resource allocation mechanism that can support both single-network and multi-homing services in a heterogeneous wireless access medium. However, there are many technical challenges, as discussed in the following.

A. Decentralized Implementation

In literature, almost all radio resource allocation mechanisms need a central resource manager in order to meet service quality requirements in such a heterogeneous wireless access medium. The need for the central resource manager for single-network services is due to the fact that a global view of the available resources at all networks is required in order to select the best available wireless access network given the MT required bandwidth. For multi-homing services, the central resource manager coordinates the allocated bandwidth from different networks such that the total bandwidth allocation to a given MT satisfies the total required resources by the MT. Hence, the central resource manager should have global information of network resource availability, and perform network selection for MTs with single-network services and bandwidth allocation for MTs with single-network and multi-homing services. However, the assumption of the central resource manager is not practical in a case that the networks are operated by different service providers. This is because the central resource manager would raise some issues [13]:

1. The central resource manager is a single point of failure. If it breaks down, the whole single-network and multi-homing services fail and this may extend to the operation of the other networks;

2. It is difficult to determine which network should be in charge of the operation and maintenance of the central resource manager, taking account that the network in charge will control the radio resources of other networks;

3. Modifications are required in different network structures in order to account for the central resource manager.

As a result, it is desirable to have decentralized radio resource allocation. In this context, an MT with single-network service can select the best wireless access network available at its location and ask for its required bandwidth from this network. While an MT with multi-homing service can ask for the required bandwidth share from each available network so as to satisfy its total required bandwidth. Each network then can perform its own bandwidth allocation and CAC without the need for a central resource manager. However, with user arrivals and departures, achieving the optimal allocation for a given connection at any point of time would trigger reallocations of a whole set of connections. This will take place with every service request arrival or departure and a considerable amount of signalling information has to be exchanged among different network entities. Hence, through network cooperation, we have developed an efficient decentralized implementation of the radio resource allocation that balances bandwidth allocation and call blocking probability with the associated signalling overhead. Through cooperative resource allocation, different networks can coordinate their resource allocation in order to support the required bandwidth of each call, satisfy a target call blocking probability, and eliminate the need for a central resource manager, while reducing the amount of signalling overhead over the air interface.

B. Service Differentiation

In general, mobile users are the subscribers of different networks. As a result, the service requests of different MTs should not be treated in the same manner by each network. Instead, it is more practical that each network gives a higher priority in allocating its resources to its own subscribers as compared to other users. Hence, a priority mechanism should be in place to enable each network to assign different priorities to MTs on its resources.

Considering the aforementioned challenges in designing a resource allocation mechanism to support both single-network and multi-homing services in a dynamic environment, we have taken the following steps to develop our radio resource allocation solutions:

1. Static multi-homing bandwidth allocation in Chapter 2 [13, 14]: In this step, we have investigated a system model with only multi-homing calls, and without considering the arrival of new calls or departure of existing ones. This simplifies the problem under consideration due to two reasons. Firstly, in the absence of a network assignment problem, we focus on finding the optimal bandwidth allocation

from each network to a given connection in order to satisfy its total required bandwidth. Secondly, due to the static nature of the system model, there are no perturbations associated with the number of MTs in the system. Hence, no bandwidth reallocations are necessary, and the signalling between MTs and BSs/APs occurs only in the call setup. We have developed a decentralized implementation of the bandwidth allocation in order to identify the role of each network entity in this architecture. In addition, we have enabled each network to give a higher priority in allocating its resources to its own subscribers as compared to other users;

2. Dynamic multi-homing bandwidth allocation and CAC in Chapter 3 [15]: We consider the stochastic mobility and call traffic models for the users and service requests, respectively. The system experiences perturbations in the call traffic load. This triggers resource reallocations for all the existing connections, and results in a considerable amount of signalling overhead. Hence, we have extended the bandwidth allocation in Chapter 2 in order to provide an efficient bandwidth allocation and CAC mechanism that can balance the resource allocation with the associated signalling overhead through short-term call traffic load prediction and network cooperation;

3. Single-network and multi-homing bandwidth allocation and CAC in Chapter 4 [11, 12]: We extend the ideas presented in Chapters 2 and 3 to include single-network calls in the system model. Hence, the radio resource allocation mechanism is of twofold: To determine the network assignment of MTs with single-network service to the available wireless access networks, and to find the corresponding bandwidth allocation to MTs with single-network and multi-homing services.

1.3.2 MT Energy Management for Multi-homing Video Transmission

A research topic that is not well investigated in the context of heterogeneous wireless networks is related to the MT battery energy management. This research topic is motivated by the increasing gap between the demand for energy and the offered MT battery capacity [16]. Hence, even if the MT is allocated sufficient bandwidth through multiple networks, service degradation is expected if the MT cannot efficiently manage its energy consumption. The problem is further complicated when other service quality aspects are taken into consideration. Consider an uplink multi-homing video transmission for posting on social network sites [6]. In multi-homing video transmission, packet scheduling determines the assignment of a packet to a radio interface, given the packet required data rate and the radio interface characteristics in terms of channel condition and available bandwidth. Video packets which missed their playback deadlines should be dropped in order not to waste the network resources. The strategy of packet dropping and assignment to different radio interfaces should minimize the total video quality distortion. Thus, a video packet scheduling mechanism should be content-aware in order to transmit the most valuable packets and, whenever necessary, drop the least valuable ones. On the other hand, MT battery energy limitation is a concern in multi-homing video transmission. The MT operational time in between battery charging has become a significant factor in the user perceived QoS [17]. In addition to developing new battery technology with improved capacity, the operational period of an MT between battery chargings can be extended through managing its energy consumption [18]. Thus, packet scheduling should be energy-aware in order to work under the MT battery limitation. However, this concern has been mostly overlooked so far in designing a video streaming packet scheduling mechanism.

Minimizing energy consumption (e.g. [19]) does not guarantee that the MT available energy can support video transmission for the target call duration, given the battery energy limitation. In addition, related works for single-path video transmission deal with an energy budget per time slot (e.g. [20]), in the presence of an energy management sub-system which can determine the energy budget per time slot to ensure a sustainable video transmission for the target call duration. However, not many details are given regarding this energy management sub-system. A simple energy management sub-system can equally distribute the MT available energy over different time slots. Given the time varying video packet encoding and channel conditions at different radio interfaces, using this uniform energy distribution will lead to inconsistent temporal fluctuations in the video quality. Instead, an appropriate energy management sub-system should use the MT energy in a way such that it can support the target call duration with a consistent video quality over different time slots, independent of varying packet encoding and channel conditions.

In literature, none of the existing works provides a statistical guarantee for multi-homing video transmission to complete the call with a consistent quality. One approach to provide performance statistical guarantee is through the effective bandwidth and effective capacity concepts, as in [21]. However, [21] mainly addresses single-network video transmission and does not provide an energy efficient design. Adopting the effective bandwidth and effective capacity concepts in providing performance statistical guarantees imposes some restrictions on the service process, in order to develop an effective capacity expression that is easy to compute and to handle. Hence, the problem formulation will not incorporate many details (i.e. MT available energy at the beginning of the call, the target call duration, radio interface characteristics in terms of offered bandwidth and time varying channel conditions, and video packet characteristics in terms of distortion impact, delay deadline, and packet encoding).

Considering the aforementioned challenges in designing an energy management mechanism for MTs to support a sustainable multi-homing video transmission, over the target call duration, in a heterogeneous wireless access medium, we have taken the following steps to develop our energy management solutions:

1. Energy and content aware multi-homing video packet scheduling in Chapter 5 [22]: In this step, an energy budget per time slot is assumed for the MT, given the MT available energy at the beginning of the call and the target call duration. We mainly focus on a single time slot with a fixed channel condition and study how to perform multi-homing video packet scheduling to improve the resulting video quality under the battery energy constraint. The energy and content aware multi-homing video transmission problem is formulated to capture i) the video packet characteristics in terms of distortion impact, delay deadlines, and packet dependence relation, ii) the characteristics of the multiple wireless interfaces in terms of the channel conditions and the allocated bandwidth, and iii) the MT battery energy limitation. The problem solution determines the power allocation for the radio interfaces, assigns the most valuable packets to different radio interfaces, and, if necessary, drops some packets given the MT energy constraint, and with the objective of minimizing video quality distortion;

2. Statistical QoS guarantee for sustainable multi-homing video transmission in Chapter 6 [23, 24]: In this step, we extend our ideas developed in Chapter 5 to consider a system with multiple time slots and time varying channel conditions over different radio interfaces. Through statistical video quality guarantee, we enable the MT to determine a target video quality lower bound that can be supported for the target call duration with a pre-defined success probability. The target video quality lower bound captures the MT available energy at the beginning of the call, the available bandwidth and time varying channel conditions at different radio interfaces, the target call duration, and the video packet characteristics in terms of distortion im-

pact, delay deadlines, and video packet encoding statistics. The MT then adapts its energy consumption to support at least the target video quality lower bound during the call.

The relation between the problems addressed in the thesis is described in Figure 1.2. For a given video call, bandwidth allocation and CAC can be established using the mechanisms given in Chapters 2 - 4. Then, power allocation and packet scheduling to different radio interfaces can be performed using the mechanisms in Chapters 5 and 6 for multi-homing video transmission.

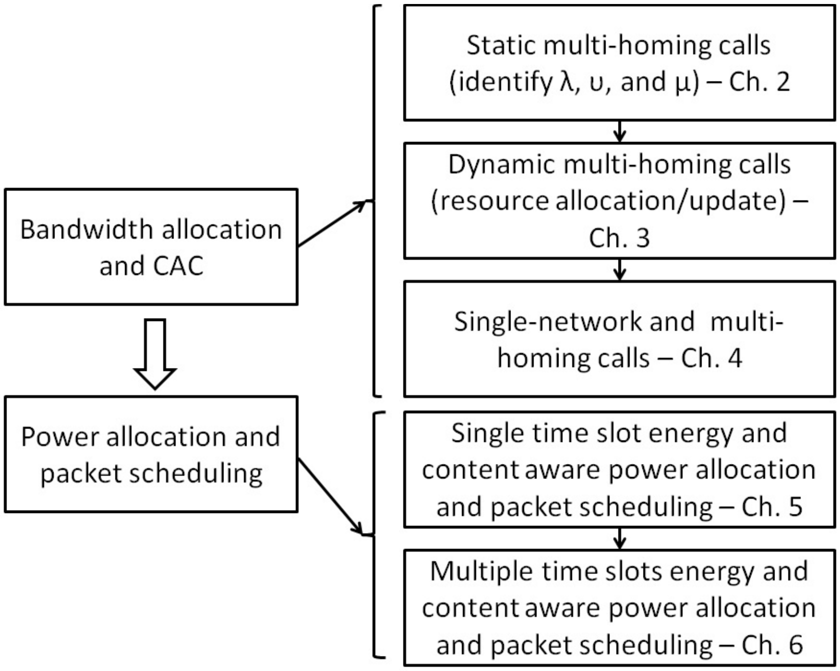

Bandwidth allocation and CAC
Power allocation and packet scheduling
Static multi-homing calls (identify λ, υ, and μ) – Ch. 2
Dynamic multi-homing calls (resource allocation/update) – Ch. 3
Single-network and multi-homing calls – Ch. 4
Single time slot energy and content aware power allocation and packet scheduling – Ch. 5
Multiple time slots energy and content aware power allocation and packet scheduling – Ch. 6

Figure 1.2: An illustration of the relation between the problems addressed in the thesis.

1.4 Outline of the Thesis

This thesis is organized as follows. Radio resource management mechanisms for bandwidth allocation and CAC are given in Chapters 2 - 4. Specifically, in Chapter 2, we present a decentralized optimal resource allocation (DORA) mechanism to support MTs with multi-homing service. The DORA mechanism is limited to a static system model, without new arrival and departure of calls in different service areas. In Chapter 3, we discuss the challenges that face the DORA mechanism in a dynamic system and propose a sub-optimal decentralized resource allocation (PBRA) mechanism that can address these challenges. In Chapter 4, we further extend the radio resource allocation problem to consider the simultaneous presence of both single-network and multi-homing services in the networking environment. We present a decentralized implementation for the radio resource allocation using a decentralized sub-optimal resource allocation (DSRA) mechanism. Energy management mechanisms for sustainable multi-homing video transmission are given in Chapters 5 and 6. In Chapter 5, we propose an energy and content aware video transmission mechanism that incorporates the energy limitation of MTs and the QoS requirements of video streaming applications, and employs the multi-homing capability in a heterogeneous wireless access medium. In Chapter 6, we extend the ideas presented in Chapter 5 to provide a statistical guarantee framework (SGF) for the resulting video quality over time varying channel conditions and video packet encoding statistics. Finally, we conclude the thesis and discuss further research in Chapter 7.

Chapter 2

Static Decentralized Multi-homing Bandwidth Allocation

Mutli-homing bandwidth allocation is considered to be a promising solution that can efficiently exploit the available resources in a heterogeneous wireless access medium to satisfy required bandwidth, reduce call blocking probability, and allow for better mobility support. The main challenge in designing a multi-homing bandwidth allocation mechanism is how to coordinate the allocation from different networks so as to satisfy the user's required bandwidth while making efficient utilization of available network resources. One simple solution is to employ a central resource manager with a global view over the available resources and the required bandwidth for different calls, which can perform the necessary coordination among different networks. However, this solution is not practical in the case that those different networks are operated by different service providers. Hence, the question now is how to coordinate the resource allocation in different networks without a central resource manager. In addition, it is more practical that every network prioritizes bandwidth allocation to its own subscribers as compared to other users. In this chapter, we present a decentralized optimal bandwidth allocation mechanism that enables each MT to coordinate the allocation from different networks in order to satisfy its required bandwidth, and allows each network to give a higher priority in allocating its resources to its own subscribers. We consider only multi-homing calls in the system model. Also, we consider a static system model, without arrivals of new calls or departures of existing ones. Our main objective is to identify the role of each entity in the heterogeneous wireless access medium to support a decentralized multi-homing bandwidth allocation.

2.1 Related Work

The problem of bandwidth allocation in heterogeneous wireless access networks is studied in [25] - [34]. The existing solutions can be classified in two categories based on whether a single radio interface or multiple radio interfaces of an MT are used simultaneously for the same application.

The single-network bandwidth allocation solutions are studied in [25] - [29]. In [25], a utility function based bandwidth allocation scheme is introduced for a single service class code division multiple access (CDMA) cellular network and WLAN. In [26], two resource management schemes are proposed for bandwidth allocation and admission control in a heterogeneous wireless access environment with different classes of service. The bandwidth allocation mechanism of [27] is to maximize the allocations under demand uncertainty while minimizing users' rejection and underutilization of different networks. In [28], a radio resource management mechanism is presented to maximize the minimum throughput among all users in the heterogeneous networks. The mechanisms provided in [25] - [28] needs a central resource manager to select the best network for a given connection from a set of available wireless networks, and then performs the bandwidth allocation for that connection from the selected network. In [29], a decentralized resource allocation mechanism is developed to find the optimum bandwidth allocation for a given set of voice users and best effort users in a heterogeneous wireless access environment. In this case, an MT selects the best network and the selected network then performs the bandwidth allocation for the connection. While a decentralized mechanism is developed in [29], only a single network is considered in obtaining the required bandwidth. In general, for centralized and decentralized mechanisms, the selection of the best available network depends on a pre-defined criterion [11]. One criterion is the received signal strength (RSS) [35], where the MT is assigned to the wireless network with the highest RSS from its BS or AP among all available networks. Another network selection criterion is the offered bandwidth [36, 37], where the MT is assigned to the network BS/AP with the largest offered bandwidth. Moreover, different network selection criteria, such as RSS, offered bandwidth, and monetary cost, can be combined in a utility function and the MT network assignment is based on the results of this function associated with the BSs/APs of the candidate networks [38, 39]. The single-network bandwidth allocation mechanisms suffer from the limitation that an incoming call is blocked if no network in the service area can individually satisfy the call required bandwidth. As a result, these mechanisms do not fully exploit the available resources from different networks.

The multi-homing bandwidth allocation solutions are studied in [30] - [34]. In [30], the concept of utility fairness is applied to allocate bandwidth to different types of traffic. In [31, 32], the problem of bandwidth allocation in a heterogeneous wireless access medium is formulated as a non-cooperative game, while in [33, 34] the problem is formulated using a cooperative game. In [30] - [34], each MT obtains its required bandwidth for a specific application from all available wireless access networks. This has the following advantages [40]. Firstly, with multi-homing capabilities, the available resources from different wireless access networks can be aggregated to support applications with high required bandwidth using multiple threads at the application layer; Secondly, these mechanisms allow for better mobility support since at least one of the used radio interfaces will remain active during the call; Finally, the multi-homing concept can reduce the call blocking rate and improve the system capacity.

However, these existing banwidth allocation mechanisms for a heterogeneous wireless access environment that support MTs with multi-homing capabilities need a central resource manager to perform the bandwidth allocation and CAC. The need for the central resource manager arises from the fact that the allocated bandwidth from each network BS/AP to a given connection should sum up to the bandwidth required by that connection. Hence, a global view of the BS/AP capacity of every network is needed to coordinate the allocations from different networks in order to satisfy the required bandwidth for that connection. This global view is provided by the central resource manager. This is not practical in a case that these networks are operated by different service providers. Hence, in such a networking environment it is desirable to have a decentralized solution that enables each network BS/AP to solve its own utility maximization problem and to perform its own bandwidth allocation, while at the same time cooperates with other available networks to support MTs with multi-homing capabilities.

In this chapter, a decentralized mechanism for bandwidth allocation in a heterogeneous wireless access medium for MTs with multi-homing capabilities is proposed. Each wireless access network BS/AP, in this mechanism, solves its own utility maximization problem to allocate its resources so that the MT requirement can be satisfied. Two classes of service are considered, namely, constant bit rate (CBR) and variable bit rate (VBR) services. When sufficient resources are available from different networks, VBR services are allocated the maximum required bandwidth. On the other hand, when all available networks with overlapped coverage areas reach their capacity limitation, the bandwidth allocation to VBR services is degraded towards the minimum required bandwidth using the resources from different overlapped networks. The work of [30] employs a utility fairness concept to ensure that all MTs with VBR service are degraded simultaneously with the same amount of resources within the same wireless access network and according to the same utility change among different wireless access networks. This, however, does not take into consideration the fact that different MTs are the subscribers of different networks and, as a result, they should not be treated equally by each network. It is more practical that each network supports first its own subscribers and ensures that they are satisfied with the maximum possible required bandwidth, while at the same time it supports the subscribers of other networks. To accomplish this, our proposed mechanism employs a priority scheme so that each network can give a higher priority in allocating its resources to its subscribers as compared to the other users.

2.2 System Model

In this section, we present the system model under consideration in terms of wireless networks, network subscribers and users, and service requests. These are discussed as follows.

A. Wireless Access Networks

Consider a geographical region with a set $\mathcal{N}$ of available wireless access networks, $\mathcal{N} = \{1, 2, \ldots, N\}$. Each network, $n \in \mathcal{N}$, is operated by a unique service provider and has a set, $\mathcal{S}_n$, of BSs/APs in the geographical region with $\mathcal{S}_n = \{1, 2, \ldots, S_n\}$. The BSs/APs of different networks have different coverage that overlaps in some areas. Hence, the geographical region is partitioned to a set $\mathcal{K}$ of service areas, $\mathcal{K} = \{1, 2, \ldots, K\}$. As shown in Figure 2.1, each service area $k \in \mathcal{K}$ is covered by a unique subset of networks' BSs/APs. Each BS/AP, $s \in \mathcal{S}_n$ for $n \in \mathcal{N}$, has a transmission capacity of C_n Mbps.

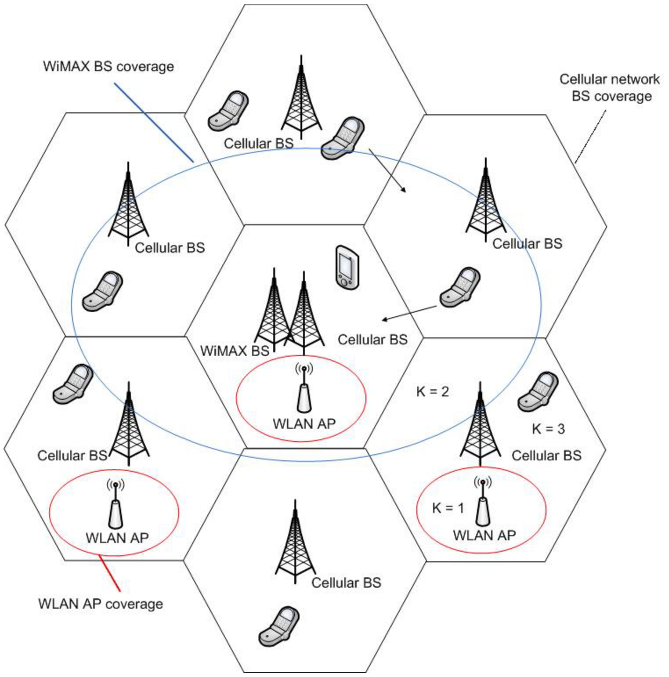


Figure 2.1: The networks' coverage areas.

B. Network Subscribers and Users

There are M MTs with multiple radio interfaces and multi-homing capabilities in the geographical region, given by the set $\mathcal{M} = \{1, 2, \dots, M\}$. Each MT has its own home network but can also get service from other available networks. Let $\mathcal{M}_{ns} \subset \mathcal{M}$ denote the set of MTs which lies in the coverage area of the sth BS/AP of the nth network. The set $\mathcal{M}_{ns}$ is further divided into two subsets, $\mathcal{M}_{ns1}$ to denote MTs whose home network is network n, and $\mathcal{M}_{ns2}$ to denote MTs whose home network is not network n. Hence, $\mathcal{M}_{ns1} \cup \mathcal{M}_{ns2} = \mathcal{M}_{ns}$, and $\mathcal{M}_{ns1} \cap \mathcal{M}_{ns2} = \emptyset$. An MT $m \in \mathcal{M}_{ns1}$ is referred to as network n subscriber, while an MT $m \in \mathcal{M}_{ns2}$ is referred to as network n user.

C. Service Requests

The service requests of MTs are expressed in terms of call required bandwidth. An MT can receive its required bandwidth from all available wireless access networks using its multi-homing capability. The allocated bandwidth from network n to an MT m through BS/AP s is given by b_{nms}, with $n \in \mathcal{N}$, $m \in \mathcal{M}_{ns}$, and $s \in \mathcal{S}_n$. Let B be a tensor of bandwidth allocation from each network n through BS/AP s to each MT m, $B = [b_{nms}], n \in \mathcal{N}, m \in \mathcal{M}, s \in \mathcal{S}_n$, with $b_{nms} = 0$ if MT m is not in the coverage area of network n BS/AP s.

The networks support both CBR and VBR services. An MT, m, with a CBR call requires a constant bandwidth B_m from all wireless access networks available at its location. On the other hand, an MT, m, with a VBR call requires a bandwidth allocation within a maximum value $B_m^{\max}$ and a minimum value $B_m^{\min}$. With sufficient available radio resources, the VBR call is allocated its maximum required bandwidth $B_m^{\max}$. When all BSs/APs reach their transmission capacity limitation C_n, the allocated bandwidth for the VBR call is degraded towards $B_m^{\min}$ in order to support more calls. Two examples of this service class are video and data calls. The key difference between video and data calls is the impact of the allocated bandwidth on the call presence in the system [41]. For video calls, the amount of the allocated bandwidth influences the perceived video quality experienced on the video terminal, while it does not affect the video call duration. On the other hand, bandwidth allocation to a data call affects its throughput and thus its duration. Let $\mathcal{M}_{r1}$ denote the set of MTs in the geographical region with CBR service, while $\mathcal{M}_{r2}$ denote the set of MTs in the geographical region with VBR service, and both are a subset of $\mathcal{M}$.

We consider call-level radio resource allocation (i.e. we do not deal with traffic on a packet-by-packet basis, but rather we treat traffic as a fluid flow with aggregate required bandwidth). The bandwidth allocation mechanism is to find the optimal resource allocation to a set of MTs in a particular service area from each of the available BSs/APs. As a first step, the resource allocation is performed according to the average call level statistics in different service areas [32]. Hence, a static system is investigated without arrivals of new calls or departures of existing ones. It is assumed that a call admission control procedure is in place [42], and a feasible resource allocation solution exists.

2.3 Formulation of the Bandwidth Allocation Problem

In this section, we discuss the problem formulation of bandwidth allocation for a static system of multi-homing MTs in the heterogeneous wireless access medium. A decentralized solution for the problem is then proposed based on the problem formulation.

The utility $u_{nms}(b_{nms})$ of network n allocating bandwidth b_{nms} to MT m through BS/AP s is given by

$$u_{nms}(b_{nms}) = \ln(1 + \eta_1 b_{nms}) - \eta_2(1 - \omega_{nms})b_{nms} \tag{2.1}$$

where η_1 and η_2 are used for scalability of b_{nms} [43], and $\omega_{nms} \in [0, 1]$ is a priority parameter set by network n BS/AP s on its resources for MT m. The attained network utility from the allocated bandwidth is a concave function of b_{nms} [44] and is given by the first term in the right hand side of (2.1) [30]. This term originates from the concept of proportionally fair resource allocation and satisfies the law of diminishing marginal utility [43]. The cost that the user pays for the allocated bandwidth is given by the second term in the right hand side of (2.1). This term is a linear function of the allocated bandwidth b_{nms}; hence, the more the allocated bandwidth, the higher the cost. The utility function of (2.1) involves a trade-off between the attained network utility and the cost that the user pays on the network radio resources [13]. The utility function of (2.1) is a concave function of the allocated bandwidth b_{nms} [43]. We employ priority parameter ω_{nms} set by network n BS/AP s to MT m to establish service differentiation among different users, which is given by

$$\omega_{nms} = \begin{cases} 1, & \forall m \in \mathcal{M}_{ns1} \\ \beta, & \forall m \in \mathcal{M}_{ns2} \end{cases} \tag{2.2}$$

where $\beta \in [0, 1)$. Using (2.2) in (2.1), the utility function for a network subscriber accounts only on the attained network utility by that subscriber, while a network user suffers from a trade-off between the attained network utility and the cost that the network sets on its resources [13]. This enables each network to give a higher priority in allocating its resources to its own subscribers than to other users. The allocated bandwidth to MTs with VBR service is reduced, when all networks in the geographical region reach their capacity limitation, in order to support more calls. However, each subscriber should be able to enjoy the resources of his/her own home network. Hence, it is desirable to differentiate the radio resource allocation performed by a network to its own subscribers and the allocation performed by that network to the other users. This is taken care of by the priority parameter ω_{nms} which gives a higher cost on the network resources for the network users than to the network subscribers. Each network, $n \in \mathcal{N}$, assigns a priority parameter value $\omega_{nms} \in [0, 1)$ on its resources for the users in its coverage area, while setting $\omega_{nms} = 1$ for its own subscribers. Hence, the subscribers of each network with VBR service enjoy their maximum required bandwidth using their home network radio resources. A network degrades its resource allocation to its own subscribers only so as not to violate the minimum required bandwidth of the other users.

The radio resource allocation objective of each network BS/AP is to maximize the total satisfaction for all MTs that lie within its coverage area, which is given by

$$U_{ns} = \sum_{m \in \mathcal{M}_{ns}} u_{nms}(b_{nms}), \quad \forall s \in \mathcal{S}_n, n \in \mathcal{N} \tag{2.3}$$

where U_{ns} is the total utility of network n BS/AP s.

The overall radio resource allocation objective of all networks in the geographical region is to find the optimal bandwidth allocation b_{nms}, $\forall n \in \mathcal{N}$, $\forall m \in \mathcal{M}$, $\forall s \in \mathcal{S}_n$, which maximizes the total utility in the region, given by

$$U = \sum_{n=1}^{N} \sum_{s=1}^{S_n} U_{ns}. \tag{2.4}$$

The total bandwidth allocation by each network n BS/AP s should be such that the total call traffic load in its coverage area is within the network BS/AP transmission capacity limitation C_n, that is

$$\sum_{m \in \mathcal{M}_{ns}} b_{nms} \leq C_n, \quad \forall s \in \mathcal{S}_n, n \in \mathcal{N}. \tag{2.5}$$

For an MT with CBR service, the total bandwidth allocation from all available wireless access networks to this MT should satisfy its application required bandwidth, that is

$$\sum_{n=1}^{N} \sum_{s=1}^{S_n} b_{nms} = B_m, \quad \forall m \in \mathcal{M}_{r1}. \tag{2.6}$$

As for an MT with VBR service, the total bandwidth allocation from all available wireless access networks to this MT should be within the application minimum required bandwidth $B_m^{\min}$ and the application maximum required bandwidth $B_m^{\max}$, that is

$$B_m^{\min} \leq \sum_{n=1}^{N} \sum_{s=1}^{S_n} b_{nms} \leq B_m^{\max}, \quad \forall m \in \mathcal{M}_{r2}. \tag{2.7}$$

Hence, the bandwidth allocation for MTs with multi-homing capabilities in the heterogeneous wireless access medium, for CBR and VBR services, can be expressed by the following optimization problem

$$\begin{aligned} &\max_{B \geq 0} \quad U \\ &s.t. \quad (2.5) - (2.7). \end{aligned} \tag{2.8}$$

Using the utility function definitions in (2.1), (2.3), and (2.4), the objective function of (2.8) is concave and the problem has linear constraints. Therefore, problem (2.8) is a convex optimization problem, and a local maximum is a global maximum as well [44]. Although problem (2.8) can be solved efficiently in polynomial time complexity in a centralized manner using a central resource manager, this is not practical in a case that these networks are operated by different service providers. Thus, it is desirable to develop a decentralized solution of (2.8).

Constraints (2.6) and (2.7) are coupling constraints that make it difficult to obtain the desirable decentralized solution of (2.8) at each network. A decentralized solution can be developed using full dual decomposition of (2.8) [45] - [49]. We can re-rewrite constraint (2.7) in the following form

$$\sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms} \leq B_m^{\max}, \quad \forall m \in \mathcal{M}_{r2} \tag{2.9}$$

$$\sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms} \geq B_m^{\min}, \quad \forall m \in \mathcal{M}_{r2}. \tag{2.10}$$

In order to develop the decentralized solution, first we find the Lagrangian function for (2.8) using constraints (2.9) and (2.10), which can be expressed as

$$\begin{aligned} L(B, \lambda, \nu, \mu^{(1)}, \mu^{(2)}) &= \sum_{n=1}^{N}\sum_{s=1}^{S_n} U_{ns} + \sum_{n=1}^{N}\sum_{s=1}^{S_n} \lambda_{ns}(C_n - \sum_{m\in\mathcal{M}_{ns}} b_{nms}) \\ &+ \sum_{m\in\mathcal{M}_{r1}} \nu_m(B_m - \sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms}) + \sum_{m\in\mathcal{M}_{r2}} \mu_m^{(1)}(B_m^{\max} - \sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms}) \\ &+ \sum_{m\in\mathcal{M}_{r2}} \mu_m^{(2)}(\sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms} - B_m^{\min}) \end{aligned} \tag{2.11}$$

with $\lambda = (\lambda_{ns} : n \in \mathcal{N}, s \in \mathcal{S}_n)$ defined to be a matrix of Lagrange multipliers corresponding to capacity constraint (2.5), and $\lambda_{ns} \geq 0$, $\nu = (\nu_m : m \in \mathcal{M}_{r1})$, $\mu^{(1)} = (\mu_m^{(1)} : m \in \mathcal{M}_{r2})$, $\mu^{(2)} = (\mu_m^{(2)} : m \in \mathcal{M}_{r2})$ are vectors of Lagrange multipliers corresponding to the required bandwidth constraints (2.6), (2.9), and (2.10) respectively, and $\mu_m^{(1)}, \mu_m^{(2)} \geq 0$. The dual function is given by

$$H(\lambda, \nu, \mu^{(1)}, \mu^{(2)}) = \max_{B \geq 0} L(B, \lambda, \nu, \mu^{(1)}, \mu^{(2)}) \tag{2.12}$$

and the dual problem corresponding to the primal problem (2.8) is expressed by

$$\min_{(\lambda, \mu^{(1)}, \mu^{(2)}) \geq 0, \nu} H(\lambda, \nu, \mu^{(1)}, \mu^{(2)}). \tag{2.13}$$

A strong duality holds since the optimization problem (2.8) is a convex optimization problem, which makes the optimal values for the primal and dual problems equal [44]. The maximization problem (2.12) can be written as

$$\begin{aligned} H(\lambda, \nu, \mu^{(1)}, \mu^{(2)}) &= \sum_{n=1}^{N} \sum_{s=1}^{S_n} \max_{B \geq 0} \{U_{ns} - \lambda_{ns} \sum_{m \in \mathcal{M}_{ns}} b_{nms} \\ &- \sum_{m \in \mathcal{M}_{r1}} \nu_m b_{nms} - \sum_{m \in \mathcal{M}_{r2}} (\mu_m^{(1)} - \mu_m^{(2)}) b_{nms}\}. \end{aligned} \tag{2.14}$$

Then, each network BS/AP can solve its own network utility maximization (NUM) problem, given by

$$\max_{B \geq 0} \{U_{ns} - \lambda_{ns} \sum_{m \in \mathcal{M}_{ns}} b_{nms} - \sum_{m \in \mathcal{M}_{r1}} \nu_m b_{nms} - \sum_{m \in \mathcal{M}_{r2}} (\mu_m^{(1)} - \mu_m^{(2)}) b_{nms}\}. \tag{2.15}$$

By applying the Karush-Kuhn-Tucker (KKT) conditions on (2.15), each network BS/AP can find the bandwidth allocation, b_{nms}, for fixed values of $\lambda, \nu, \mu^{(1)}$, and $\mu^{(2)}$. Thus, we have

$$\frac{\partial U_{ns}}{\partial b_{nms}} - \lambda_{ns} - \nu_m - (\mu_m^{(1)} - \mu_m^{(2)}) = 0 \tag{2.16}$$

which results in

$$b_{nms} = [(\frac{\eta_1}{\lambda_{ns} + \nu_m + \eta_2(1 - \omega_{nms})} - 1)/\eta_1]^+, \quad \forall m \in \mathcal{M}_{r1} \tag{2.17}$$

$$b_{nms} = [(\frac{\eta_1}{\lambda_{ns} + (\mu_m^{(1)} - \mu_m^{(2)}) + \eta_2(1 - \omega_{nms})} - 1)/\eta_1]^+, \quad \forall m \in \mathcal{M}_{r2} \tag{2.18}$$

where the notion $[\cdot]^+$ is a projection on the positive quadrature to account for the fact that $B \geq 0$. By solving the dual problem (2.13), we can obtain the optimal values for the Lagrange multipliers that results in the optimal bandwidth allocation b_{nms} of (2.17) and (2.18). For a fixed bandwidth allocation B, the dual problem can be written as

$$\sum_{n=1}^{N}\sum_{s=1}^{S_n} \min_{\lambda \geq 0}\{\lambda_{ns}(C_n - \sum_{m \in \mathcal{M}_{ns}} b_{nms})\} + \sum_{m \in \mathcal{M}_{r1}} \min_{\nu}\{\nu_m(B_m - \sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms})\} + \sum_{m \in \mathcal{M}_{r2}} \min_{\mu^{(1)} \geq 0}\{\mu_m^{(1)}(B_m^{\max} - \sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms})\} + \sum_{m \in \mathcal{M}_{r2}} \min_{\mu^{(2)} \geq 0}\{\mu_m^{(2)}(\sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms} - B_m^{\min})\}. \tag{2.19}$$

For a differentiable dual function, a gradient descent method can be applied so as to find the optimal values for the Lagrangian multipliers [44], which is given by

$$\lambda_{ns}(j+1) = [\lambda_{ns}(j) - \alpha_1(C_n - \sum_{m \in \mathcal{M}_{ns}} b_{nms}(j))]^+ \tag{2.20}$$

$$\nu_m(j+1) = \nu_m(j) - \alpha_2(B_m - \sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms}(j)) \tag{2.21}$$

$$\mu_m^{(1)}(j+1) = [\mu_m^{(1)}(j) - \alpha_3(B_m^{\max} - \sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms}(j))]^+ \tag{2.22}$$

$$\mu_m^{(2)}(j+1) = [\mu_m^{(2)}(j) - \alpha_4(\sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms}(j) - B_m^{\min})]^+ \tag{2.23}$$

where j is an iteration index and α_e, $e \in \{1, 2, 3, 4\}$, is a fixed sufficiently small step size. As the gradient of (2.19) satisfies the Lipchitz continuity condition, the convergence of (2.20) - (2.23) towards the optimal solution is guaranteed [44]. Hence, the radio resource allocation b_{nms} of (2.17) and (2.18) converges to the optimal solution.

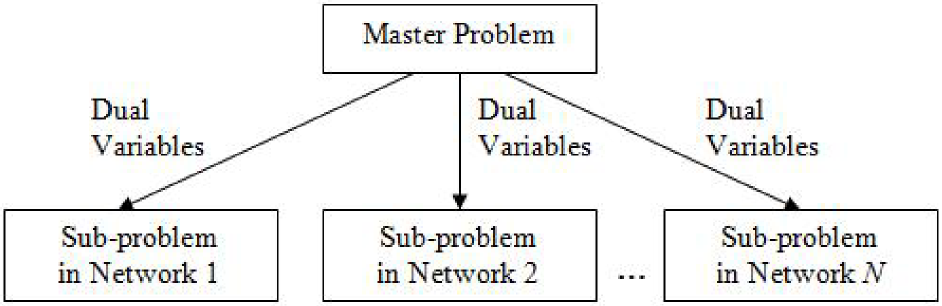


Figure 2.2: Decomposition of Problem (2.8).

2.4 A Decentralized Resource Allocation Mechanism

The decomposition approach for optimization problem (2.8) is defined in two levels. The first one is a lower level that is executed at each network, $n \in \mathcal{N}$, BS/AP, $s \in \mathcal{S}_n$, so as to find the optimal bandwidth allocation b_{nms} for each MT $m \in \mathcal{M}_{ns}$. This optimal bandwidth allocation is found by solving the sub-problems given in (2.15) by BSs/APs, which results in the solution of (2.17) for MTs with CBR service and (2.18) for MTs with VBR service. The other is a higher level, where the master problem is solved. The master problem is given in (2.19) and its optimal solution is obtained using the iterative method introduced in (2.20) - (2.23). The role of the master problem is to set the dual variables $\lambda, \nu, \mu^{(1)}$, and $\mu^{(2)}$ so as to coordinate the solution of the sub-problems at each network BS/AP. This is illustrated in Figure 2.2.

Following the classical interpretation of λ_{ns} in economics as the resource price [45], we refer to λ_{ns} as the link access price for network n BS/AP s. Basically, λ_{ns} serves as an indication of the capacity limitation experienced by network n link resources in BS/AP s. Hence, when the total call traffic load in network n BS/AP s ($\sum_{m \in \mathcal{M}_{ns}} b_{nms}$) reaches the capacity limitation (C_n), the link access price (λ_{ns}) increases to denote that it is expensive to use that link. The rest of the Lagrangian multipliers, namely ν_m which

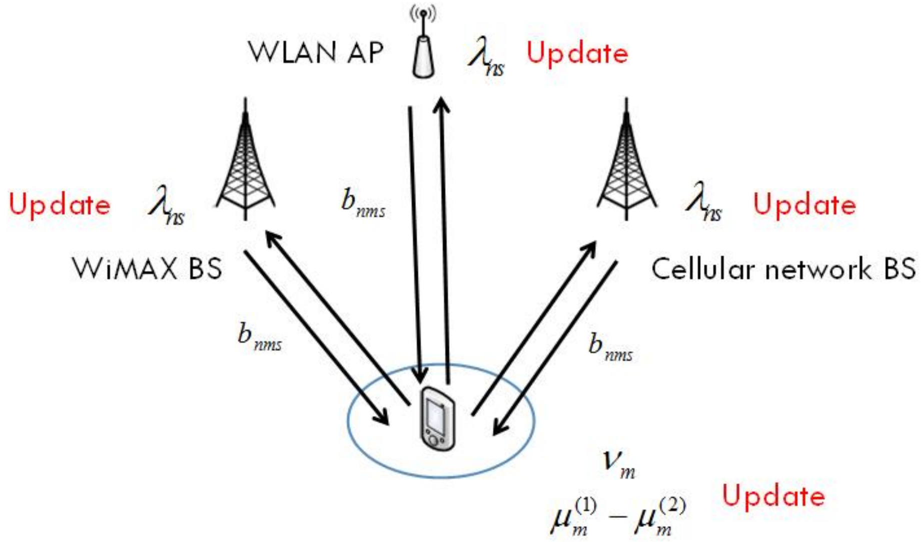


Figure 2.3: Decentralized bandwidth allocation.

is used by MTs with CBR service, and $\mu_m^{(1)}$ and $\mu_m^{(2)}$ which are used by MTs with VBR service, are coordination parameters. Hence, ν_m is used by MT m to coordinate the allocations by the available BSs/APs so as to ensure that the required bandwidth B_m is met. Similarly, $\mu_m^{(1)}$ and $\mu_m^{(2)}$ are used by MT m to coordinate the BS/AP resource allocations of different networks so as to ensure that the allocated resources lie within the specified required bandwidth range $[B_m^{\min}, B_m^{\max}]$.

The link access price λ_{ns} is calculated at each network BS/AP according to its capacity limitation and the BS/AP total call traffic load. The coordination parameter ν_m is calculated at each MT with CBR service, while the coordination parameters $\mu_m^{(1)}$ and $\mu_m^{(2)}$ are calculated by each MT with VBR service. All coordination parameters are calculated based on the allocated bandwidth from different wireless access networks and the MT total required bandwidth. The decentralized optimal radio resource allocation (DORA) mechanism can be explained using the scenario given in Figure 2.3. Consider an MT which lies in the coverage area of a WLAN AP and cellular network and WiMAX

BSs. Each BS/AP defines an initial feasible value for its link access price λ_{ns}. Similarly, the MT defines an initial feasible value for its coordination parameter(s). Each BS/AP performs its bandwidth allocation to the MT based on the network BS/AP link access price, the MT priority parameter, and its coordination parameter values. Each BS/AP then updates its link access price value based on its capacity limitation and its total call traffic load (due to the previous iteration resource allocation). Also, the MT updates its coordination parameter(s) (ν_m for MT with CBR service and $\mu_m^{(1)}$ and $\mu_m^{(2)}$ for MT with VBR service) based on the difference between its required bandwidth and the previous iteration total resource allocation. The updated coordination parameter for the new iteration (ν_m or the difference $\mu_m^{(1)} - \mu_m^{(2)}$) is broadcasted by the MT to the different available wireless access networks through the MT different radio interfaces so as to coordinate the resource allocation from different networks. As a result, each BS/AP can update its bandwidth allocation to the MT (using the updated link access price and coordination parameter values). The process continues over a number of iterations until the MT required bandwidth can be met eventually. The detailed DORA mechanism is given in Table 2.1, where ψ is a small tolerance.

For an incoming call, in case that no feasible resource allocation solution exists, the call will be blocked. Since a CAC procedure is in place, the call blocking probability is kept below a desired value. A blocked call may then try to ask for a lower service class (i.e. lower required bandwidth) for feasible resource allocation and admission.

2.5 Numerical Results and Discussion

This section presents numerical results for the radio resource allocation problem (2.8) using the DORA mechanism given in Table 2.1. We consider a simplified system model with a geographical region that is entirely covered by an IEEE 802.16e WiMAX BS and partially covered by a cellular network BS and an IEEE 802.11b WLAN AP [30],

Table 2.1: DORA Mechanism.

```
**Input:** $C_n \; \forall n \in \mathcal{N}$, $B_m \; \forall m \in \mathcal{M}_{r1}$, $[B_m^{\min}, B_m^{\max}] \; \forall m \in \mathcal{M}_{r2}$;
**Initialization**: $j \longleftarrow 1$; $\{\lambda_{ns}(1), \nu_m(1), \mu_m^{(1)}(1), \mu_m^{(2)}(1)\} \geq 0$, $b_{nms}(0) = \{\}$, $y = 0$;
**while** $y = 0$ **do**
    **for** $n \in \mathcal{N}$ **do**    // Bandwidth Allocation at Each Network BS/AP
        **for** $m \in \mathcal{M}$ **do**
            **for** $s \in \mathcal{S}_n$ **do**
                **if** $m \in \mathcal{M}_{ns}$ **then**
                    $b_{nms}(j) = [(\frac{\eta_1}{\lambda_{ns}(j)+\nu_m(j)+\eta_2(1-\omega_{nms})} - 1)/\eta_1]^+, \quad m \in \mathcal{M}_{r1};$
                    $b_{nms}(j) = [(\frac{\eta_1}{\lambda_{ns}(j)+(\mu_m^{(1)}(j)-\mu_m^{(2)}(j))+\eta_2(1-\omega_{nms})} - 1)/\eta_1]^+, \; m \in \mathcal{M}_{r2};$
                **end if**
            **end for**
        **end for**
    **end for**
    **if** $|b_{nms}(j) - b_{nms}(j-1)| > \psi$ **then**
        **for** $n \in \mathcal{N}$ **do**    // Update of Link Access Price at Each Network BS/AP
            **for** $s \in \mathcal{S}_n$ **do**
                $\lambda_{ns}(j+1) = [\lambda_{ns}(j) - \alpha_1(C_n - \sum_{m \in \mathcal{M}_{ns}} b_{nms}(j))]^+;$
            **end for**
        **end for**
        **for** $m \in \mathcal{M}$ **do**    // Update of Coordination Parameters at Each MT
            $\nu_m(j+1) = \nu_m(j) - \alpha_2(B_m - \sum_{n=1}^{N} \sum_{s=1}^{S_n} b_{nms}(j)), \quad \forall m \in \mathcal{M}_{r1}$
            $\mu_m^{(1)}(j+1) = [\mu_m^{(1)}(j) - \alpha_3(B_m^{\max} - \sum_{n=1}^{N} \sum_{s=1}^{S_n} b_{nms}(j))]^+, \quad \forall m \in \mathcal{M}_{r2};$
            $\mu_m^{(2)}(j+1) = [\mu_m^{(2)}(j) - \alpha_4(\sum_{n=1}^{N} \sum_{s=1}^{S_n} b_{nms}(j) - B_m^{\min})]^+, \quad \forall m \in \mathcal{M}_{r2};$
        **end for**
        $j \longleftarrow j + 1$
    **else**
        $y = 1;$
    **end if**
**end while**
```

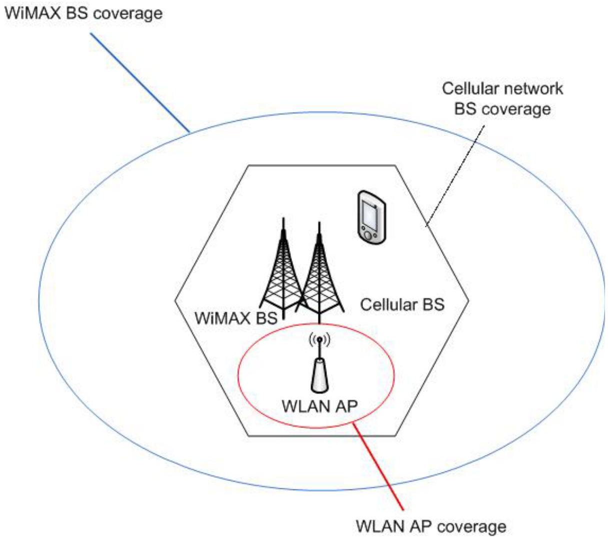


Figure 2.4: Service areas under consideration.

as shown in Figure 2.4. Thus, $\mathcal{N} = \{1, 2, 3\}$, with the WiMAX, cellular network, and WLAN indexed as $1, 2,$ and 3, respectively. Each network has only one BS/AP in the geographical region, i.e. $S_n = \{1\}$, $\forall n \in \mathcal{N}$. As a result, the geographical region is described by three service areas, $\mathcal{K} = \{1, 2, 3\}$. In service area 1, only the WiMAX BS coverage is available. In service area 2, both the WiMAX and cellular network coverages are available. In service area 3, all three networks are available. The transmission capacities of the three networks are given by $C_1 = 20$ Mbps, $C_2 = 2$ Mbps, and $C_3 = 11$ Mbps.

For the priority mechanism, different networks can set different costs on their resources through the priority parameter ω_{nms}. As the cellular network has the lowest transmission capacity among all the available networks, it sets the highest cost on its resources. Both the WiMAX and WLAN have a high transmission capacity, however, the WiMAX covers a larger area with more users. Hence, the WiMAX sets a higher cost on its resources than the WLAN with its limited coverage area. So, for network users we set $\omega_{1m1} = 0.6$, $\omega_{2m1} = 0.5$, and $\omega_{3m1} = 0.8$.

Table 2.2: Number of subscribers of different networks in different service areas

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
M_{111}	10	M_{122}	7	M_{221}	8	M_{232}	5
M_{112}	10	M_{131}	5	M_{222}	8	M_{332}	5
M_{121}	7	M_{132}	5	M_{231}	5	M_{331}	Variable

Let the required bandwidth allocation be 256 Kbps for an MT with CBR service, while for an MT with VBR service the required bandwidth allocation lies in the range $[256, 512]$ Kbps. Let the number of subscribers for network n in service area k with service r be M_{nkr} with $r = 1$ for CBR service and $r = 2$ for VBR service. We vary the number of WLAN subscribers with CBR calls in service area 3 (M_{331}) and fix all other parameters to study the system performance. The number of different network subscribers in all service areas are given in Table 2.2.

Figures 2.5 - 2.8 depict various bandwidth allocation results versus the number of ongoing CBR calls for the WLAN subscribers in service area 3 (M_{331}).

Figure 2.5 shows the total allocated bandwidth by each network BS/AP. Both the WiMAX and cellular network BSs reach their capacity limitation, independent of M_{331}. On the other hand, the WLAN AP increases its total allocated bandwidth with M_{331} so as to accommodate more subscribers. The WLAN AP reaches its capacity limitation at $M_{331} = 14$.

In the following results, we study the total allocated bandwidth from each network BS/AP to subscribers of different networks in all three service areas.

Figure 2.6a shows the total allocated bandwidth by each network BS/AP for the CBR WLAN subscribers in service area 3. Because of the priority mechanism, the WLAN AP supports its own subscribers with all their required bandwidth in order to avoid the

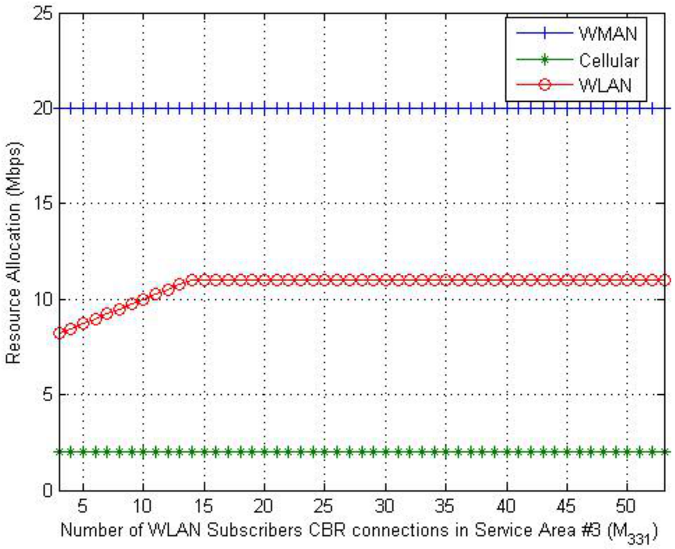
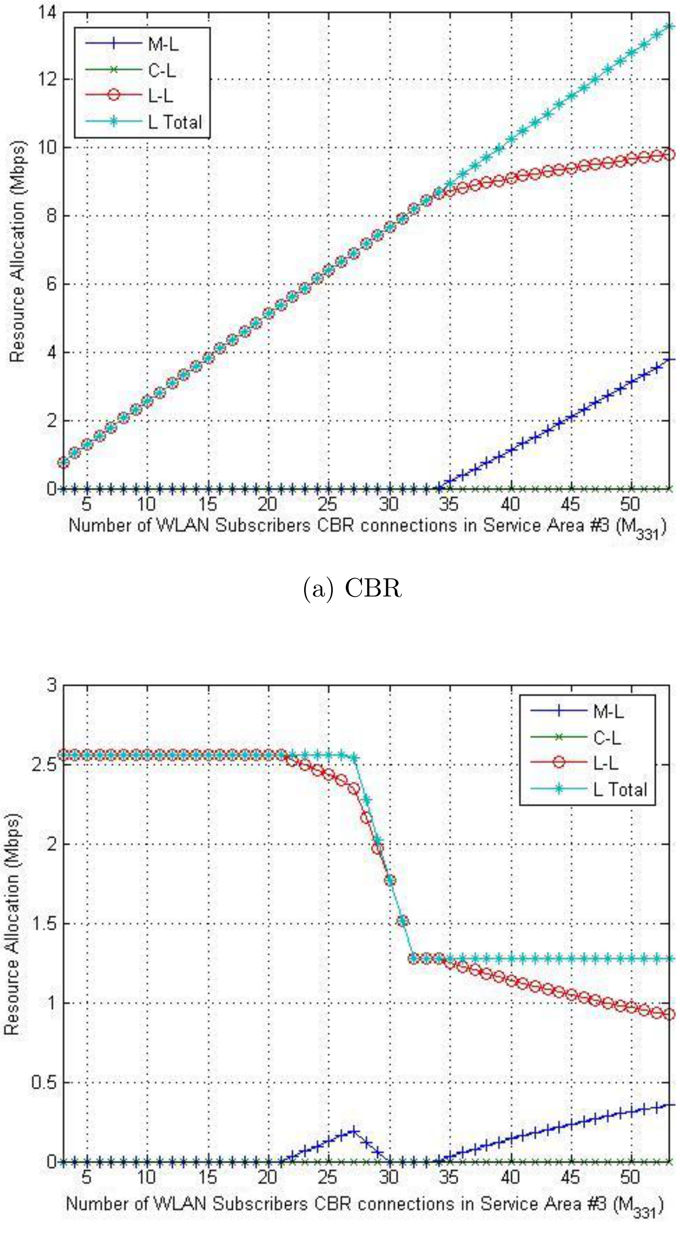


Figure 2.5: Total bandwidth allocation by each network BS/AP.

associated high cost of the BS resources of WiMAX and cellular network. Hence, The bandwidth allocation for the WLAN subscribers from the WiMAX (M-L) and cellular network (C-L) BSs is equal to zero, while the WLAN AP allocated bandwidth (L-L) increases with M_{331} so as to accommodate more subscribers. For $M_{331} > 34$, there is no sufficient resources at the WLAN AP to support individually its own subscribers. Hence, the WiMAX BS increases its bandwidth allocation to support the WLAN subscribers. The support comes only from the WiMAX BS as it sets a lower cost on its resources than the cellular network BS.

Figure 2.6b shows the allocated bandwidth by each network BS/AP for the VBR WLAN subscribers in service area 3. For $M_{331} \geq 22$, the WLAN AP decreases its allocated bandwidth to the VBR subscribers (L-L) in order to support the increasing number of the CBR subscribers. This is compensated by an increase in the bandwidth allocation from the WiMAX BS (M-L) in order to keep the total bandwidth allocation constant at the call maximum required bandwidth (512 Kbps for each VBR call). For



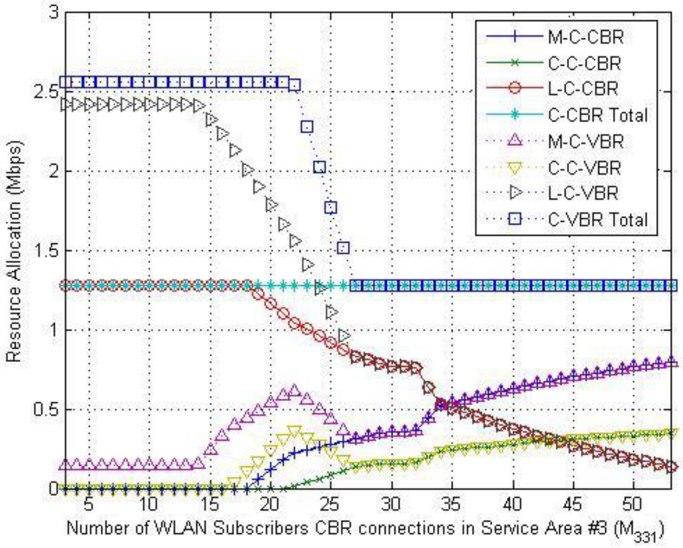
(a) CBR

(b) VBR

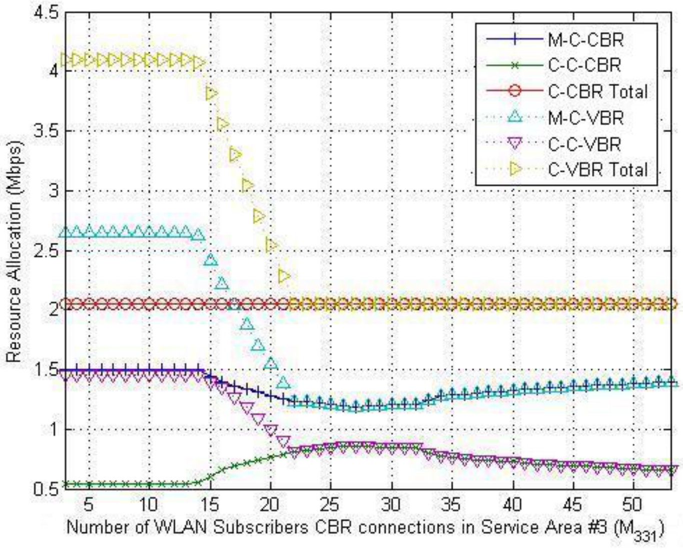
Figure 2.6: Total bandwidth allocation by each network BS/AP to (a) CBR and (b) VBR WLAN subscribers.

$M_{331} > 27$, any further increase in the bandwidth allocation from the WiMAX BS to the WLAN subscribers would degrade the WiMAX BS bandwidth allocation to its own VBR subscribers. This is not allowed, however, by the priority mechanism as it gives higher priority on the WiMAX BS resources to the WiMAX subscribers. Hence, the WiMAX BS decreases its allocated bandwidth to the VBR WLAN subscribers which reduces the VBR call total bandwidth allocation towards the call minimum required bandwidth. For $M_{331} > 34$, the WLAN AP decreases its bandwidth allocation to its VBR subscribers in order to support the increasing number of its CBR subscribers. Hence, the WiMAX BS increases its bandwidth allocation to the WLAN VBR subscribers so as not to violate their minimum required bandwidth (256 Kbps for each VBR call).

Figure 2.7a shows the total allocated bandwidth by each network BS/AP to the cellular network subscribers, with CBR and VBR calls, in service area 3. The total allocated bandwidth of CBR cellular network subscribers (C-CBR Total) comes from the WLAN AP (L-C-CBR). The allocated bandwidth from the cellular network BS (C-C-CBR) is zero, as it uses its bandwidth to support its own subscribers in service area 2 (which is covered only by the cellular network BS, and the WiMAX BS with a higher cost for bandwidth). As for the WiMAX BS zero bandwidth allocation (M-C-CBR), it is due to the higher cost that the WiMAX BS sets on its resources as compared to the WLAN AP. For $M_{331} > 18$, the WLAN AP decreases its bandwidth allocation to the CBR cellualr network subscribers in order to support its increasing number of subscribers (M_{331}). Hence, the WiMAX BS increases its allocation to the CBR cellular network subscribers in order to keep the total bandwidth allocation constant at the required bandwidth (256 Kbps for each CBR call). For $M_{331} > 21$, more allocated bandwidth is required from the WiMAX BS to keep the CBR cellular network subscriber total allocation constant; however, this would increase the associated cost due to the WiMAX BS low priority parameter for the network users. Hence, the cellular network



(a) Area 3



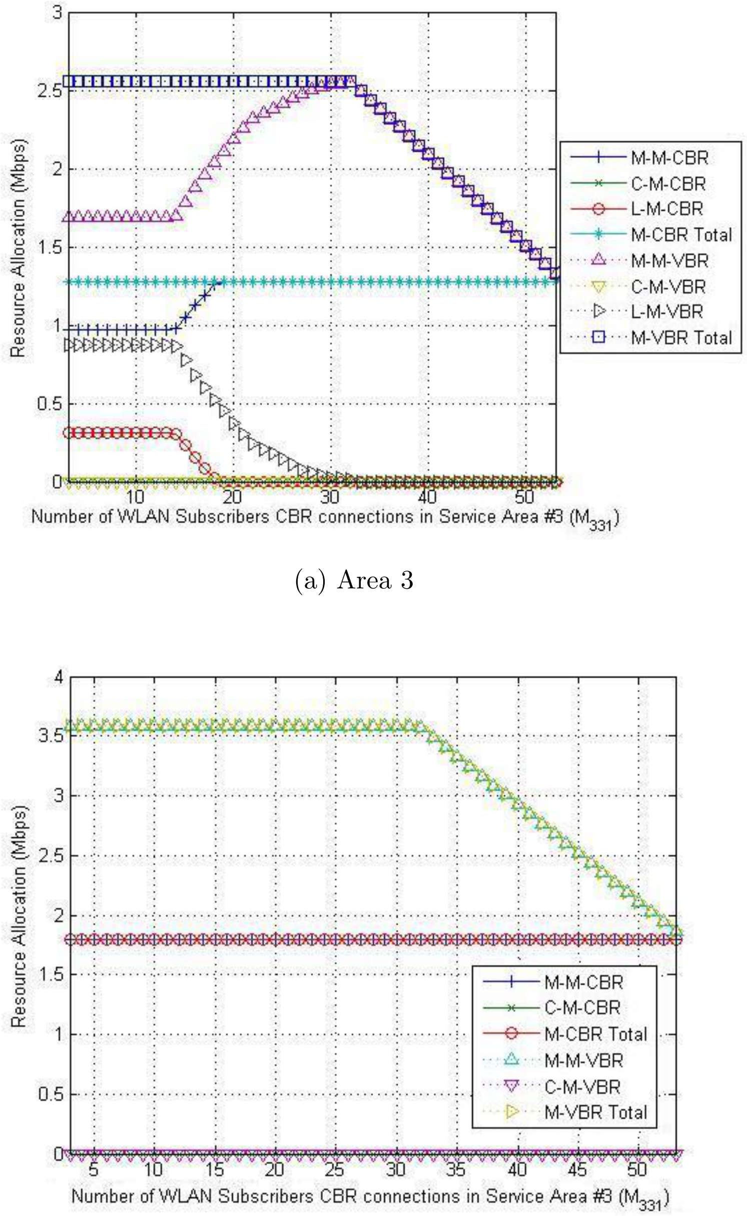
(b) Area 2

Figure 2.7: Total bandwidth allocation by each network BS/AP to the cellular network subscribers in (a) Area 3 and (b) Area 2.

BS increases its allocated bandwidth to support its own CBR subscribers. As shown in the figure, the total bandwidth allocation is always constant at the call required bandwidth. For the VBR subscribers, the WLAN AP decreases its bandwidth allocation to the VBR cellular network subscribers with M_{331} in order to support its own subscribers. This is compensated for by an increase in the WiMAX BS bandwidth allocation to keep the total allocated bandwidth (C-VBR-Total) at its maximum required bandwidth (512 Kbps for each VBR call). For $M_{331} > 17$, the cellular network BS increases its bandwidth allocation to its VBR subscribers in order to reduce the amount of required bandwidth from the WiMAX BS due to the associated high cost. For $M_{331} > 22$, any further increase in the allocated bandwidth from the WiMAX BS to the VBR cellular network subscribers would reduce the WiMAX BS allocation to its own VBR subscribers. Hence, the WiMAX BS decreases its allocated bandwidth to the VBR cellular network subscribers. Also, the cellular network BS decreases its allocated bandwidth to its VBR subscribers to support its CBR subscribers in this area. As a result, the total allocated bandwidth to the VBR cellular network subscribers starts to decrease towards the minimum required bandwidth. For $M_{331} > 26$, the WiMAX and cellular network BSs increase their bandwidth allocation to the VBR cellular network subscribers in order to compensate for the reduction in the allocated bandwidth from the WLAN AP and keep the total bandwidth allocation constant at the call minimum required bandwidth.

Figure 2.7b shows the total allocated bandwidth by each network BS/AP to the cellular network subscribers in service area 2. The allocated bandwidth comes only from the WiMAX and cellular network BSs since the MTs are out of the coverage area of the WLAN AP. For the CBR subscribers with $M_{331} > 14$, the WiMAX BS reduces its allocated bandwidth to the CBR cellular network subscribers to support its own subscribers with their maximum required bandwidth. As a result, the cellular network BS increases its allocated bandwidth. For $M_{331} > 32$, the cellular network BS reduces its bandwidth allocation to support its subscribers in area 3 (refer to Figure 2.7a). This is compensated for by an increase in the WiMAX BS allocated bandwidth to the CBR cellular network subscribers. In all the cases, the total bandwidth allocation (C-CBR Total) is constant at the required bandwidth (256 kbps for each CBR user). For the VBR subscribers with $M_{331} > 14$, the cellular network BS cannot further keep its VBR subscribers in area 2 at their maximum required bandwidth, and has to decrease its allocated bandwidth to support the CBR cellular network subscribers in this area. Also, the WiMAX BS has to decrease its bandwidth allocation to satisfy its own VBR subscribers with their maximum required bandwidth. Therefore, the total bandwidth allocation (C-VBR Total) starts to decrease towards the minimum required bandwidth. As in the CBR bandwidth allocation, for $M_{331} > 32$, the cellular network BS reduces its allocated bandwidth to its VBR subscribers in area 2 to support its subscribers in area 3. As a result, the WiMAX BS increases its bandwidth allocation to keep the total allocated bandwidth constant at the minimum required bandwidth.

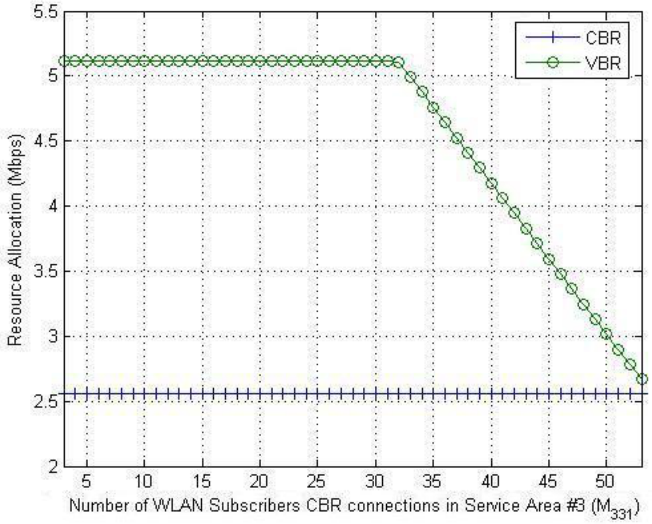
Figure 2.8a shows the total allocated bandwidth by each network BS/AP to the WiMAX subscribers in service area 3. For both CBR and VBR calls, most of the allocated bandwidth comes from the WiMAX BS (M-M-CBR and M-M-VBR), so as to reduce the associated cost of the WLAN bandwidth allocation. The allocated bandwidth from the cellular network BS (C-M-CBR and C-M-VBR) is zero, as it allocates radio resources to its own subscribers in service areas 2 and 3. For $M_{331} > 13$, the WLAN AP decreases its allocated bandwidth to the VBR WiMAX subscribers in order to support its own subscribers. Hence, the WiMAX BS increases its allocated bandwidth to support its own subscribers. For $M_{331} > 18$, all the required bandwidth to support the CBR calls (M-CBR-Total) in service area 3 comes from the WiMAX BS. For $M_{331} > 32$, the WiMAX BS reduces its bandwidth allocation to the VBR WiMAX subscribers towards the minimum required bandwidth to support the WLAN subscribers (refer to Figure 2.6).



(a) Area 3

(b) Area 2

Figure 2.8: Total bandwidth allocation by each network BS/AP to the WiMAX subscribers in (a) Area 3, (b) Area 2, and (c) Area 1.



(c) Area 1

Figure 2.8: Cont. Total bandwidth allocation by each network BS/AP to the WiMAX subscribers in (a) Area 3, (b) Area 2, and (c) Area 1.

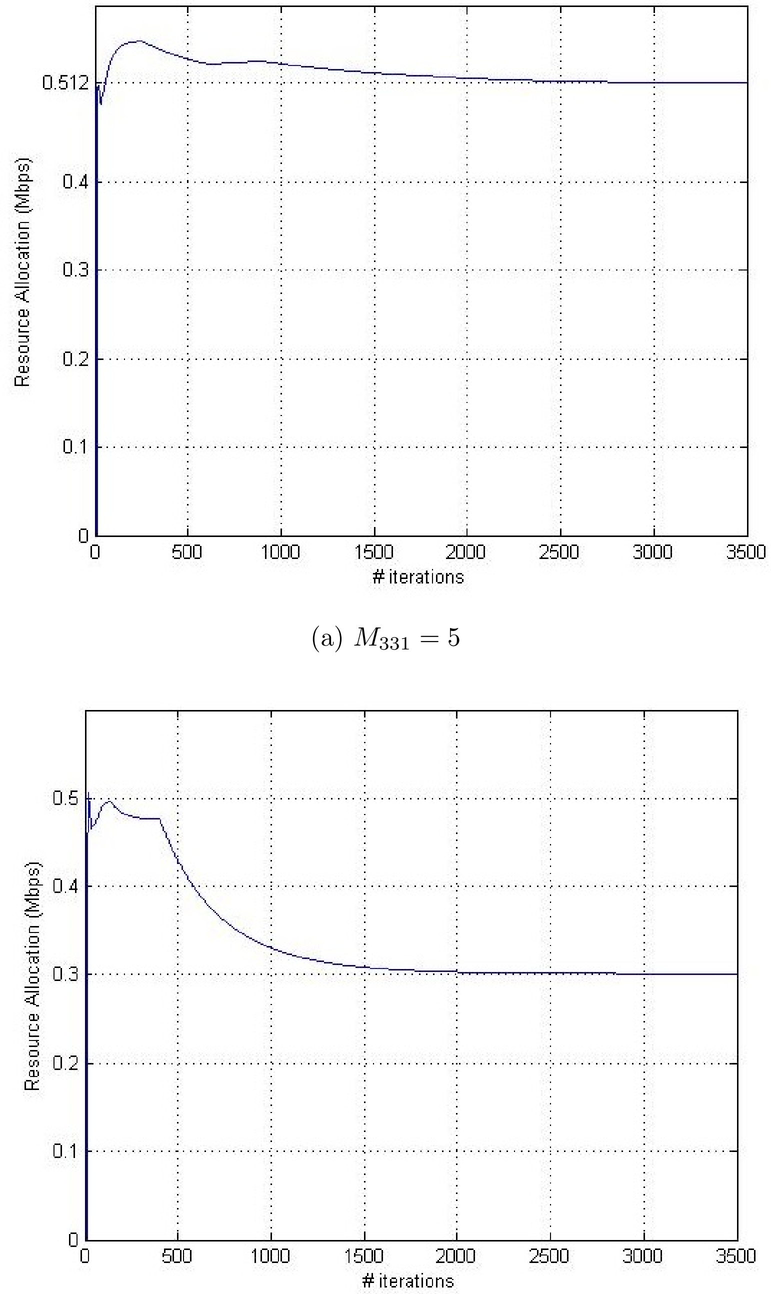
Figure 2.8b shows the total allocated bandwidth by each network BS/AP to the WiMAX subscribers in service area 2. The total allocated bandwidth comes only from the WiMAX BS (M-M-CBR and M-M-VBR) although the MTs lie in the coverage area of the cellular network. This is due to the associated high cost of the cellular network bandwidth. Again, as in Figure 2.8a, for $M_{331} > 32$, the WiMAX BS decreases its allocated bandwidth to the VBR subscribers to support the WLAN subscribers in service area 3.

Figure 2.8c shows the total allocated bandwidth by each network BS/AP to the WiMAX subscribers in service area 1. Since the MTs are outside the coverage areas of the cellular network BS and WLAN AP, the total bandwidth allocation comes only from the WiMAX BS. For $M_{331} > 32$, the WiMAX BS allocated bandwidth to the VBR calls is reduced to support the WLAN subscribers in area 3.

From the results in Figures 2.6 - 2.8, service degradation of VBR calls starts from the cellular network subscribers as these users depend heavily on other networks in order to satisfy their required bandwidth. Because of the priority mechanism, these networks give higher priority in allocating their resources to their own subscribers, leading to a reduced bandwidth allocated to the VBR calls of cellular network subscribers.

The rate of convergence for a WiMAX subscriber in service area 1 towards its total required bandwidth is shown in Figures 2.9a and 2.9b for $M_{331} = 5$ and $M_{331} = 50$, respectively. With a small number of WLAN subscribers ($M_{331} = 5$), the WiMAX subscriber converges to the maximum required bandwidth (512 Kbps) after a number of iterations. As the number of WLAN subscribers increases ($M_{331} = 50$), resource allocation to the WiMAX subscriber converges to a lower bandwidth within the required range ($[256, 512]$ Kbps).

Figure 2.10 shows the variation in the link access price (λ_{ns}). For $M_{331} < 14$, the WLAN AP has not yet reached its capacity limitation, resulting in its link access price value equals to zero. On the other hand, the WiMAX and the cellular network BSs have a high value of link access price as they reach their capacity limitation (refer to Figure 2.5). The cellular network BS has the highest link access price value due to its lowest capacity. For $M_{331} \geq 14$, the BSs/AP of three networks reach their capacity limitation. This calls for a higher link access prices for all three networks. As M_{331} increases, the link access price value increases to indicate that it is more expensive to use these links. These results follow the complementary slackness condition [44]. Normally, the WLAN AP has a lower link access price than the WiMAX BS, since the number of users supported by the WLAN AP is less than those supported by the WiMAX BS in the three areas. But as the WLAN AP gives a lower cost on its resources using the priority parameter ω_{3m1}, most of the users in area 3 use its bandwidth, and the WLAN subscribers in area 3 are mainly supported by the WLAN AP, which causes the link access price for the WLAN



(a) $M_{331} = 5$

(b) $M_{331} = 50$

Figure 2.9: Rate of convergence towards the required bandwidth for a WiMAX subscriber in service area 1 with (a) $M_{331} = 5$ and (b) $M_{331} = 50$.

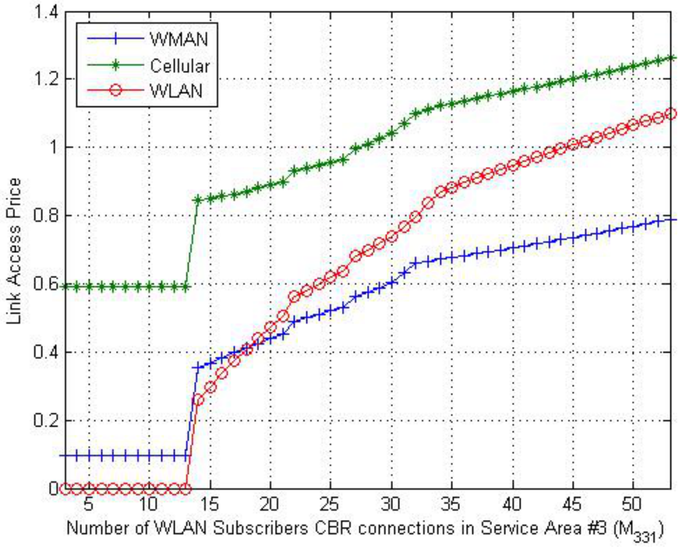


Figure 2.10: Link access price.

AP to increase above the link access price value of the WiMAX BS for $M_{331} > 18$.

2.6 Summary

In this chapter, a decentralized optimal resource allocation (DORA) mechanism is presented in a heterogeneous wireless access environment. The mechanism has the following features:

1. It is a decentralized mechanism. Each network BS/AP solves its own NUM problem and performs its resource allocation. No central resource manager is required;

2. It supports MTs with multi-homing capabilities for multi-services, namely, CBR and VBR services;

3. It allows for service differentiation, among the network subscribers and the other

users. As a result, the network subscribers enjoy their maximum required bandwidth using their home network resources;

4. The MTs play an active role in the resource allocation operation by coordinating the available wireless access networks to satisfy their required bandwidth.

The DORA mechanism is limited to a static system with no arrival of new calls or departure of existing ones with the objective of identifying the role of different network entities in such a decentralized architecture model. In the next chapter, we discuss the main limitations of the DORA mechanism in a dynamic system with call arrivals and departures and present some modifications to address these limitations.

Chapter 3

Dynamic Cooperative Bandwidth Allocation

In a dynamic environment, call arrivals and departures in different service areas may trigger resource reallocations for all MTs in service. In a decentralized architecture, this is translated to a heavy signalling overhead between the MTs and different BSs/APs with every call arrival and/or departure in any service area. Hence, the main challenge is how to develop an efficient decentralized bandwidth allocation mechanism that reduces the associated signalling overhead with call arrivals and departures. In this chapter, concepts of call traffic load prediction and network cooperation are introduced to address the challenges that face the decentralized bandwidth allocation in a dynamic environment.

3.1 Introduction

In Chapter 2, the DORA mechanism is presented to support MTs with multi-homing capabilities in a heterogeneous wireless access medium. The DORA mechanism mainly identifies the role of different entities in the heterogeneous wireless access medium in order to enable a decentralized architecture. Specifically, the main role of a network, $n \in \mathcal{N}$, BS/AP, $s \in \mathcal{S}_n$, in the decentralized architecture is to update a link access price value (λ_{ns}) that indicates the capacity limitation experienced by this BS/AP. On the other hand, the main role of an MT, m, is to update its coordination parameter(s) (ν_m for MT with CBR service, or $\mu_m^{(1)} - \mu_m^{(2)}$ for MT with VBR service) in order to satisfy its required bandwidth. Both link access price values for different BSs/APs and coordination parameter(s), together with the priority parameter ω_{nms}, determine the allocated resources from each network BS/AP so as to satisfy the MT total required bandwidth. The DORA mechanism is an iterative one that relies on signalling exchange between an MT and different BSs/APs in order to reach the optimal resource allocation from each BS/AP to the MT. This includes the exchange of the current iteration MT coordination parameter (from MT to BSs/APs) and the corresponding BS/AP resource allocation b_{nms} (from each BS/AP to MT). The DORA mechanism is proposed for a static environment without arrival of new calls or departure of existing ones.

As illustrated in Figure 2.10, due to the complimentary slackness condition for (2.5), we have the following observations:

1. When the total call traffic load ($\sum_{m \in \mathcal{M}_{ns}} b_{nms}$) carried by network n BS/AP s is less than the BS/AP transmission capacity limitation C_n, the corresponding optimal link access price value $\lambda_{ns}^* = 0$. This results in allocating the maximum required bandwidth for all VBR calls under this BS/AP jurisdiction;

2. When the carried call traffic load reaches the BS/AP transmission capacity limitation, $\lambda_{ns}^* > 0$. Hence, the allocated bandwidth to each of the VBR calls in service is reduced towards the call minimum required bandwidth so as to support new incoming calls.

In a dynamic system, with call arrivals to and departures from different service areas, the carried call traffic load by each BS/AP fluctuates over time. This in turn results in a fluctuating (time-varying) optimal value for the link access price λ^*_{ns} and hence a fluctuating bandwidth allocation matrix B^*, with every call arrival and/or departure. This results in the following limitations for the DORA mechanism [15]:

1. A fluctuating link access price λ_{ns} triggers bandwidth reallocations to the existing calls. Let J denote the number of iterations required by the DORA mechanism to reach the optimal bandwidth allocation. In the decentralized architecture, information exchange between MTs and BSs/APs for coordination parameter updates is required for the J iterations in order to support an optimal bandwidth reallocation. This signalling exchange should take place, for all MTs in service, with every call arrival to and/or departure from any service area k. In general, the signalling overhead is a function of the call arrival and departure rates, the numbers of existing calls in different service areas, and the number of iterations required for the mechanism to converge to the optimal bandwidth allocation. Hence, excessive signalling overhead is needed for information exchange between existing MTs and different BSs/APs, which makes the DORA mechanism too expensive to implement in a dynamic system;

2. Due to the random nature of call arrivals and departures in different service areas, it is possible that an arrival and/or departure event occurs during the J iterations. Hence, the DORA mechanism may not converge in practice to an optimal bandwidth allocation;

3. The signalling information exchange for the J iterations between an MT and different BSs/APs takes place on both up and down links. Let σ denote the total time duration of the signalling exchange for the J iterations. When there exists a

network with contention based medium access control among the available wireless networks, it is expected that σ increases with the call arrival rates as more MTs will be involved in the signalling procedure. Hence, the DORA mechanism can lead to high handoff latency, which is not desirable for seamless service provision.

In this chapter, we aim to extend the DORA mechanism so as to account for the system dynamics in terms of call arrivals and departures, and hence to perform an efficient bandwidth allocation. We set two objectives for this bandwidth allocation: 1) To significantly reduce the required resource reallocations to existing calls and the associated signalling overhead over the air interface in the decentralized network architecture, with call arrivals to and departures from different service areas; and 2) To achieve an acceptable call blocking probability and a sufficient amount of allocated bandwidth per VBR call. These objectives are achieved through a prediction based resource allocation (PBRA), that is presented in this chapter and relies on concepts of call traffic load prediction, network cooperation, convex optimization, and decomposition theory.

3.2 Related Work

In addition to the bandwidth allocation mechanisms presented from literature in Chapter 2, we review in this section some existing CAC mechanisms for a heterogeneous wireless access medium, as we target a dynamic system model [50]. The CAC mechanisms can be classified, based on the admission domain, to either a single-network or a multi-homing admission.

Single-network CAC mechanisms are studied in [42] and [51] - [53]. In [42], the CAC mechanism provides service differentiation among voice and data services in an integrated cellular/WLAN network using restricted access principle. The CAC mechanism in [51] aims to provide service differentiation among multiple service classes using virtual partitioning with pre-emption and prioritize handoff connections (horizontal and vertical) over new connection requests in an integrated cellular/WLAN network. In [52], the CAC mechanism is to improve the connection level performance of the system by prioritizing handoff connection requests (horizontal and vertical) over new connection requests, using the cut-off priority and fractional guard channel schemes [54, 55] in an integrated cellular/WLAN network. The CAC mechanisms in [42, 51, 52] are based on a central resource manager. The central controller decides whether to accept or reject the call. It selects one of the available networks based on a predefined criterion and decides whether to admit or reject the requested call in this network. If the call request is rejected in the selected network, another network is selected, and so on. In [53], a decentralized CAC mechanism is developed in an integrated cellular/WLAN network for voice and data services.

Multi-homing CAC mechanisms are studied in [31] - [34], based on a central resource manager. The CAC mechanisms in [31, 33, 34] aim to ensure that the amount of bandwidth allocated from all networks to an incoming connection satisfies the corresponding user requirement. In [32], the CAC mechanism ensures that the amount of bandwidth allocated from all networks satisfies an incoming connection requirement and prioritizes handoff calls (horizontal and vertical) over new calls.

In this chapter, a decentralized mechanism for bandwidth allocation and CAC in a heterogeneous wireless access medium for MTs with multi-homing capabilities is proposed. The MT determines the required bandwidth allocation from each network in order to satisfy its total required bandwidth. If the networks provide the MT with its required bandwidth, the call is admitted, else the call is rejected.

3.3 System Model

In this section, we present the modifications to the system model presented in Chapter 2 to account for the system dynamics in terms of users' mobility and call traffic models.

A. Wireless Access Networks

Consider the geographical region given in Figure 2.1. Let $\mathcal{N}_k$ denote the set of networks available at service area k, and $\mathcal{S}_{nk}$ denotes the set of BSs/APs from network n covering service area k. The subset $\mathcal{S}_k$ denotes the BSs/APs from all networks covering service area k, with cardinality $|\mathcal{S}_k|$. An identification (ID) beacon is broadcasted by each BS/AP, which is used in the MT attachment procedure [56]. It is assumed that different networks are connected through a backbone to exchange their signalling information.

B. Call Traffic Models

In this chapter and the next one, we only focus on VBR video calls as they are more challenging to support in a dynamic system due to the requirement of providing the call with a bandwidth allocation that is as close as possible to the maximum required bandwidth. The extension of the proposed bandwidth allocation mechanism (PBRA) is straight forward, to include CBR calls.

There exists a set $\mathcal{L}$ of service classes, $\mathcal{L} = \{1, 2, \ldots, L\}$. Each service class, $l \in \mathcal{L}$, has unique $B_l^{\min}$ and $B_l^{\max}$ values. For subscribers of a given network n, let M_{lk}^n denote the number of existing calls of service class l in service area k. It is assumed that there exists sufficient capacity through the available BSs/APs in the geographical region to satisfy a target call blocking probability for each service class l in each service area k for network n subscribers. Let C_{lk}^n denote the maximum number of calls of each service class l which can be supported in each service area k for network n subscribers, given the transmission capacities of available BSs/APs. A capacity analysis similar to the one in [42] can be used to determine this maximum number of calls.

A Poisson process is used to model video call arrivals, which is a widely adopted assumption [42]. In particular, a Poisson process with parameter υ_{lk} is used to model the arrival process of both new and handoff video calls from service class l to service area k. Following the statistics of on-demand video streaming [57, 58], the video call duration is very likely to be heavy-tailed. The 'mice-elephants' phenomenon is a very important feature of heavy-taildness [59]. This implies that, with respect to the video call duration, most video calls have quite short duration, while a small fraction of video calls have an extremely large duration. Yet, performance analysis is complex with heavy-tailed distributions. Hence, it is proposed in [60], for effective analysis, to fit a large class of heavy-tailed distributions with hyper-exponential distributions. For simplicity, a two-stage hyper-exponential distribution is used to model the video call duration. Thus, a video call of MT m that belongs to class l has a call duration with a mean $\bar{T}_c^l$ and a probability density function (PDF), $f_{T_c^l}(t)$, which is given by

$$f_{T_c^l}(t) = \frac{\varsigma_l}{\varsigma_l + 1} \cdot \frac{\varsigma_l}{\bar{T}_c^l} \cdot e^{-\frac{\varsigma_l}{\bar{T}_c^l} t} + \frac{1}{\varsigma_l + 1} \cdot \frac{1}{\varsigma_l \bar{T}_c^l} \cdot e^{-\frac{1}{\varsigma_l \bar{T}_c^l} t}, \quad \varsigma_l \geq 1, t \geq 0. \tag{3.1}$$

The parameter ς_l in (3.1) can characterize the mice-elephant feature. A large fraction of calls $\frac{\varsigma_l}{\varsigma_l+1}$ has a duration with mean time $\frac{\bar{T}_c^l}{\varsigma_l}$, while the other fraction $\frac{1}{\varsigma_l+1}$ has a duration with mean time $\varsigma_l \bar{T}_c^l$.

D. Mobility Models and Channel Holding Time

User residence time is used to characterize the user mobility within a given service area $k \in \mathcal{K}$, and is assumed to follow an exponential distribution. The PDF of the user residence time T_r^k, with mean $\bar{T}_r^k$, in service area $k \in \mathcal{K}$, is given by

$$f_{T_r^k}(t) = \frac{1}{\bar{T}_r^k} e^{-\frac{t}{\bar{T}_r^k}}, \quad t \geq 0. \tag{3.2}$$

In a given service area $k \in \mathcal{K}$, the channel holding time is given by $T_h^{lk} = \min(T_c^l, T_r^k)$, where T_c^l and T_r^k are independent of each other. Then,

$$\Pr\{\min(T_c^l, T_r^k) > t\} = \Pr\{T_c^l > t, T_r^k > t\} = \Pr\{T_c^l > t\} \cdot \Pr\{T_r^k > t\} \tag{3.3}$$

where $\Pr\{\cdot\}$ is the probability of occurence of event $\{\cdot\}$. This results in a channel holding time with a PDF given by

$$f_{T_h^{lk}}(t) = f_{T_c^l}(t)[1 - F_{T_r^k}(t)] + f_{T_r^k}(t)[1 - F_{T_c^l}(t)], \quad t \geq 0 \tag{3.4}$$

where $F_{T_c^l}(t)$ and $F_{T_r^k}(t)$ are the cumulative distribution functions (CDFs) for the call duration and user residence time, respectively. From (3.1) and (3.2), we have

$$\begin{aligned} f_{T_h^{lk}}(t) &= \frac{\varsigma_l}{\varsigma_l + 1} \cdot (\frac{1}{\bar{T}_r^k} + \frac{\varsigma_l}{\bar{T}_c^l}) \cdot e^{-(\frac{1}{\bar{T}_r^k} + \frac{\varsigma_l}{\bar{T}_c^l})t} \\ &+ \frac{1}{\varsigma_l + 1} \cdot (\frac{1}{\bar{T}_r^k} + \frac{1}{\varsigma_l \bar{T}_c^l}) \cdot e^{-(\frac{1}{\bar{T}_r^k} + \frac{1}{\varsigma_l \bar{T}_c^l})t}, \qquad t \geq 0. \end{aligned} \tag{3.5}$$

3.4 Constant Price Resource Allocation

For an efficient decentralized bandwidth allocation in a dynamic network environment, one strategy is to avoid solving problem (2.8) for every call arrival to and/or departure from any service area k. Meanwhile, our main objective is to satisfy the required bandwidth allocation per call for a target call blocking probability. This can be achieved through employing fixed link access price values for bandwidth allocation at different BSs/APs, independent of call arrivals and departures. Using time-invariant BS/AP link access price values, the corresponding bandwidth allocation is referred to as constant price resource allocation (CPRA) [15]. The CPRA works in two phases, namely set-up phase and operation phase. The set-up phase takes place only once at the initial operation time of the networks, while the operation phase takes place every time a new MT joins the networks.

A. The Set-up Phase

The main objective of this phase is to determine the fixed BS/AP link access price values that will be used during the operation phase. These are based on steady-state statistics of call traffic and user mobility so as to achieve satisfactory performance in terms of the allocated bandwidth per call and call blocking probability in the operation phase.

Consider the geographical region shown in Figure 2.1. In the set-up phase, let the number of calls of each service class l in each service area k for subscribers of a given network n, M_{lk}^n, equals to a target value $\widehat{M_{lk}^n}$. The corresponding optimal link access price value for each BS/AP in the geographical region can be determined using the DORA mechanism with $\widehat{M_{lk}^n}$ values for all $n \in \mathcal{N}$ subscribers of different networks and $\forall l \in \mathcal{L}, k \in \mathcal{K}$. The radio resources of all networks will be distributed exactly over $\widehat{M_{lk}^n}$ calls $\forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K}$, if we employ these BS/AP link access price values for bandwidth allocation in the operation phase. Thus, for subscribers of a given network n, when $M_{lk}^n = \widehat{M_{lk}^n}$ in the operation phase, any incoming call from a network n subscriber with service class l to service area k will be blocked. This means that the choice of the target value $\widehat{M_{lk}^n}$ for all networks' subscribers and $\forall l \in \mathcal{L}, k \in \mathcal{K}$, and in turn the corresponding BS/AP link access price λ_{ns} $\forall n \in \mathcal{N}, s \in \mathcal{S}_n$, in the set-up phase determines the geographical region overall performance in terms of the allocated bandwidth per call and the call blocking probability in the operation phase. Hence, the value of $\widehat{M_{lk}^n}$ should be properly chosen to achieve target performance in the resource allocation. For a dynamic system, M_{lk}^n is a random variable. Alternatively, we can represent $\widehat{M_{lk}^n}$ by a design parameter ϵ_{lk}^n using the probability distribution of M_{lk}^n for subscribers of every network n and $\forall l \in \mathcal{L}, k \in \mathcal{K}$, such that

$$\Pr(M_{lk}^n > \widehat{M_{lk}^n}) \leq \epsilon_{lk}^n, \quad \forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K} \tag{3.6}$$

where $\epsilon_{lk}^n \in [0, 1]$. It is evident that the value of $\widehat{M_{lk}^n}$ depends on both ϵ_{lk}^n and the distribution of M_{lk}^n. Indeed, from (3.6), ϵ_{lk}^n gives an upper bound of the call blocking probability for subscribers of a given network n with service class l in service area k when $\widehat{M_{lk}^n} \leq C_{lk}^n$. Otherwise, let $\widehat{M_{lk}^n} = C_{lk}^n$, and both the optimal solution of (2.8) and the CPRA result in the same call blocking performance. Hence, the value of $\widehat{M}_{lk}^{n}$ can be chosen based on the requirement on call blocking probability for a given network n with service class l in service area k.

As call arrivals of service class l to service area k follow a Poisson process, the channel holding time follows a general distribution, and all calls are served simultaneously without queuing, an $M/G/\infty$ model [61] can be used to determine $\widehat{M}_{lk}^{n}$ for network n subscribers and $\forall l \in \mathcal{L}, k \in \mathcal{K}$ in the set-up phase, using the steady-state call traffic and user mobility statistics. Let υ_{lk}^{n} denote the arrival rate of new and handoff calls from network n subscribers with service class l in service area k. A BS/AP in k can determine υ_{lk}^{n} by counting the number of new and handoff call arrivals from network n subscribers with service class l to service area k and divide it by the total elapsed time. Then, the number of calls for network n subscribers with service class l that are simultaneously present in service area k, M_{lk}^{n}, follows a Poisson distribution with mean $\kappa_{lkn} = \upsilon_{lk}^{n}.E[T_h^{lk}]$ [61], where $E[T_h^{lk}]$ denotes the average channel holding time of service class l in service area k and can be calculated using (3.5) as

$$E[T_h^{lk}] = \frac{\varsigma_l}{\varsigma_l + 1} \cdot \frac{1}{\frac{1}{\bar{T}_r^k} + \frac{\varsigma_l}{\bar{T}_c^l}} + \frac{1}{\varsigma_l + 1} \cdot \frac{1}{\frac{1}{\bar{T}_r^k} + \frac{1}{\varsigma_l \bar{T}_c^l}}, \quad \forall l \in \mathcal{L}, k \in \mathcal{K}. \tag{3.7}$$

Hence, from (3.6), $\widehat{M}_{lk}^{n}$ is the minimum integer which satisfies [61]

$$\sum_{m=0}^{\widehat{M}_{lk}^{n}} \frac{\kappa_{lkn}^{m} e^{-\kappa_{lkn}}}{m!} \geq (1 - \epsilon_{lk}^{n}), \quad \forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K}. \tag{3.8}$$

For a given ϵ_{lk}^{n}, using $\widehat{M}_{lk}^{n}$ $\forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K}$, problem (2.8) can be solved using the DORA mechanism in order to find the corresponding optimal link access price values $\widehat{\lambda}_{ns}$ $\forall n \in \mathcal{N}, s \in \mathcal{S}_n$.

B. The Operation Phase

The main objective of this phase is to perform the bandwidth allocation process for each user joining the networks based on the following four steps.

Table 3.1: Calculation of bandwidth share from each available network BS/AP at MT m.

```
Input: $\widehat{\lambda}_{ns}\ \forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$, $[B_m^{\min}, B_m^{\max}]$, $m \in \mathcal{M}$;
Initialization: $\mu_m^{(1)}(1) \geq 0$; $\mu_m^{(2)}(1) \geq 0$;
for $j = 1 : J$ do
    for $n \in \mathcal{N}_k$ do
        for $s \in \mathcal{S}_{nk}$ do
            $b_{nms}(j) = [(\frac{\eta_1}{\widehat{\lambda}_{ns}+(\mu_m^{(1)}(j)-\mu_m^{(2)}(j))+\eta_2(1-\omega_{nms})} - 1)/\eta_1]^+$;
        end for
    end for
    $\mu_m^{(1)}(j+1) = [\mu_m^{(1)}(j) - \alpha_1(B_m^{\max} - \sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms}(j))]^+$;
    $\mu_m^{(2)}(j+1) = [\mu_m^{(2)}(j) - \alpha_2(\sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms}(j) - B_m^{\min})]^+$;
end for
Output: The required $b_{nms}\ \forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$.
```

Step 1: Each network BS/AP in the geographical region fixes its link access price value to the value calculated in the set-up phase, $\widehat{\lambda}_{ns}$, independent of call arrivals and departures. This fixed value, $\widehat{\lambda}_{ns}$, is broadcasted by each network $n \in \mathcal{N}$ BS/AP $s \in \mathcal{S}_n$ via its ID beacon.

Step 2: An incoming MT listens to the link access price values of the BSs/APs available at its location through its multiple radio interfaces.

Step 3: The link access price values are then used by the MTs in order to solve for the bandwidth share from each network BS/AP such that the total amount of allocated resources from all BSs/APs satisfies the call required bandwidth. This can be calculated at MT, m, with service class l in service area k, using the mechanism in Table 3.1, which is based on the DORA mechanism.

Step 4: MT, m, asks BS/AP s of network n, $\forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$, for the calculated bandwidth share b_{nms}. The BS/AP performs the required bandwidth allocation if it has sufficient resources. The MT call is blocked if the call total required bandwidth is not satisfied by the total allocated radio resources.

In the CPRA, no resource reallocations to existing calls are required since the BS/AP link access price values are independent of call arrivals to and departures from different service areas. Moreover, the required J iterations to reach the desired resource allocations from all BSs/APs to satisfy the call total required bandwidth is solved locally at each MT. Hence, no information exchange is required between the MTs and the BSs/APs for every iteration as in the DORA mechanism. Thus, almost no signalling overhead is required in the CPRA in order to reach the required bandwidth from each BS/AP[1]. The convergence of the CPRA follows the convergence of the DORA mechanism which is given in Chapter 2. However, unlike the DORA mechanism, the CPRA provides a sub-optimal solution to problem (2.8) since the link access price value is not updated with every call arrival and departure.

In the CPRA, a low call blocking probability can be obtained in the operation phase using a small value of ϵ_{lk}^{n}. However, this corresponds to a large $\widehat{M}_{lk}^{n}$ value. This results in a large BS/AP link access price values, which leads to a low amount of bandwidth allocation per call in the operation phase. On the other hand, a large ϵ_{lk}^{n} value results in a high call blocking probability and a large amount of bandwidth allocation per call in the operation phase. As a result, the value of ϵ_{lk}^{n} should be chosen so as to balance the trade-off between the allocated bandwidth per call and the call blocking probability.

[1] This is apart from the required overhead in broadcasting the fixed link access price value $\widehat{\lambda}_{ns}$ by every BS/AP on its ID beacon. However, the contribution of broadcasting this value to the overhead is negligible.

Using an appropriate choice of ϵ_{lk}^{n}, the CPRA with its setup and operation phases can allocate bandwidth for a target call blocking probability in the decentralized network architecture with dynamic call arrivals and/or departures.

3.5 Prediction Based Resource Allocation

The CPRA is performed based on $\widehat{M_{lk}^{n}}$ which is calculated according to the steady-state (long-term) call traffic and user mobility statistics. However, in a dynamic environment, with call arrivals and departures, M_{lk}^{n} can deviate from $\widehat{M_{lk}^{n}}$ for some time. Yet, the allocated bandwidth in the operation phase does not adapt to the short-term dynamics in the call traffic load. Hence, even if there exist sufficient resources in the BSs/APs that can be used to improve a video call quality, the call can be allocated only its minimum required bandwidth. In CPRA, these unutilized extra resources (at a low call traffic load) are actually reserved for possible incoming calls so as to satisfy the target call blocking probability. A bandwidth allocation adaptive to a short-term call traffic load (via bandwidth reallocation to the calls in service) can help to provide a better service quality compromise between the existing calls (in terms of the amount of allocated bandwidth to each call) and the potential incoming calls (in terms of the call blocking probability). Towards this end, in the following, we propose to update $\widehat{M_{lk}^{n}}$ $\forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K}$ in the operation phase periodically with period τ, and hence update the corresponding BS/AP link access price values, based on the instantaneous M_{lk}^{n} value at time t, $M_{lk}^{n}(t)$. We refer to the corresponding bandwidth allocation as prediction based resource allocation (PBRA) [15].

Let the time be partitioned into a set of periods, $\mathcal{T}$, of constant duration τ, $\mathcal{T} = \{T_1, T_2, \ldots,\ T_o, \ldots\}$. Let t_o denote the beginning of each period T_o. A time vector of arrival events for calls of network n subscribers with service class l in service area k

during period T_o is denoted by $\vec{\mathcal{T}}^o_{lkn}$. The PBRA mechanism is carried out in the following six steps.

Step 1: Given a new call arrival at time instant $t^o_\pi \in \vec{\mathcal{T}}^o_{lkn}$, $\pi = \{1, 2, \ldots, \left|\vec{\mathcal{T}}^o_{lkn}\right|\}$, in period T_o, the number of calls of network n subscribers with service class l in service area k at the time instant, $M^n_{lk}(t^o_\pi)$, is used by the BSs/APs in this service area to probabilistically predict the number of calls at time instant $t^o_\pi + \tau$ in the next time period T_{o+1}. Hence, τ is referred to as the prediction duration. The predicted number, $\widetilde{M}^n_{lk}(t^o_\pi + \tau)$, should satisfy

$$\Pr(M^n_{lk}(t^o_\pi + \tau) > \widetilde{M}^n_{lk}(t^o_\pi + \tau) | M^n_{lk}(t^o_\pi)) \leq \epsilon^n_{lk}, \quad \forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K}. \tag{3.9}$$

In order to determine $\widetilde{M}^n_{lk}(t^o_\pi + \tau)$, the conditional probability mass function (PMF) of $M^n_{lk}(t^o_\pi + \tau)$ given $M^n_{lk}(t^o_\pi)$, $f_{M^n_{lk}(t^o_\pi+\tau)|M^n_{lk}(t^o_\pi)}(m)$, is calculated using the transient distribution of the $M/G/\infty$ model [62]. First, we present the following definitions under the assumption of stationary call arrival and departure processes:

- ϕ^{lkn}_τ - The probability that a call of network n subscribers with service class l which is in service area k at time t^o_π is still present in the same service area at time $t^o_\pi + \tau$;
- θ^{lkn}_τ - The probability that a call of network n subscribers with service class l that arrives in service area k during $(t^o_\pi, t^o_\pi + \tau]$ is still present in the same service area at time $t^o_\pi + \tau$;
- $\rho(\varkappa_1, \varkappa_2)$ - A binomial random variable with parameters $\varkappa_1$ and $\varkappa_2$;
- $\xi(\varkappa)$ - A Poisson random variable with mean $\varkappa$.

Using $M^n_{lk}(t^o_\pi)$, we have [62]

$$M^n_{lk}(t^o_\pi + \tau) =_d \rho(M^n_{lk}(t^o_\pi), \phi^{lkn}_\tau) + \xi(\upsilon^n_{lk} \tau \theta^{lkn}_\tau) \tag{3.10}$$

where $=_d$ denotes equality in distribution. The probabilities ϕ_τ^{lkn} and θ_τ^{lkn} are given as [62]

$$\phi_\tau^{lkn} = \frac{1}{E[T_h^{lk}]}\int_\tau^\infty \Pr(T_h^{lk} > y)dy = \frac{1}{E[T_h^{lk}]}\int_\tau^\infty (1 - F_{T_h^{lk}}(y))dy \tag{3.11}$$

$$\theta_\tau^{lkn} = \int_0^\tau \frac{1}{\tau}\Pr(T_h^{lk} > y)dy = \int_0^\tau \frac{1}{\tau}(1 - F_{T_h^{lk}}(y))dy = \frac{E[T_h^{lk}]}{\tau}(1 - \phi_\tau^{lkn}) \tag{3.12}$$

where $F_{T_h^{lk}}(y) = \int_0^y f_{T_h^{lk}}(t)dt$ is the CDF of T_h^{lk}. The conditional PMF, $f_{M_{lk}(t_\pi^o+\tau)|M_{lk}(t_\pi^o)}(m)$, can be calculated using (3.10) - (3.12). Hence, $\widetilde{M}_{lk}(t_\pi^o + \tau)$ can be calculated using (3.9) as the minimum integer satisfying

$$\sum_{m=0}^{\widetilde{M}_{lk}^n(t_\pi^o+\tau)} f_{M_{lk}^n(t_\pi^o+\tau)|M_{lk}^n(t_\pi^o)}(m) \geq (1 - \epsilon_{lk}^n), \quad \forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K}. \tag{3.13}$$

Step 2: Each BS/AP in the geographical region records the predicted values of $\widetilde{M}_{lk}^n(t_\pi^o + \tau)$, $\forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K}$ for $\pi = \{1, 2, \ldots, \left|\vec{\mathcal{T}}_{lkn}^o\right|\}$ in a vector $\vec{\mathcal{M}}_{lkn}^{o+1}$.

Step 3: For the subscribers of each network $n \in \mathcal{N}$, the maximum predicted number of calls from each service class $l \in \mathcal{L}$ in each service area $k \in \mathcal{K}$ during T_{o+1}, $\widetilde{M}_{lk}^n(T_{o+1})$, is calculated at t_{o+1} from $\vec{\mathcal{M}}_{lkn}^{o+1}$. Hence, $\widetilde{M}_{lk}^n(T_{o+1}) = \max(\vec{\mathcal{M}}_{lkn}^{o+1})$ if it is less than or equal to C_{lk}^n, otherwise $\widetilde{M}_{lk}^n(T_{o+1}) = C_{lk}^n$. This ensures that for $\widetilde{M}_{lk}(T_{o+1}) \leq C_{lk}$, we have

$$\begin{aligned} &\Pr(M_{lk}^n(t_\pi^{o+1}) > \widetilde{M}_{lk}^n(T_{o+1})) \leq \epsilon_{lk}^n, \\ &\qquad\qquad \forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K}, \pi \in \{1, 2, \ldots, \left|\vec{\mathcal{T}}_{lkn}^{o+1}\right|\}. \end{aligned} \tag{3.14}$$

Step 4: Through cooperative networking, different BSs/APs in the geographical region exchange their information regarding $\widetilde{M}_{lk}^n(T_{o+1})$ $\forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K}$. Problem (2.8) can be solved at each BS/AP to update its link access price value which is fixed over T_{o+1}, independent of call arrivals to and departures from different service areas, and is broadcasted on the BS/AP ID beacon.

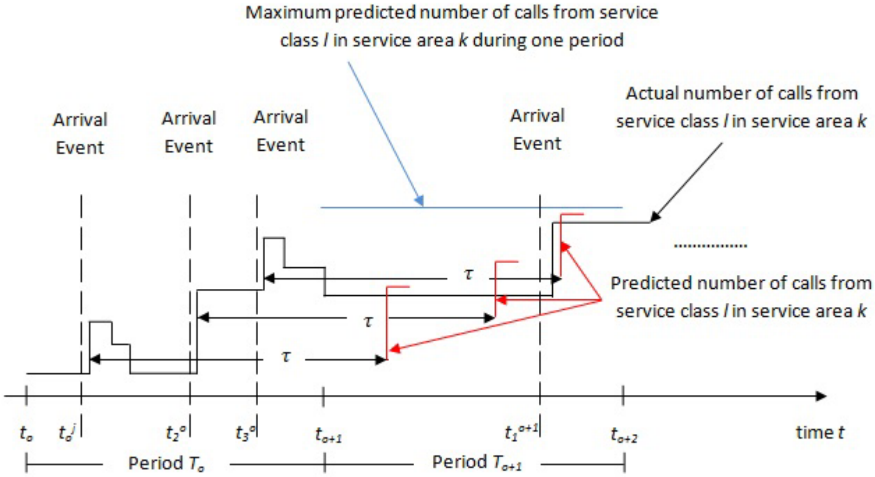

Maximum predicted number of calls from service class l in service area k during one period
Arrival Event
Arrival Event
Arrival Event
Arrival Event
Actual number of calls from service class l in service area k
Predicted number of calls from service class l in service area k
τ
τ
τ
t_o
t_o^j
t_2^o
t_3^o
t_{o+1}
t_1^{o+1}
t_{o+2}
time t
Period T_o
Period T_{o+1}

Figure 3.1: Illustration of PBRA Events.

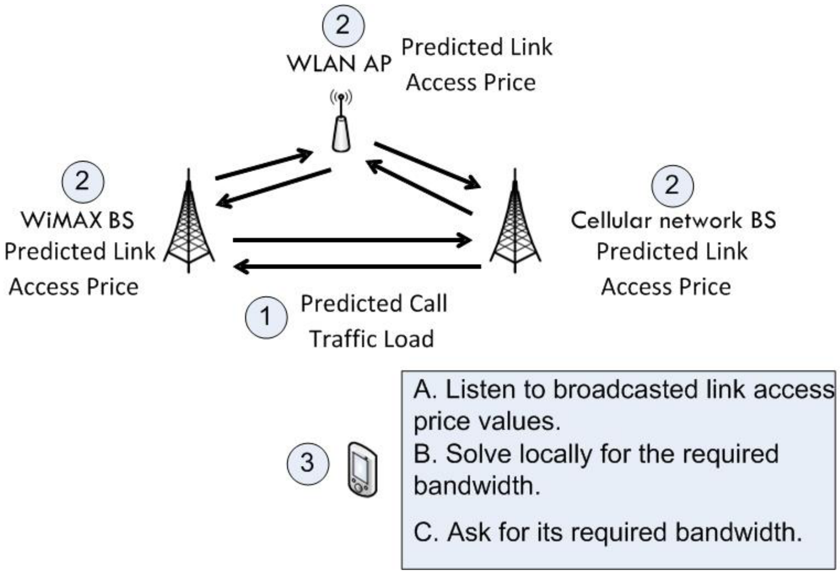
Figure 3.1 illustrates the call arrival times, the actual and predicted numbers of calls for network n subscribers with service class l in service area k associated with the steps 1-4.

Step 5: During T_{o+1}, each MT in the geographical region, including both incoming and already existing ones, uses the broadcasted BS/AP link access price values received at its location during this period to determine and ask for a bandwidth share from each available BS/AP. This is achieved following steps 2-4 in the CPRA[2].

Step 6: Each MT reports to the BSs/APs available at its location its service class, home network, and a list of the BS/AP IDs that the MT can receive. BSs/APs of different networks use this information so as to predict $\widetilde{M}_{lk}^{n}(T_{o+2})$, $\forall n \in \mathcal{N}, l \in \mathcal{L}, k \in \mathcal{K}$, during the next period T_{o+2} to update their link access price values at time t_{o+2}.

While the CPRA uses the target $\widehat{M}_{lk}^{n}$ value from the set-up phase based on steady-state (long-term) statistics to perform the bandwidth allocation in the operation phase,

[2] In Table 3.1, $\widehat{\lambda}_{ns}$ is replaced by the updated link access price value the MT receives during T_{o+1}.



A. Listen to broadcasted link access price values.

B. Solve locally for the required bandwidth.

C. Ask for its required bandwidth.

Figure 3.2: The PBRA procedure.

the PBRA updates the target value by $\widetilde{M}_{lk}^n(T_o)$ every period T_o, $o = \{1, 2, \ldots\}$, using the current number of calls in service. Using this extra information, the PBRA can make a better prediction of the call traffic load carried in the geographical region in a short-term, and hence an improved bandwidth allocation is expected over the CPRA. The PBRA mechanism provides an improved sub-optimal solution to problem (2.8) as compared to the CPRA mechanism. The convergence of the PBRA mechanism to this sub-optimal solution follows the convergence of the DORA mechanism which is given in Chapter 2. As the BS/AP link access price values during period T_o are based on $\widetilde{M}_{lk}^n(T_o)$, the BSs/APs allocate their available resources exactly among $\widetilde{M}_{lk}^n(T_o)$ calls during period T_o. Thus, following the definitions in (3.9) and (3.14) and using the same argument of CPRA, ϵ_{lk}^n serves as an upper bound on the call blocking probability for $\widetilde{M}_{lk}^n(T_o) \leq C_{lk}^n$.

The PBRA procedure is illustrated in Figure 3.2. The main differences between the DORA and PBRA operations, which are made clear by comparing Figures 2.3 and 3.2, are summarized in the following:

1. In the DORA mechanism, the link access price values for different BSs/APs are updated with every call arrival to and/or departure from any service area. This requires bandwidth reallocations for all existing MTs, which results in high signalling overhead. On the other hand, the PBRA updates the BSs/APs link access price values every τ, fix and broadcast them during τ. This can significantly reduce the amount of signalling overhead over the air interface, and is achieved through short-term call traffic load prediction and network cooperation. Specifically, cooperative networking allows different networks to exchange the necessary information required so as to enable each BS/AP to calculate and broadcast the predicted link access price value for the next τ duration.

2. In the DORA mechanism, each MT plays an active role in the bandwidth allocation operation by coordinating different BSs/APs bandwidth allocations so as to satisfy the call total required bandwidth. While in the PBRA, the MT active role is to calculate the required bandwidth share from each BS/AP to satisfy the call total required bandwidth. Hence, in the PBRA, all the necessary information for the calculations are made locally available to the MT, unlike the DORA mechanism, which again significantly reduces the amount of signalling overhead required over the air interface in order to determine the required bandwidth share from each network BS/AP.

Different BSs/APs can use a look-up table for implementation simplicity of the PBRA mechanism. Specifically, an BS/AP can store the link access price value corresponding to a given number of MTs with different service classes in different service areas. Hence, when BSs/APs exchange their information regarding the predicted call traffic load for different service classes in different service areas, every BS/AP can easily find the corresponding link access price value. By broadcasting the corresponding link access price value, an MT can determine/update its required bandwidth share from the networks available at its location to satisfy its total required bandwidth.

In the following section, we present a complexity analysis for the DORA implementation in a dynamic system, the CPRA, and the PBRA.

3.6 Complexity Analysis

The complexity analysis in this section examines both signalling overhead and processing time complexity for the DORA implementation in a dynamic environment, the CPRA, and the PBRA.

A. Signalling Overhead

In order to implement the DORA mechanism in a dynamic system, information signalling needs to be exchanged between all existing MTs and BSs/APs with every call arrival to and/or departure from any service area in order to reach the optimal bandwidth allocation. This signalling overhead is a function of the call arrival and departure rates, the number of existing calls in different service areas, and the required number of iterations J for the DORA mechanism to converge to the optimal bandwidth allocation. Denote χ_1 and χ_2 as the average number of call arrivals and departures over a period, respectively. Hence, for the DORA implementation in a dynamic system, the signalling overhead on the air interface scales as $O(\chi_1 + \chi_2)$ over the period. As a result, for high call arrival/departure rates, a high signalling overhead is expected. On the other hand, for the CPRA and PBRA, the link access price values for different BSs/APs are independent of call arrivals and departures. Thus, in order to reach the required resource allocation, their signalling overhead on the air interface scales as $O(1)$. As a result, the signalling overhead for the CPRA and the PBRA scales well with the call arrival and departure rates, as compared with the DORA implementation in a dynamic system.

B. Processing Time

In the DORA mechanism, MTs and BSs/APs exchange signalling information for J iterations in order to reach an optimal bandwidth allocation. Let σ denote the total amount of time required for the signalling exchange completion for the J iterations. The signalling exchange for J iterations should take place with every call arrival to and/or departure from any service area. Then, the time duration between two successive execution of the J-iteration signalling exchange is expressed as $\delta = \min(\text{call inter-arrival time}, \text{call departure time})$. Since the call arrivals follow a Poisson process with parameter υ_{lk}, the call inter-arrival time follows an exponential distribution with PDF $f_{T^{lk}}(t)$. The channel holding time gives the call departure time, which follows a hyper-exponential distribution with PDF $f_{T_h^{lk}}(t)$. Using the same analysis as given in (3.3) - (3.5), the PDF of δ, $f_\delta(t)$, is expressed as

$$\begin{aligned} f_\delta(t) \;&= \tfrac{\varsigma_l}{\varsigma_l+1} \cdot (\tfrac{1}{\bar{T}_r^k} + \tfrac{\varsigma_l}{\bar{T}_c^l} + \upsilon_{lk}) \cdot e^{-(\frac{1}{\bar{T}_r^k} + \frac{\varsigma_l}{\bar{T}_c^l} + \upsilon_{lk})t} + \tfrac{1}{\varsigma_l+1} \cdot (\tfrac{1}{\bar{T}_r^k} + \\ &\tfrac{1}{\varsigma_l \bar{T}_c^l} + \upsilon_{lk}) \cdot e^{-(\frac{1}{\bar{T}_r^k} + \frac{1}{\varsigma_l \bar{T}_c^l} + \upsilon_{lk})t}, \quad t \geq 0. \end{aligned} \tag{3.15}$$

Using (3.15), the average of δ is given by

$$\bar{\delta} = \frac{\varsigma_l}{\varsigma_l + 1} \cdot \frac{1}{\frac{1}{\bar{T}_r^k} + \frac{\varsigma_l}{\bar{T}_c^l} + \upsilon_{lk}} + \frac{1}{\varsigma_l + 1} \cdot \frac{1}{\frac{1}{\bar{T}_r^k} + \frac{1}{\varsigma_l \bar{T}_c^l} + \upsilon_{lk}}. \tag{3.16}$$

It is clear from (3.16) that the DORA mechanism processing time does not scale with the call arrival and departure rates, since $\bar{\delta}$ is inversely proportional to them. As $\bar{\delta}$ decreases with increasing arrival and/or departure rates while σ increases with increasing arrival rates (this is especially true in a case that a contention based medium access control network is among the available networks), $\bar{\delta}$ can be smaller than σ. Hence, the DORA mechanism does not converge to an optimal allocation whenever $\bar{\delta}$ is smaller than σ. On the other hand, for the CPRA and the PBRA, the J iterations are solved locally at the MTs and no signalling information is exchanged for each iteration, unlike the DORA mechanism. As a result, both the CPRA and the PBRA reach the required bandwidth allocation from each BS/AP independent of the call arrival and departure rates.

The CPRA and the PBRA require that the BS/AP link access price values to be broadcasted by each BS/AP on its ID beacon. Furthermore, the PBRA requires an exchange of the predicted call traffic load among different BSs/APs with overlapped coverage every τ. However, unlike the DORA mechanism, this signalling exchange does not take place on the air interface, but is executed over the signalling backbone connecting different networks.

Since the link access price value for different BSs/APs are updated every τ, the choice of the τ duration should reflect some change in the call traffic load in the geographical region. Hence, as a guideline, the time duration τ can be chosen such that the probability $\Pr[\delta > \tau]$ is less than a small threshold.

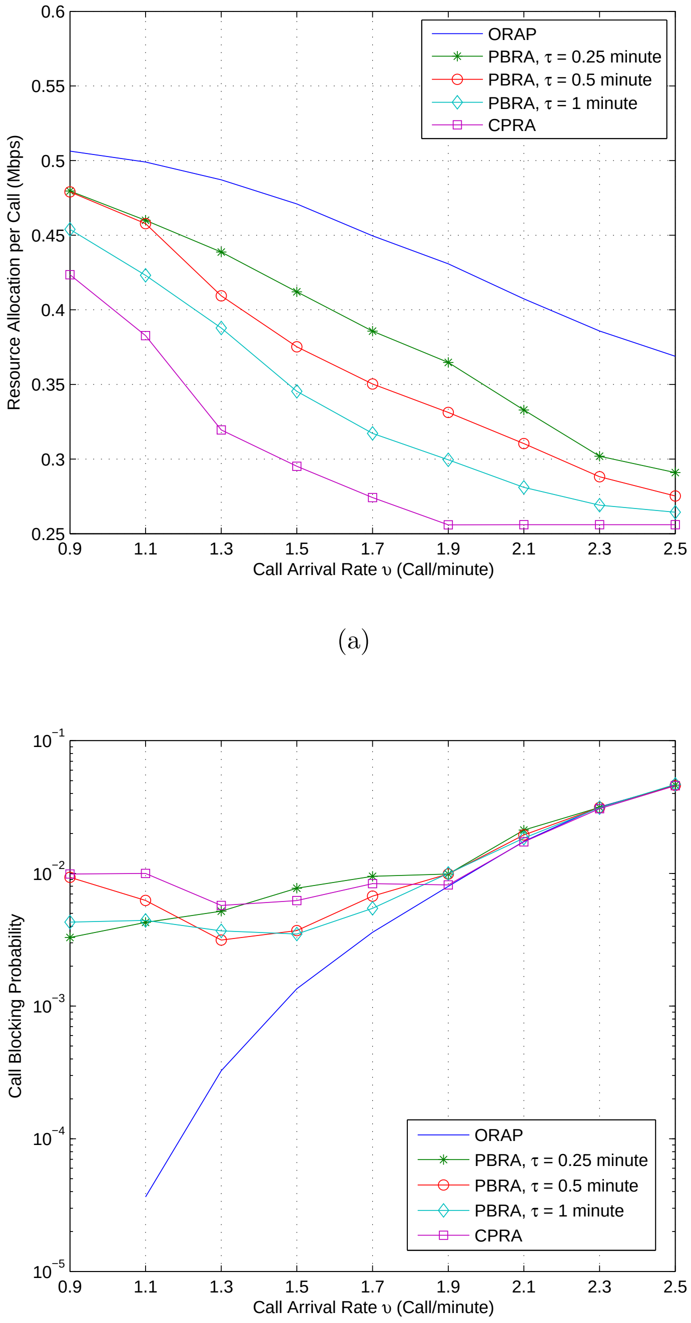
3.7 Simulation Results and Discussion

In this section, we present simulation results for the bandwidth allocation in a heterogeneous wireless access medium for MTs with multi-homing capabilities, using the PBRA mechanism as compared to problem (2.8) exact solution and the CPRA. Consider the geographical region of Figure 2.4. A single VBR service class ($l = 1$) is considered and we study the performance of the PBRA mechanism in the service area ($k = 1$) which is covered by all three networks, in terms of the allocated bandwidth per call and the call blocking probability. For simplicity, it is assumed that only subscribers of one network are present, and all networks treat them in the same manner (i.e. $\omega_{nms} = 1$ from all networks). The transmission capacity allocated from network n BS/AP to the service area under consideration is given by $C_1 = 2$ Mbps, $C_2 = 0.656$ Mbps, $C_3 = 1.184$ Mbps. A total of 15 VBR calls with required bandwidth allocation $[0.256, 0.512]$ Mbps for MTs with multi-homing capabilities can be supported in the service area under consideration using the given C_n values, that is $C = 15$ (indices n, l, k are dropped for simplicity). The new and handoff video call arrival process is modeled by a Poisson process with parameter υ (call/minute - indices n, l, k are dropped for simplicity). A hyper-exponential distribution is used to model the video call duration, with the PDF given in (3.1) and $\varsigma_1 = 6$. The average video call duration $\bar{T}_c = 20$ minutes. The user residence time in the service area under consideration follows an exponential distribution with the PDF given in (3.2) having an average time $\bar{T}_r = 15$ minutes [42]. The parameters η_1 and η_2 are set to 1 [43].

A. Performance Comparison

In the following, the performance of the PBRA mechanism is compared with the optimal solution of problem (2.8) in terms of the allocated bandwidth per call and the call blocking probability. The optimal solution of (2.8) is referred to as ORAP and can be obtained using a centralized resource allocation (which gives the same resource allocation of the DORA mechanism if the signalling overhead and time complexity issues are neglected). Although it is not appropriate for practical implementation when different networks are operated by different service providers, the ORAP is used to serve as an upper bound for the system performance in terms of the allocated bandwidth per call and a lower bound for the system performance in terms of the call blocking probability. The CPRA is also considered in the comparison, where no update of the link access price values takes place.

Figure 3.3 shows performance comparison among the CPRA, PBRA, and ORAP versus the call arrival rate υ, with $\epsilon = 1\%$ and $\tau = 0.25, 0.5,$ and 1 minute. At a low call arrival rate, the predicted number of simultaneously present calls is low, thus the predicted link access price value is low and the bandwidth allocation amounts per call



(a)

(b)

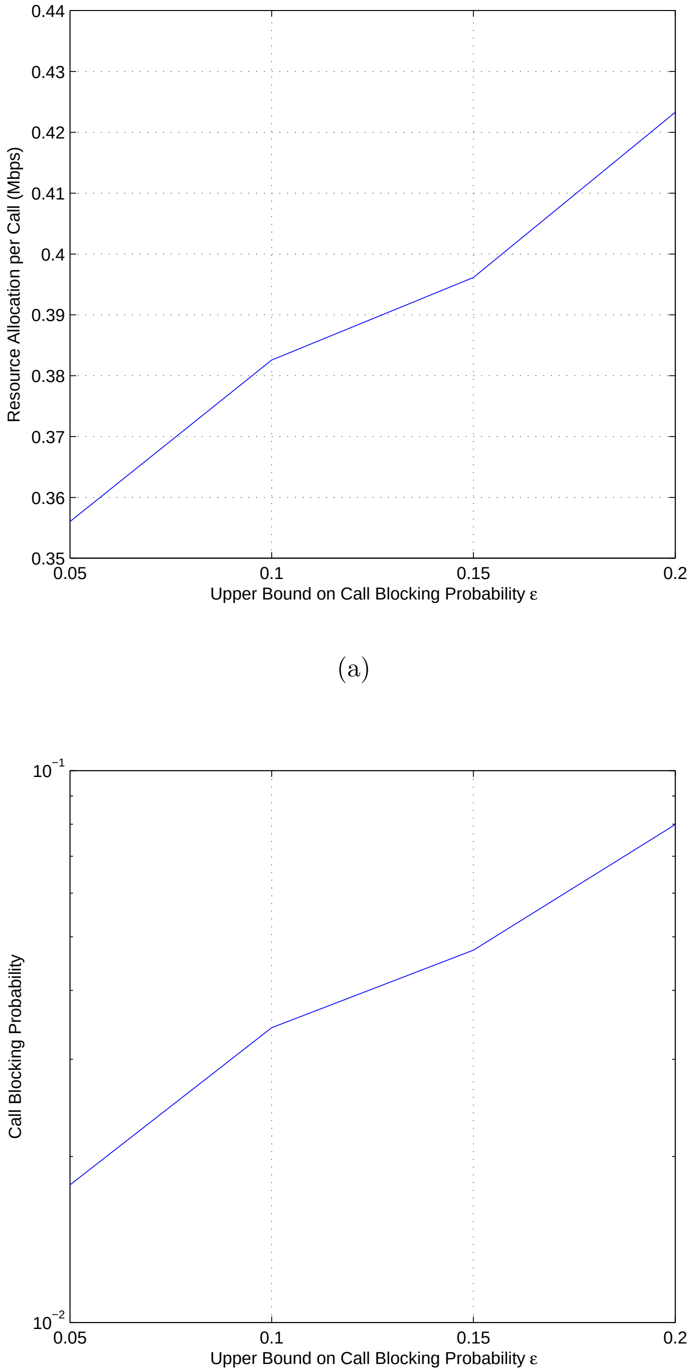
Figure 3.3: Performance comparison: (a) Bandwidth allocation per call; (b) Call blocking probability.

using the PBRA mechanism for the different τ values are high. On the other hand, at a high call arrival rate, the predicted number of simultaneously present calls in the system is high. For a larger τ value, less bandwidth is allocated per call as explained in the next subsection. The CPRA provides a lower bound of the performance in terms of the allocated bandwidth, as it does not update the BS/AP link access price values. For the ORAP, there is no call blocking probability for a call arrival rate $\upsilon < 1.1$ call/minute. All three mechanisms achieve the desired upper bound for call blocking probability, ϵ, for $\upsilon \leq 1.9$ call/minute. The predicted number of calls simultaneously present in the system is larger than C for $\upsilon > 1.9$ call/minute. As a result, according to the CPRA and the PBRA, the predicted number is made equal to C, and the mechanisms achieve the same call blocking probability as the ORAP. Overall, the PBRA performance lies between CPRA and ORAP performance, as expected. By properly choosing the τ value, the PBRA mechanism can achieve a desired compromise between performance and implementation complexity.

B. Performance of The PBRA Mechanism

In the following, the performance of the PBRA mechanism is studied versus its two parameters, namely the upper bound on the call blocking probability ϵ and the prediction duration τ.

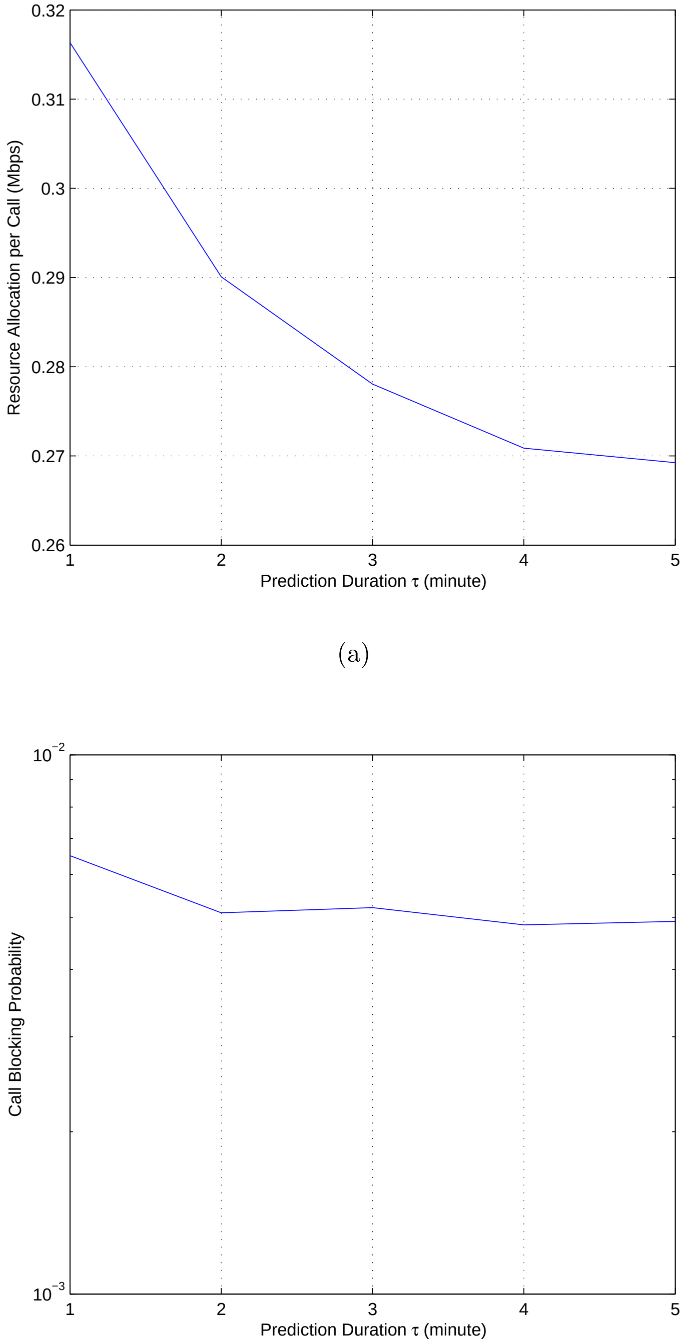
Figure 3.4 shows the performance of the PBRA mechanism in terms of the amount of bandwidth allocation per call and call blocking probability versus ϵ, with the call arrival rate $\upsilon = 1.7$ call/minute and the prediction duration $\tau = 1$ minute. As ϵ increases, the PBRA mechanism accounts for the simultaneous presence of less calls in service in the next τ in its calculation of the link access price value. This results in an increase in the call blocking probability with ϵ. In general, the call blocking probability does not exceed its upper bound ϵ as shown in Figure 3.4b. However, the bandwidth allocation per call is improved with ϵ, since less resources are reserved for incoming calls which will, more



(a)

(b)

Figure 3.4: The PBRA mechanism performance versus ϵ: (a) Bandwidth allocation per call; (b) Call blocking probability.



(a)

(b)

Figure 3.5: The PBRA mechanism performance versus τ: (a) Bandwidth allocation per call; (b) Call blocking probability.

likely, be blocked. Thus, a trade-off exists between these two performance metrics.

Figure 3.5 shows the performance of the PBRA in terms of the amount of bandwidth allocation per call and call blocking probability versus the prediction duration τ, with the call arrival rate $\upsilon = 1.7$ call/minute and $\epsilon = 1\%$. With a larger prediction duration τ, the PBRA mechanism updates the BS/AP link access price less frequently and a larger number of simultaneously present calls is predicted. As a result, the amount of allocated bandwidth per call is reduced. Again, the call blocking probability does not exceed its upper bound ϵ with the different τ values.

3.8 Summary

In this chapter, the limitations of the DORA mechanism in a dynamic system are discussed. A prediction based resource allocation (PBRA) mechanism is presented to address these limitations. The PBRA objective is to perform an efficient bandwidth allocation in a dynamic system that can reduce the signalling overhead required over the air interface for bandwidth allocation in a decentralized architecture while achieving an acceptable call blocking probability and a sufficient amount of allocated bandwidth per call. In order to achieve the objectives, the PBRA mechanism relies on short-term call traffic load prediction and network cooperation. There are two parameters in the PBRA mechanism, namely ϵ_{lk}^{n} and τ, that can be properly chosen to strike a balance between the desired performance in terms of the allocated bandwidth per call and the call blocking probability, and between the performance and the implementation complexity. In the PBRA mechanism, each MT plays an active role in the bandwidth allocation operation by requesting a bandwidth share from each available network based on the available resources at the network, such that the total allocated bandwidth from different networks satisfies the MT service requirement. However, the proposed PBRA mechanism supports only MTs with multi-homing service. It is envisioned that both single-network and multi-homing services will co-exist in the future heterogeneous wireless communication network. Hence, in the next chapter, we extend the concepts presented in Chapters 2 and 3 to include the presence of MTs with single-network service in the networking environment and propose a bandwidth allocation mechanism that can support MTs with single-network and multi-homing services in a decentralized manner.

Chapter 4

Resource Allocation for Single-network and Multi-homing Services

In the future wireless communication network, it is envisioned that both single-network and multi-homing services will co-exist. Hence, it is required to develop radio resource allocation mechanisms that can support both service types. In this case, the radio resource allocation mechanism is to determine the optimal network assignment for MTs with single-network service and the corresponding bandwidth allocation for MTs with single-network and multi-homing services. In this chapter, we discuss how to achieve these objectives in a decentralized network architecture with call arrivals and departures. Concepts of call traffic load prediction and cooperative networking, presented in Chapter 3, are employed to enable vertical handovers for single-network calls and to satisfy multi-homing calls required bandwidth in such a decentralized network architecture.

4.1 Introduction

In Chapters 2 and 3, we have presented a set of mechanisms to support a decentralized bandwidth allocation for MTs in a heterogeneous wireless access medium. In these mechanisms, an MT plays an active role in the resource allocation operation, whether by coordinating the bandwidth allocation from different networks or by calculating the required bandwidth share from each network and asking for this share to satisfy its total required bandwidth. However, the mechanisms can support only MTs with multi-homing capabilities. It is expected that both single-network and multi-homing services will coexist in future wireless networks. Many reasons support this vision. Firstly, not all calls require high data rates that call for a multi-homing support, and hence these calls can resort to a single-network service. In addition, not all MTs are currently equipped with multi-homing capabilities, thus they can only support a single-network service. Moreover, an MT with insufficient available energy can switch from a multi-homing service to a single-network service and turn off all its radio interfaces, except for the one with the best available wireless network, in order to save energy. Hence, it is required to develop a decentralized radio resource allocation mechanism that can support both single-network and multi-homing services. In such a decentralized architecture, an MT with single-network service should be able to select the best available wireless access network at its location and ask for its required bandwidth from this network. In addition, the MT should be able to perform a vertical handover whenever necessary so as to remain best connected. On the other hand, an MT with multi-homing service can determine the required bandwidth share from each network to satisfy its total required bandwidth. Hence, the objective of the radio resource allocation mechanism is twofold: First, to determine the optimal network assignment vector for MTs with single-network service; Second, to determine the corresponding optimal bandwidth allocation for MTs with single-network and multi-homing services. Towards this end, we first present a centralized optimal radio resource allocation (CORA) mechanism that can satisfy the aforementioned objectives. Then, based on the centralized mechanism and the concepts introduced in Chapter 3 for call traffic load prediction and network cooperation, we present a decentralized sub-optimal resource allocation (DSRA) mechanism.

4.2 System Model

In this section, we present the modifications to the system model presented in Chapter 3 to account for the presence of both single-network and multi-homing calls in the geographical region of Figure 2.1.

A. Service Types

Let $\mathcal{M}_k$ denote the subset of MTs in a given service area, k. Two service types are considered in the geographical region, i.e. single-network and multi-homing services. Let $\mathcal{M}_{vk}$ denote the subset of MTs with same service type in a given service area k, where $v = 1$ for single-network service and $v = 2$ for multi-homing service. An MT with single-network service, $m \in \mathcal{M}_{1k}$ in service area k, is assigned to a single network n BS/AP $s \in \mathcal{S}_{nk}$. Let $A = [a_1, \ldots, a_m, \ldots, a_{|\mathcal{M}_{1k}|}]$ denote the network assignment vector in the geographical region for MTs with single-network service, where $a_m = ns$ is the assignment of MT $m \in \mathcal{M}_{1k}$ to network n BS/AP s. For instance, $a_1 = 12$ is the assignment of MT 1 to network 1 BS/AP 2.

On the other hand, an MT, $m \in \mathcal{M}_{2k}$, with multi-homing service in a given service area k, receives its required bandwidth from all BSs/APs available at its location, $s \in \mathcal{S}_k$, using its multi-homing capability. The set $\mathcal{M}_{ns}$ of MTs assigned to network n BS/AP s includes both multi-homing and single-network MTs.

B. Call Traffic Models

There exists a set, $\mathcal{L}_v = \{1, 2, \ldots, L_v\}$, of service classes for each service type, v. In general, service class $\mathcal{L}_2$ for an MT with multi-homing service type requires larger bandwidth than service class $\mathcal{L}_1$ for an MT with single-network service. The allocated bandwidth from network n to MT m via BS/AP s, b_{nms}, is zero if MT $m \notin \mathcal{M}_{ns}$, and for single-network MT if $a_m \neq ns$. For subscribers of a given network, let M_{lvk} denote the number of existing calls of service type v and service class l in service area k and C_{lvk} is the maximum number of calls of each service type v and service class l which can be supported in each service area k for subscribers of the given network. A call admission control procedure is in place, which guarantees that $M_{lvk} \leq C_{lvk}$, such that feasible resource allocation solutions exist with sufficient resources for a target call traffic load.

The arrival process of both new and handoff calls of service type v and class l to service area k is modeled by a Poisson process with parameter υ_{lvk}. A two-stage hyper-exponential distribution is used to approximate the PDF of the video call duration, T_c^{lv}, with mean $\bar{T}_c^{lv}$, which is given by [42]

$$f_{T_c^{lv}}(t) = \frac{\varsigma_{lv}}{\varsigma_{lv}+1} \cdot \frac{\varsigma_{lv}}{\bar{T}_c^{lv}} \cdot e^{-\frac{\varsigma_{lv}}{\bar{T}_c^{lv}}t} + \frac{1}{\varsigma_{lv}+1} \cdot \frac{1}{\varsigma_{lv}\bar{T}_c^{lv}} \cdot e^{-\frac{1}{\varsigma_{lv}\bar{T}_c^{lv}}t}, \quad \varsigma_{lv} \geq 1, t \geq 0. \tag{4.1}$$

C. Mobility Models and Channel Holding Time

The user residence time within service area k is modeled by an exponential distribution with mean $\bar{T}_r^k$. Hence, the channel holding time for a given service type v with service class l in service area k, $T_h^{lvk} = \min(T_c^{lv}, T_r^k)$, has a PDF that is given by

$$\begin{aligned} f_{T_h^{lvk}}(t) \quad &= \quad \frac{\varsigma_{lv}}{\varsigma_{lv}+1} \cdot (\frac{1}{\bar{T}_r^k} + \frac{\varsigma_{lv}}{\bar{T}_c^{lv}}) \cdot e^{-(\frac{1}{\bar{T}_r^k} + \frac{\varsigma_{lv}}{\bar{T}_c^{lv}})t} \\ &+ \quad \frac{1}{\varsigma_{lv}+1} \cdot (\frac{1}{\bar{T}_r^k} + \frac{1}{\varsigma_{lv}\bar{T}_c^{lv}}) \cdot e^{-(\frac{1}{\bar{T}_r^k} + \frac{1}{\varsigma_{lv}\bar{T}_c^{lv}})t}, \quad t \geq 0. \end{aligned} \tag{4.2}$$

4.3 Centralized Optimal Resource Allocation

In this section, the radio resource allocation problem is formulated for MTs with single-network and multi-homing services in the heterogeneous wireless access medium. Based on the problem formulation, a centralized optimal resource allocation (CORA) mechanism is then presented.

A. Problem Formulation

The utility of network n allocating bandwidth b_{nms} to MT m via BS/AP s, $u_{nms}(b_{nms})$, is given by[1]

$$u_{nms}(b_{nms}) = \ln(1 + \eta_1 b_{nms}) - \eta_2(1 - \omega_{nms})b_{nms}. \tag{4.3}$$

Given a network assignment vector A, the overall resource allocation objective of all networks in the geographical region is to determine the optimal bandwidth allocation $b_{nms}, \forall n \in \mathcal{N}, m \in \mathcal{M}_{ns}, s \in \mathcal{S}_n$ which maximizes the total utility in the region, U, given by

$$U = \sum_{n=1}^{N} \sum_{s=1}^{S_n} \sum_{m \in \mathcal{M}_{ns}} u_{nms}(b_{nms}). \tag{4.4}$$

The allocated bandwidth from network n BS/AP s should satisfy the BS/AP capacity constraint given by

$$\sum_{m \in \mathcal{M}_{ns}} b_{nms} \leq C_n, \quad \forall s \in \mathcal{S}_n, n \in \mathcal{N}. \tag{4.5}$$

Given a network assignment vector A, for MTs with single-network service, the allocated bandwidth from the assigned network n BS/AP $s \in \mathcal{S}_{nk}$ to MT $m \in \mathcal{M}_{1k}$ in service area k should satisfy the application required bandwidth, given by

$$B_m^{\min} \leq b_{nms} \leq B_m^{\max}, \quad \forall m \in \mathcal{M}_{1k}, k \in \mathcal{K}. \tag{4.6}$$

[1] Equations (4.3), (4.5), and (4.7) are the same as equations (2.1), (2.5), and (2.7), respectively. They are re-stated for convenience.

While for MTs with multi-homing service, the total allocated bandwidth from all available BSs/APs in $\mathcal{S}_k$ to MT $m \in \mathcal{M}_{2k}$ in service area k should satisfy the application total required bandwidth, which is given by

$$B_m^{\min} \leq \sum_{n \in \mathcal{N}_k} \sum_{s \in \mathcal{S}_{nk}} b_{nms} \leq B_m^{\max}, \quad \forall m \in \mathcal{M}_{2k}, k \in \mathcal{K}. \tag{4.7}$$

In order to determine the optimal network assignment vector A and the corresponding optimal bandwidth allocation matrix B for single-network and multi-homing MTs, the radio resource allocation problem is expressed by the following optimization problem

$$\begin{aligned} \max_{A} \{ \max_{B \geq 0} \quad & U \\ s.t. \quad & (4.5) - (4.7). \} \end{aligned} \tag{4.8}$$

Given a network assignment vector A, the bandwidth allocation problem (i.e. the inner maximization problem of (4.8)) is a convex optimization problem that can be solved efficiently using polynomial time algorithms [44]. However, finding the optimal vector A (i.e. the outer maximization problem of (4.8)) incurs high computational complexity. In a given service area k with a total of $|\mathcal{M}_{1k}|$ MTs with single-network service and $|\mathcal{S}_k|$ BSs/APs available from different networks, there exist $|\mathcal{S}_k|^{|\mathcal{M}_{1k}|}$ distinct assignment vectors. As a result, the total number of distinct assignment vectors in the whole geographical region is $\prod_k |\mathcal{S}_k|^{|\mathcal{M}_{1k}|}$. For instance, consider one service area with a total of 50 MTs with single-network service and 3 BSs/APs having overlapped coverage. A total of $3^{50} = 7 * 10^{23}$ distinct network assignments exist in this service area. For the whole geographical region, it is expected that the inner maximization problem of (4.8) needs to be solved for a huge number of times so as to determine the optimal radio resource allocation (i.e. the optimal network assignment vector A and bandwidth allocation matrix B). As a result, it is desirable to develop a less complex formulation rather than the max-max formulation of problem (4.8). Towards this end, a binary assignment variable

w_{nms} is introduced [63], that is determined from the network assignment vector A for MT $m \in \mathcal{M}_{1k}$ by

$$w_{nms} = \begin{cases} 1, & \text{if } a_m = ns \\ 0, & \text{otherwise.} \end{cases} \tag{4.9}$$

while $w_{nms} = 1$ for MTs with multi-homing service in service area k for all $s \in \mathcal{S}_k$. Using the binary assignment variable, the problem of (4.8) can be reformulated as

$$\begin{aligned} \max_{w_{nms}, b_{nms} \geq 0} \quad & \sum_{n=1}^{N} \sum_{s=1}^{S_n} \sum_{m \in \mathcal{M}_{ns}} \{\ln(1 + \eta_1 w_{nms} b_{nms}) - \eta_2 (1 - \omega_{nms}) w_{nms} b_{nms}\} \\ s.t. \quad & \sum_{m \in \mathcal{M}_{ns}} w_{nms} b_{nms} \leq C_n, \quad \forall s \in \mathcal{S}_n, n \in \mathcal{N} \\ & B_m^{\min} \leq \sum_{n=1}^{N} \sum_{s \in \mathcal{S}_{nk}} w_{nms} b_{nms} \leq B_m^{\max}, \quad \forall m \in \mathcal{M}_k, k \in \mathcal{K} \\ & w_{nms} \in \{0, 1\}, \quad \forall m \in \mathcal{M}_{1k}, n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}, k \in \mathcal{K} \\ & \sum_{n=1}^{N} \sum_{s \in \mathcal{S}_{nk}} w_{nms} = 1, \quad \forall m \in \mathcal{M}_{1k}, k \in \mathcal{K} \\ & w_{nms} = 1, \quad \forall m \in \mathcal{M}_{2k}, n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}, k \in \mathcal{K}. \end{aligned} \tag{4.10}$$

The fourth constraint ensures that an MT with single-network service is assigned to one and only one BS/AP available at its location, while the last constraint allows an MT with multi-homing service to obtain its required bandwidth from all wireless networks available at its location. The problem of (4.10) is a non-convex mixed integer non-linear programming (MINLP) problem. In general, MINLP problems combine the difficulty of optimizing over integer variables with the handling of non-linear functions which makes them difficult to solve [64]. This is especially true when the objective and/or constraint functions are non-convex, which is the case in (4.10). Several new methods are proposed recently for solving MINLP problems [65]. Two classes of algorithms that solve MINLP problems can be distinguished. The first class includes deterministic algorithms such as branch and bound, outer approximation, generalized benders decomposition, and extended cutting plane [64, 65]. Non-convexities in MINLP problems can be addressed by global optimization approaches which are developed using convex envelopes or underestimators to formulate lower-bounding convex MINLP problems [65]. One example of deterministic global optimization methods for MINLP problems is branch and reduce [66], and other methods can be found in [65]. The second class of MINLP algorithms includes stochastic (heuristic) optimization algorithms such as the extended ant colony optimization [67].

The different algorithms of solving MINLP problems have been available through many solvers [68]. Deterministic solvers that claim to guarantee global optimality for non-convex general MINLP problems include AlphaBB, BARON, COUENNE, and LINDOGLOBAL [68]. On the other hand, stochastic solvers include MIDACO [69], however there is no guarantee for global optimality [68]. The BARON solver [70], which is available through GAMS [71, 72], has proven to be the most robust one among the currently available global solvers [73]. The BARON solver implements deterministic global optimization algorithms which integrate conventional branch and bound with a wide variety of range reduction tests [70]. The BARON solver guarantees to provide global optima under fairly general assumptions which include the availability of finite lower and upper bounds on the variables and their expressions in the MINLP to be solved [70]. Hence, to solve the radio resource allocation problem (4.10), we use the BARON solver through GAMS.

Figure 4.1 illustrates a centralized implementation of the radio resource allocation (CORA) mechanism based on the formulation of (4.10). In the CORA mechanism, each MT reports to all BSs/APs available at its location about its service type, service class,

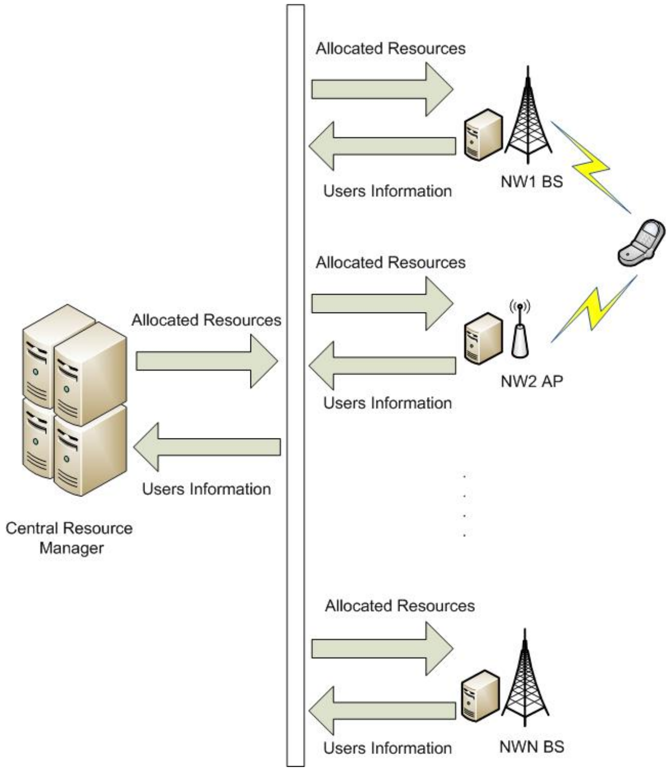


Figure 4.1: Centralized implementation of the CORA mechanism.

and home network using its multiple radio interfaces. This information then is made available to the central resource manager via different BSs/APs. Hence, the central resource manager has the information regarding the service area k for each MT, MT minimum and maximum required bandwidth, and MT priority parameter. Given the transmission capacities of all the BSs/APs, the central resource manager solves (4.10) so as to determine the optimal network assignment and bandwidth allocations for new incoming MTs with single-network and multi-homing services, updates bandwidth allocations and initiates vertical handovers for existing MTs if necessary.

B. Numerical Results and Discussion

This section presents numerical results for problem (4.10) using the BARON/GAMS solver. The GDXMRW utilities [74] are used to create an interface between GAMS and MATLAB in order to make use of GAMS as a powerful optimization platform and the MATLAB visualization tools. Consider the simplified system model given in Figure 2.4. We study the radio resource allocation in service area 2 which is covered by the WiMAX (network 1) and cellular network (network 2). For the service area under consideration, let the transmission capacity of each network BS be 4 Mbps for network 1 and 1.248 Mbps for network 2. The transmission capacities of different BSs are chosen such that they can support a total of 12 MTs with VBR calls of required bandwidth in $[64, 128]$ Kbps of single-network service, and a total of 17 MTs with VBR calls of required bandwidth in $[256, 512]$ Kbps of multi-homing service. The number of subscribers from network n with service v is given by M_{nv}, where $v = 1$ represents a single-network service while $v = 2$ represents a multi-homing service. With $M_{11} = 6$, $M_{21} = 6$, $M_{22} = 8$, we vary the number of network 1 subscribers with multi-homing service, M_{12}, in order to study the performance of the CORA mechanism as the call traffic load of the subscribers of the network with the larger capacity varies. Using the priority parameter ω_{nms}, the two networks set different costs on their resources. Since the cellular network (network 2) has a smaller transmission capacity than the WiMAX (network 1), it sets a higher cost on its resources so that it can devote its resources to its own subscribers [11]. As a result, let $\omega_{1m1} = 0.8$ and $\omega_{2m1} = 0.6$ for network users, while $\omega_{nm1} = 1$ for network subscribers with $n \in \mathcal{N}$. Let η_1 and η_2 equal 1. Let the number of assigned subscribers of network n, with single-network service, to network $n^{'}$ be $L_{nn^{'}}$.

Figure 4.2 shows the allocated bandwidth per call for MTs with single-network service versus the number M_{12} of network 1 subscribers with multi-homing service. As M_{12} increases, the allocated bandwidth for network 2 subscribers is reduced first towards the

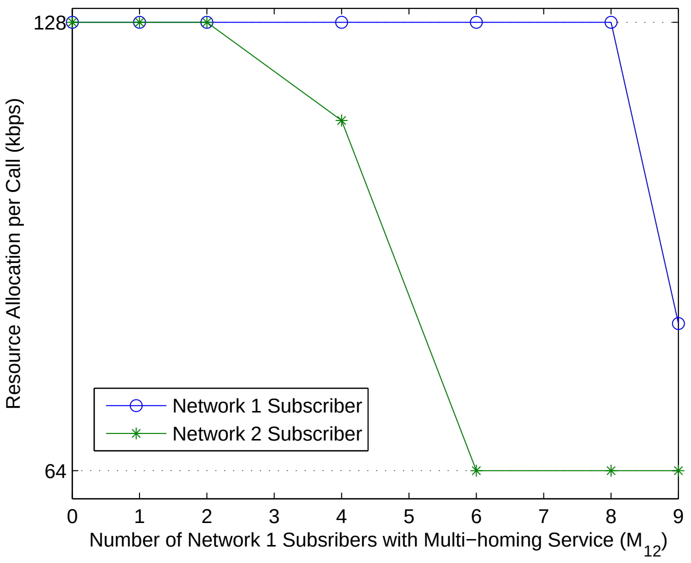


Figure 4.2: Bandwidth allocation for MTs with single-network service.

minimum required bandwidth. This is because network 2 subscribers rely heavily on network 1 resources in addition to their home network in order to support their high required bandwidth, while network 1 gives a higher priority to its own subscribers on its resources using the priority mechanism. The allocated resources to network 1 subscribers is then reduced so as to accommodate more multi-homing subscribers (M_{12}) from this network. Overall, the bandwidth allocation guarantees the desired bandwidth range for the VBR calls.

Table 4.1 shows the numbers of MTs with single-network services assigned to each BS/AP for network 1 and network 2 subscribers versus the number M_{12} of network 1 subscribers with multi-homing service. Due to the larger capacity of network 1, its subscribers are always assigned to their home network (L_{11}) which provides them with high allocated bandwidth (refer to Figure 4.2). As for network 2 subscribers, their network assignment varies with M_{12}. At a small number of M_{12} (from 0 to 2), all network 2

Table 4.1: Network assignments for network 1 and network 2 subscribers with single-network service.

M_{12}	L_{11}	L_{12}	L_{21}	L_{22}
0	6	0	6	0
1	6	0	6	0
2	6	0	6	0
4	6	0	4	2
6	6	0	3	3
8	6	0	3	3
9	6	0	2	4

subscribers with single-network service are assigned to network 1 (L_{21}), as it provides them with their maximum required bandwidth (refer to Figure 4.2). As the call traffic load increases in network 1 (due to an increase in M_{12}), more subscribers from network 2 are assigned to their home network (L_{22}), as network 1 gives higher priority to its own subscribers on its resources.

Figure 4.3 shows the allocated bandwidth per call for MTs with multi-homing service from each available network versus the number M_{12} of network 1 subscribers with multi-homing service. The total allocated bandwidth to network 1 subscribers ($N1$) comes from network 1 ($N1-1$). The allocated bandwidth from network 2 ($N2-1$) is zero, since network 2 devotes its resources to support its own subscribers using the priority parameter ω_{2m1}. The total allocated bandwidth per call for network 1 subscribers ($N1$) decreases with M_{12} towards the minimum required bandwidth to accommodate more subscribers. For network 2 subscribers, the allocated bandwidth from network 1 ($N1-2$) decreases as M_{12} increases, since network 1 uses its resources to support its own subscribers. This is compensated by an increase in the bandwidth allocation from network 2 ($N2-2$)

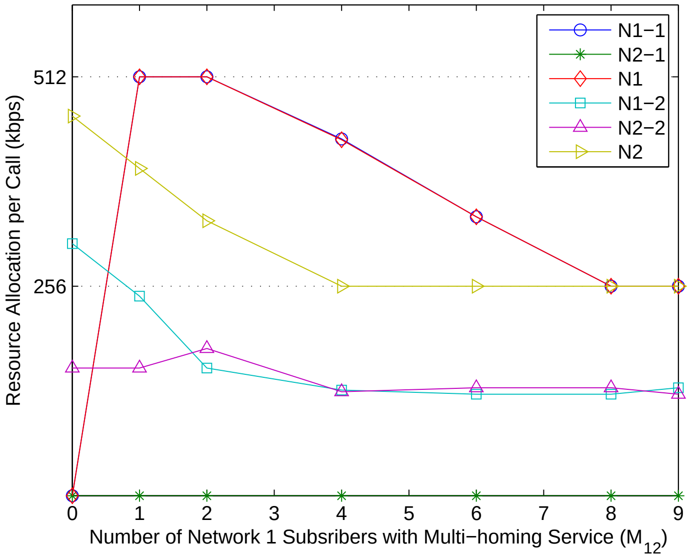


Figure 4.3: Bandwidth allocation for MTs with multi-homing service.

to improve the allocated bandwidth to its own subscribers. However, for $M_{12} > 2$, network 2 decreases its allocated bandwidth to its subscribers with multi-homing service, since more single-network subscribers are assigned to its BS (refer to Table 4.1). As a result, the total allocated bandwidth per call for network 2 subscribers ($N2$) decreases with M_{12} towards the minimum required bandwidth. The total allocated bandwidth per call for network 1 and network 2 subscribers with multi-homing services ($N1$ and $N2$, respectively) are within the desired bandwidth range for the VBR calls.

4.4 Decentralized Sub-optimal Resource Allocation

In this section, a decentralized sub-optimal resource allocation (DSRA) mechanism is presented for the radio resource allocation problem. The DSRA mechanism is desirable when different networks are operated by different service providers.

In problem (4.8), if the network assignment vector A is known, the problem is reduced to finding the optimal bandwidth allocation matrix B which is a convex optimization problem that can be solved in a decentralized manner using the decomposition approach discussed in Chapter 2. In order to find the network assignment vector A, call traffic load prediction and network cooperation concepts presented in Chapter 3 can be employed. Hence, in the DSRA mechanism, time is partitioned into a set of periods $\mathcal{T} = \{T_1, T_2, \ldots, T_o, \ldots\}$ of constant duration τ. At each BS/AP, the call traffic load at current period, T_o, is used to predict the call traffic load during the next period, T_{o+1}. By exchanging their predicted call traffic load information for the next period, cooperative BSs/APs can determine the distribution of the total call traffic load in the geographical region (i.e. network assignment vector A) for the next period, T_{o+1}. Based on the predicted call traffic load, every BS/AP broadcasts a parameter (a predicted link access price) which enables incoming and existing MTs to perform network selection and bandwidth request without the need for a central resource manager. As in Chapter 3, the call traffic load prediction is a probabilistic one which ensures that the prediction error is lower than a target value ϵ, and ϵ is chosen based on the target call blocking probability of the system. The DSRA mechanism can be carried out in the following 8 steps.

Step 1: For clarity of presentation, we focus our discussion in steps 1 - 3 on one network subscribers, and the same steps hold for subscribers of other networks. Consider video calls of service type v and class l in service area k. Let $\vec{\mathcal{T}}^{o}_{lvk}$ be a time vector of call arrival events for calls of service type v and service class l in service area k during period T_o. With a call arrival event at time instant $t^o_\pi \in \vec{\mathcal{T}}^{o}_{lvk}$, $\pi = \{1, 2, \ldots, \left|\vec{\mathcal{T}}^{o}_{lvk}\right|\}$, in period T_o, the number of calls at the time instant, $M_{lvk}(t^o_\pi)$, is used by the BSs/APs in the service area to probabilistically predict the number of calls at time instant $t^o_\pi + \tau$ in the next time period T_{o+1}. The predicted number is given by $\widetilde{M}_{lvk}(t^o_\pi + \tau)$. As the number of calls at t, $M_{lvk}(t)$, is a random variable, using the probability distribution of

$M_{lvk}(t_\pi^o + \tau)$ given $M_{lvk}(t_\pi^o)$, we can represent $\widetilde{M}_{lvk}(t_\pi^o + \tau)$ by a design parameter ϵ_{lvk}, such that

$$\Pr(M_{lvk}(t_\pi^o + \tau) > \widetilde{M}_{lvk}(t_\pi^o + \tau)|M_{lvk}(t_\pi^o)) \leq \epsilon_{lvk}, \quad \forall v, l \in \mathcal{L}, k \in \mathcal{K}. \tag{4.11}$$

Similar to ϵ_{lk}^n in Chapter 3, the design parameter $\epsilon_{lvk} \in [0,1]$ denotes the probability that $M_{lvk}(t_\pi^o + \tau)$ exceeds the predicted number $\widetilde{M}_{lvk}(t_\pi^o + \tau)$. The predicted number $\widetilde{M}_{lvk}(t_\pi^o + \tau)$ can be determined using the conditional PMF of $M_{lvk}(t_\pi^o + \tau)$ given $M_{lvk}(t_\pi^o)$, $f_{M_{lvk}(t_\pi^o+\tau)|M_{lvk}(t_\pi^o)}(m)$. Again, the transient distribution of the $M/G/\infty$ model [62] can be used to calculate $f_{M_{lvk}(t_\pi^o+\tau)|M_{lvk}(t_\pi^o)}(m)$, since call arrivals follow a Poisson process, the channel holding time follows a general distribution, and all calls are served simultaneously without queuing. We redefine ϕ_τ^{lkn} and θ_τ^{lkn} introduced in Chapter 3, under the assumption of stationary call arrival and departure processes:

- ϕ_τ^{lvk} - The probability that a call of service class l and service type v which is in service area k at time t_π^o is still present in the same service area at time $t_\pi^o + \tau$;
- θ_τ^{lvk} - The probability that a call of service class l and service type v that arrives in service area k during $(t_\pi^o, t_\pi^o + \tau]$ is still present at the same service area at time $t_\pi^o + \tau$;

while $\rho(\varkappa_1, \varkappa_2)$ and $\xi(\varkappa)$ are the same as in Chapter 3. At time instant t_π^o, given the number of calls, $M_{lvk}(t_\pi^o)$, we have [62]

$$M_{lvk}(t_\pi^o + \tau) =_d \rho(M_{lvk}(t_\pi^o), \phi_\tau^{lvk}) + \xi(\upsilon_{lvk}^n \tau \theta_\tau^{lvk}) \tag{4.12}$$

where υ_{lvk}^n denotes the arrival rate of new and handoff calls to network n in service area k. In oder to determine υ_{lvk}^n for BS/AP of network n, a BS/AP can count the number of its new call arrivals to service area k (excluding vertical handoff calls, since these calls are not arrivals to service area k) and divide it by the total elapsed time. In (4.12), the probabilities ϕ_τ^{lvk} and θ_τ^{lvk} are given by [62]

$$\phi_\tau^{lvk} = \frac{1}{E[T_h^{lvk}]} \int_\tau^\infty (1 - F_{T_h^{lvk}}(y))dy \tag{4.13}$$

$$\theta_\tau^{lvk} = \frac{E[T_h^{lvk}]}{\tau}(1 - \phi_\tau^{lvk}) \tag{4.14}$$

where $E[T_h^{lvk}]$ denotes the average channel holding time which can be calculated similar to (3.7). From (4.12), $f_{M_{lvk}(t_\pi^o+\tau)|M_{lvk}(t_\pi^o)}(m)$ can be found, and hence $\widetilde{M}_{lvk}(t_\pi^o + \tau)$ can be calculated using (4.11) as the minimum integer which satisfies

$$\sum_{m=0}^{\widetilde{M}_{lvk}(t_\pi^o+\tau)} f_{M_{lvk}(t_\pi^o+\tau)|M_{lvk}(t_\pi^o)}(m) \geq (1 - \epsilon_{lvk}), \quad \forall v, l \in \mathcal{L}, k \in \mathcal{K}. \tag{4.15}$$

Step 2: Each BS/AP in service area k records the predicted values of $\widetilde{M}_{lvk}(t_\pi^o + \tau)$, $\forall v, l \in \mathcal{L}, k \in \mathcal{K}$ and $\pi = \{1, 2, \ldots, \left|\vec{\mathcal{T}}_{lvk}^o\right|\}$, in a vector $\vec{\mathcal{M}}_{lvk}^{o+1}$.

Step 3: At the beginning of period T_{o+1}, the maximum predicted number of calls of each service type v and service class l in each service area k during T_{o+1}, $\widetilde{M}_{lvk}(T_{o+1})$, can be found using $\vec{\mathcal{M}}_{lvk}^{o+1}$. That is, $\widetilde{M}_{lvk}(T_{o+1}) = \max(\vec{\mathcal{M}}_{lvk}^{o+1})$ if it is less than or equal to C_{lvk}, otherwise $\widetilde{M}_{lvk}(T_{o+1}) = C_{lvk}$. This guarantees that for $\widetilde{M}_{lvk}(T_{o+1}) \leq C_{lvk}$ we have

$$\begin{aligned} &\Pr(M_{lvk}(t_\pi^{o+1}) > \widetilde{M}_{lvk}(T_{o+1})) \leq \epsilon_{lvk}, \\ &\qquad \forall v, l \in \mathcal{L}, k \in \mathcal{K}, \pi \in \{1, 2, \ldots, \left|\vec{\mathcal{T}}_{lvk}^{o+1}\right|\}. \end{aligned} \tag{4.16}$$

Step 4: The cooperating BSs/APs in the geographical region exchange their information regarding $\widetilde{M}_{lvk}(T_{o+1})$ $\forall v, l \in \mathcal{L}, k \in \mathcal{K}$ for all subscribers. As a result, $\mathcal{M}_{lvk}$ can be determined and hence problem (4.10) can be solved at each BS/AP so as to determine the binary assignment variable w_{nms}^{o+1} for all MTs with single-network service in the geographical region during T_{o+1} and the corresponding bandwidth allocation matrix B^{o+1}. Therefore, the network assignment vector A^{o+1} for single-network MTs during T_{o+1} can be determined. Based on the network assignment vector A^{o+1}, each BS/AP s can determine the maximum number of MTs with single-network calls and service class l in service area k which can be supported by this BS/AP during T_{o+1}, f_{lks}^{o+1}, $\forall l \in \mathcal{L}, k \in \mathcal{K}$, given B^{o+1}.

Step 5: Given the network assignment vector A^{o+1}, during T_{o+1}, calculated in step 4, problem (4.8) is reduced to

$$\begin{aligned} \max_{B \geq 0} \quad & U \\ s.t. \quad & (4.5) - (4.7). \end{aligned} \tag{4.17}$$

Problem (4.17) is a convex optimization problem, on which full dual decomposition can be applied (this helps in the decentralized resource allocation as described in the next step). As in Chapter 2, in order to apply full dual decomposition, we first find the Lagrangian function, $L(B, \lambda, \nu^{(1)}, \nu^{(2)}, \mu^{(1)}, \mu^{(2)})$, of (4.17), where $\lambda = (\lambda_{ns} : n \in \mathcal{N}, s \in \mathcal{S}_n)$ is defined to be a matrix of Lagrangian multipliers corresponding to capacity constraint (4.5), and $\lambda_{ns} \geq 0$, $\nu^{(1)} = (\nu_m^{(1)} : m \in \mathcal{M}_{1k}, \forall k \in \mathcal{K})$ and $\nu^{(2)} = (\nu_m^{(2)} : m \in \mathcal{M}_{1k}, \forall k \in \mathcal{K})$ are vectors of Lagrangian multipliers corresponding to the maximum and minimum required bandwidth constraints of (4.6) for MTs with single-network service and $\nu_m^{(1)}, \nu_m^{(2)} \geq 0$, and $\mu^{(1)} = (\mu_m^{(1)} : m \in \mathcal{M}_{2k}, \forall k \in \mathcal{K})$ and $\mu^{(2)} = (\mu_m^{(2)} : m \in \mathcal{M}_{2k}, \forall k \in \mathcal{K})$ are vectors of Lagrangian multipliers corresponding to the required bandwidth constraints for MTs with multi-homing service (4.7) and $\mu_m^{(1)}, \mu_m^{(2)} \geq 0$. The dual function then is given by

$$H(\lambda, \nu^{(1)}, \nu^{(2)}, \mu^{(1)}, \mu^{(2)}) = \max_{B \geq 0} L(B, \lambda, \nu^{(1)}, \nu^{(2)}, \mu^{(1)}, \mu^{(2)}) \tag{4.18}$$

and the dual problem corresponding to the primal problem of (4.17) is given by

$$\min_{(\lambda, \nu^{(1)}, \nu^{(2)}, \mu^{(1)}, \mu^{(2)}) \geq 0} H(\lambda, \nu^{(1)}, \nu^{(2)}, \mu^{(1)}, \mu^{(2)}). \tag{4.19}$$

The maximization problem (4.18) gives the bandwidth allocation matrix B for fixed value of the Lagrangian multipliers, which can be solved using the KKT conditions, and hence we have

$$b_{nms} = [(\frac{\eta_1}{\lambda_{ns} + (\nu_m^{(1)} - \nu_m^{(2)}) + \eta_2(1 - \omega_{nms})} - 1)/\eta_1]^+, \quad \forall m \in \bigcup_k \mathcal{M}_{1k} \tag{4.20}$$

$$b_{nms} = [(\frac{\eta_1}{\lambda_{ns} + (\mu_m^{(1)} - \mu_m^{(2)}) + \eta_2(1 - \omega_{nms})} - 1)/\eta_1]^+, \quad \forall m \in \bigcup_k \mathcal{M}_{2k} \tag{4.21}$$

The optimal values of the Lagrangian multipliers which result in the optimal bandwidth allocation can be found by solving the dual problem of (4.19). For a differentiable dual function, a gradient descent method can be applied to determine the optimum values for the Lagrangian multipliers, which results in

$$\lambda_{ns}(j+1) = [\lambda_{ns}(j) - \alpha_1(C_n - \sum_{m \in \mathcal{M}_{ns}} b_{nms}(j))]^+ \tag{4.22}$$

$$\nu_m^{(1)}(j+1) = [\nu_m^{(1)}(j) - \alpha_2(B_m^{\max} - b_{nms}(j))]^+ \tag{4.23}$$

$$\nu_m^{(2)}(j+1) = [\nu_m^{(2)}(j) - \alpha_3(b_{nms}(j) - B_m^{\min})]^+ \tag{4.24}$$

$$\mu_m^{(1)}(j+1) = [\mu_m^{(1)}(j) - \alpha_4(B_m^{\max} - \sum_{n=1}^{N} \sum_{s=1}^{S_n} b_{nms}(j))]^+ \tag{4.25}$$

$$\mu_m^{(2)}(j+1) = [\mu_m^{(2)}(j) - \alpha_5(\sum_{n=1}^{N} \sum_{s=1}^{S_n} b_{nms}(j) - B_m^{\min})]^+ \tag{4.26}$$

where j is the iteration index and α_e with $e = \{1, \dots, 5\}$ is a fixed sufficiently small step size. Convergence towards the optimal solution is guaranteed as the gradient of (4.19) satisfies the Lipchitz continuity condition.

As in Chapters 2 and 3, λ_{ns} is a link access price that is used as an indication of the capacity limitation experienced by each network BS/AP, while $\mu_m^{(1)}$ and $\mu_m^{(2)}$ are used by MTs with multi-homing calls to guarantee that the total bandwidth allocated from all BSs/APs satisfy the call total required bandwidth. On the other hand, $\nu_m^{(1)}$ and $\nu_m^{(2)}$ are used by MTs with single-network calls to guarantee that the bandwidth allocated from the assigned network satisfies the call required bandwidth.

Given the predicted maximum number of calls during T_{o+1}, $\widetilde{M}_{lvk}(T_{o+1})$ $\forall v, l \in \mathcal{L}, k \in \mathcal{K}, n \in \mathcal{N}$, each BS/AP can determine its predicted link access price value $\widetilde{\lambda}_{ns}^{o+1}$ using the BARON solver while solving (4.10) at the beginning of T_{o+1} using $\widetilde{M}_{lvk}(T_{o+1})$.

Step 6: At the beginning of T_{o+1}, each BS/AP updates its link access price value with $\widetilde{\lambda}_{ns}^{o+1}$ and this value is fixed over T_{o+1}, independent of call arrivals to and departures from different service areas, and is broadcasted on the BS/AP ID beacon. In addition, a flag bit, fb_{lks}, is set to 1 if $M_{lvk} < f_{lks}$, for $v = 1$, and is broadcasted by each BS/AP s on its ID beacon to denote that a new incoming call from subscribers of a given network with single-network service and service class l in service area k can be admitted by the BS/AP. Otherwise, $fb_{lks} = 0$.

The fixed link access price values, $\widetilde{\lambda}_{ns}^{o+1}$ $\forall n \in \mathcal{N}, s \in \mathcal{S}_n$ which are broadcasted during T_{o+1}, distribute the radio resources of all networks exactly over the maximum predicted number of calls $\widetilde{M}_{lvk}(T_{o+1})$ $\forall v, l \in \mathcal{L}, k \in \mathcal{K}$. Hence, during T_{o+1}, when $M_{lvk} = \widetilde{M}_{lvk}(T_{o+1})$, any incoming call from subscribers of a given network with service type v and service class l in service area k will be blocked. Hence, similar to ϵ_{lk}^{n}, from (4.11), ϵ_{lvk} is the upper bound of the call blocking probability for subscribers of a given network, given that $\widetilde{M}_{lvk}(T_{o+1}) \leq C_{lvk}$. Otherwise, $\widetilde{M}_{lvk}(T_{o+1}) = C_{lvk}$, and both the CORA and DSRA mechanisms achieve the same call blocking probability.

Step 7: An incoming MT to service area k during T_{o+1} listens to the link access price values $\widetilde{\lambda}_{ns}^{o+1}$ $\forall n \in \mathcal{N}, s \in \mathcal{S}_n$ using its multiple radio interfaces. Based on its service type, the MT then performs the following.

First, consider MTs with single-network service. An MT, $m \in \mathcal{M}_{1k}$, uses the link access price values to solve for the allocated bandwidth from BS/AP available at its location with $fb_{lks} = 1$. This can be done at MT, m, with a call from service class l in service area k, using the mechanism in Table 4.2, where J denotes the number of iterations required for the mechanism to converge to the required bandwidth allocation.

Table 4.2: Calculation of bandwidth allocation from each available network BS/AP at MT m with single-network service.

```
**Input:** $\widetilde{\lambda}_{ns}^{o+1}\ \forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$, $B_m$, $m \in \mathcal{M}$;
**Initialization**: $\nu_m^{(1)}(1) \geq 0$; $\nu_m^{(2)}(1) \geq 0$;
**for** $j = 1 : J$ **do**
    $b_{nms}(j) = [(\frac{\eta_1}{\widetilde{\lambda}_{ns}^{o+1} + (\nu_m^{(1)}(j) - \nu_m^{(2)}(j)) + \eta_2(1-\omega_{nms})} - 1)/\eta_1]^+$;
    $\nu_m^{(1)}(j+1) = [\nu_m^{(1)}(j) - \alpha_1(B_m^{\max} - b_{nms}(j))]^+$;
    $\nu_m^{(2)}(j+1) = [\nu_m^{(2)}(j) - \alpha_2(b_{nms}(j) - B_m^{\min})]^+$;
**end for**
**Output:** $b_{nms}$.
```

The MT asks the BS/AP for the b_{nms} resource allocation. The BS/AP provides the required bandwidth allocation if it has sufficient resources. Otherwise, the MT repeats the process with another BS/AP with $fb_{lks} = 1$. If no BS/AP in k can provide the MT with its required bandwidth, the call is blocked. For MTs which are already in service, the link access price values $\widetilde{\lambda}_{ns}^{o+1}\ \forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$ with $fb_{lks} = 1$, are used at the beginning of T_{o+1} in a similar way as described before in order to perform a vertical handover if necessary.

Next, consider MTs with multi-homing services. During T_{o+1}, each MT in the geographical region, including both incoming and existing ones, uses the broadcasted link access price values received at its location to determine the required bandwidth share from each available BS/AP, such that the total amount of allocated resources from all the BSs/APs satisfies its required bandwidth. This is performed at MT, m, with service class l in service area k using the mechanism in Table 4.3. The MT then asks for the required bandwidth share b_{nms} from BS/AP s of network $n\ \forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$, which al-

Table 4.3: Calculation of bandwidth share from each available network BS/AP at MT m with multi-homing service.

```
Input: $\widetilde{\lambda}_{ns}^{o+1}\ \forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk},\ B_m,\ m \in \mathcal{M}$;
Initialization: $\mu_m^{(1)}(1) \geq 0$; $\mu_m^{(2)}(1) \geq 0$;
for $j = 1 : J$ do
    for $n \in \mathcal{N}_k$ do
        for $s \in \mathcal{S}_{nk}$ do
            $b_{nms}(j) = [(\frac{\eta_1}{\widetilde{\lambda}_{ns}^{o+1}+(\mu_m^{(1)}(j)-\mu_m^{(2)}(j))+\eta_2(1-\omega_{nms})} - 1)/\eta_1]^{+}$;
        end for
    end for
    $\mu_m^{(1)}(j+1) = [\mu_m^{(1)}(j) - \alpha_3(B_m^{\max} - \sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms}(j))]^{+}$;
    $\mu_m^{(2)}(j+1) = [\mu_m^{(2)}(j) - \alpha_4(\sum_{n=1}^{N}\sum_{s=1}^{S_n} b_{nms}(j) - B_m^{\min})]^{+}$;
end for
Output: The required $b_{nms}\ \forall n \in \mathcal{N}_k, s \in \mathcal{S}_{nk}$.
```

locates the required bandwidth if it has sufficient resources. The incoming call is blocked if the total allocated resources from all BSs/APs do not satisfy its required bandwidth.

Step 8: Each MT reports to its serving BSs/APs its home network, service type, service class, and a list of the BS/AP IDs that the MT can receive signal from. This information is used by BSs/APs to predict $\widetilde{M}_{lvk}(T_{o+2})\ \forall v, l \in \mathcal{L}, k \in \mathcal{K}$ for every network subscribers, during the next period T_{o+2} in order to update their link access price values at the beginning of T_{o+2}.

The DSRA procedure can be illustrated using a figure similar to Figure 3.2 for the PBRA mechanism. The main difference between the PBRA and DSRA operations is that the PBRA can support only multi-homing calls, while the DSRA deals with the simultaneous presence of both single-network and multi-homing calls.

As in the PBRA mechanism, the link access price value for BSs/APs of different networks are updated every τ which should reflect some change in the call traffic load in the geographical region. Let δ_{lvk} be the minimum of durations to the arrival of a new call and to the departure of an existing call for service class l with service type v in service area k for subscribers of a given network. Define $\delta=\min(\delta_{lvk})$ $\forall l, v, k$ and subscribers of different networks. Thus, as a guideline, the time duration τ is chosen such that the probability $\Pr[\delta > \tau]$ is less than a small threshold.

The DSRA mechanism can be implemented using a look-up table stored at BSs/APs in a way similar to the implementation of the PBRA mechanism, as discussed in Chapter 3.

4.5 Simulation Results and Discussion

This section presents simulation results for the radio resource allocation problem in a heterogeneous wireless access medium for MTs with single-network and multi-homing services. Consider the geographical region given in Figure 2.4. A single service class ($l = 1$) is considered for each service type v (single-network and multi-homing) and we study the performance of the proposed mechanisms in the service area that is covered by the WiMAX and cellular network BSs ($k = 2$) in terms of the allocated bandwidth per call and the call blocking probability. As a proof of concept, we only show the results of resource allocation for the cellular network subscribers. For simplicity, we consider a complete partitioning strategy for each network BS transmission capacity [75], where the total capacity of each BS is divided into two separate parts, dedicating to single-network and multi-homing services respectively[2]. The allocated transmission capacity from network n BS/AP to the service area under consideration for cellular network subscribers with service type v, C_{nv}, is given by $C_{11} = 1.344$ Mbps, $C_{12} = 2.864$ Mbps, $C_{21} = 0.576$ Mbps, and $C_{22} = 2$ Mbps. The C_{nv} values can support a total of 30 VBR calls with required bandwidth allocation $[0.064, 0.128]$ Mbps for single-network MTs, i.e. $C_{112} = 30$, and 19 VBR calls with required bandwidth allocation $[0.256, 0.512]$ Mbps for multi-homing MTs, i.e. $C_{122} = 19$. The arrival process of new and handoff calls to the service area under consideration is modeled as a Poisson process with parameter υ_{112} (call/minute) for single-network MTs and υ_{122} (call/minute) for multi-homing MTs. The video call duration is modeled by a two-stage hyper-exponential distribution with the PDF given in (3.1) and $\varsigma_{1v} = 1.5$. The average call duration for single-network MTs $\bar{T}_c^{11}$ is 15 minutes and for multi-homing MTs $\bar{T}_c^{12}$ is 10 minutes. The user residence time in the service area under consideration follows an exponential distribution with an average duration $\bar{T}_r = 20$ minutes [42]. The parameters η_1 and η_2 are both set to 1 [43]. The WiMAX and cellular networks set different costs on their resources using the priority parameter $\omega_{1m1} = 0.8$, $\omega_{2m1} = 0.6$ for network users, while $\omega_{nms} = 1$ for network subscribers [12]. The GDXMRW utilities [74] are used to create an interface between GAMS and MATLAB to make use of the BARON solver of GAMS in solving the optimization problem of (4.10) while using the MATLAB simulation and visualization tools.

A. Performance Comparison

In the following, the performance of the DSRA mechanism is compared to the CORA mechanism. While it is not appropriate for practical implementation when different networks are operated by different service providers, the CORA mechanism is used as a

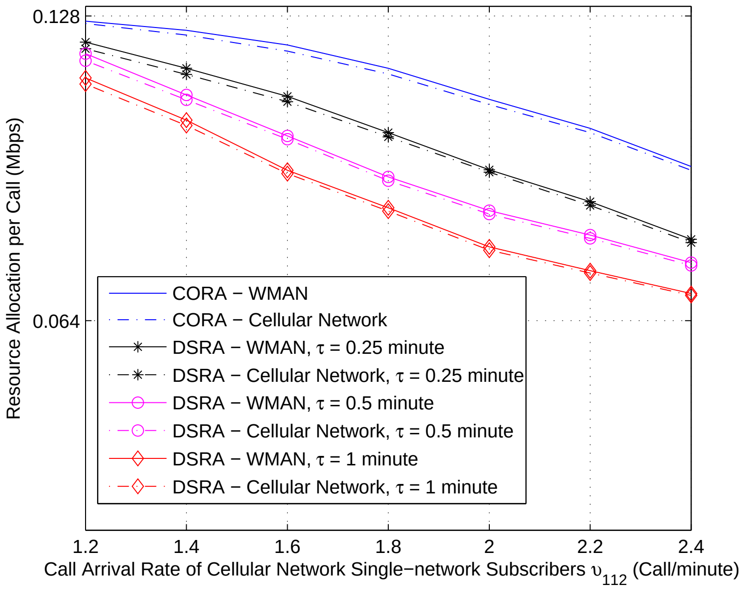
[2] The numerical results in Section 4.3 investigates a complete sharing strategy for each BS/AP transmission capacity [75] where both service types can occupy up to the total capacity of each BS/AP.

performance bound for the allocated bandwidth per call and the call blocking probability. In the simulation, we set the upper bounds on call blocking probability ϵ_{112}, ϵ_{122} to 1% and the prediction duration τ to 0.25, 0.5, and 1 minute. We only show the results for single-network service and similar observations hold for multi-homing service.

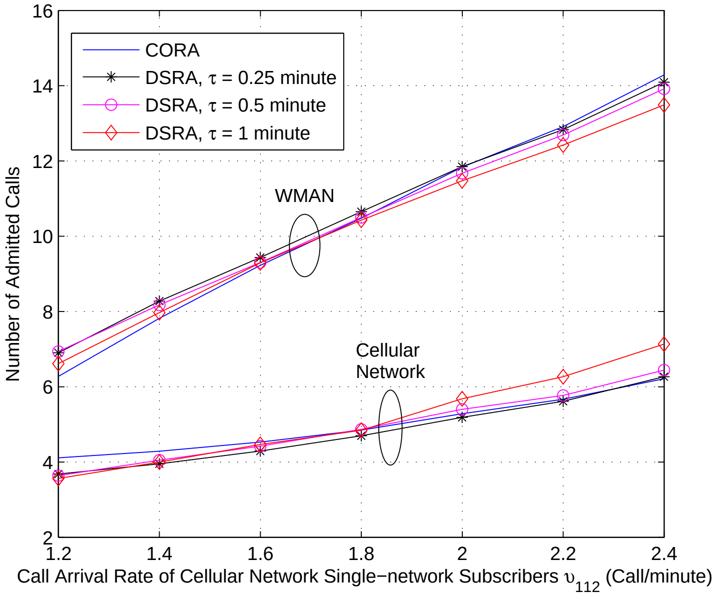
Figure 4.6 shows performance comparison between the DSRA and CORA mechanism for MTs with single-network service versus the call arrival rate υ_{112}. Figure 4.4a shows the bandwidth allocation per call for MTs assigned to the WiMAX and MTs assigned to the cellular network. At a low call arrival rate, the predicted number of simultaneously present calls is low, which results in a high allocated bandwidth per call using the DSRA mechanism for different τ values. At a high call arrival rate, the predicted number of simultaneously present users is high, as a result less bandwidth is allocated to each call. Furthermore, less bandwidth is allocated per call for larger values of τ as explained in the next sub-section. Figure 4.4b shows that more MTs with single-network service are assigned to the WiMAX BS as compared to the cellular network BS due to the WiMAX BS larger capacity C_{11}. In Figure 4.4c, using the CORA mechanism, there is no call blocking probability for $\upsilon_{112} < 1.6$ call/minute. For call arrival rate $\upsilon_{112} < 2.2$ call/minute, the DSRA mechanism does not exceed the target upper bound on call blocking probability of 1%. For call arrival rate $\upsilon_{112} \geq 2.2$ call/minute, the predicted number of calls simultaneously present in the service area under consideration is larger than C_{112}. Hence, according to the DSRA mechanism, the predicted number of calls is made equal to C_{112}, and both the DSRA and the CORA mechanisms achieve the same call blocking probability.

B. Performance of The DSRA Mechanism

In the following, we study the performance of the DSRA mechanism versus its two design parameters, namely the upper bound on call blocking probability ϵ_{lvk} and the prediction duration τ. We only show the results for multi-homing service and the same

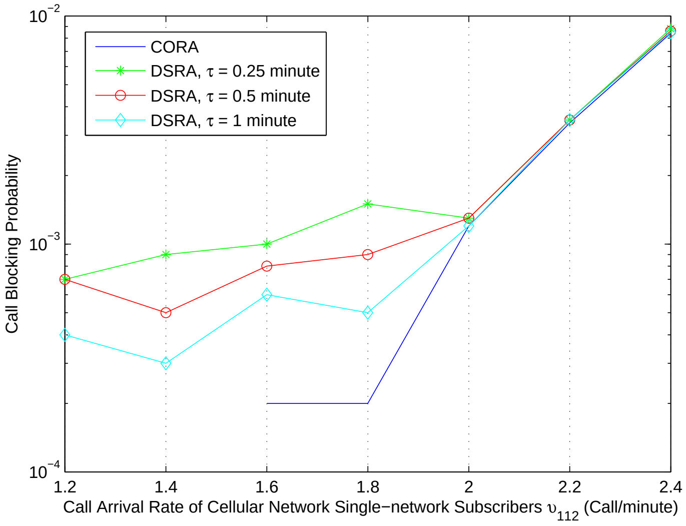


(a)



(b)

Figure 4.4: Performance comparison for single-network service: (a) Bandwidth allocation per call; (b) Number of admitted calls; (c) Call blocking probability.



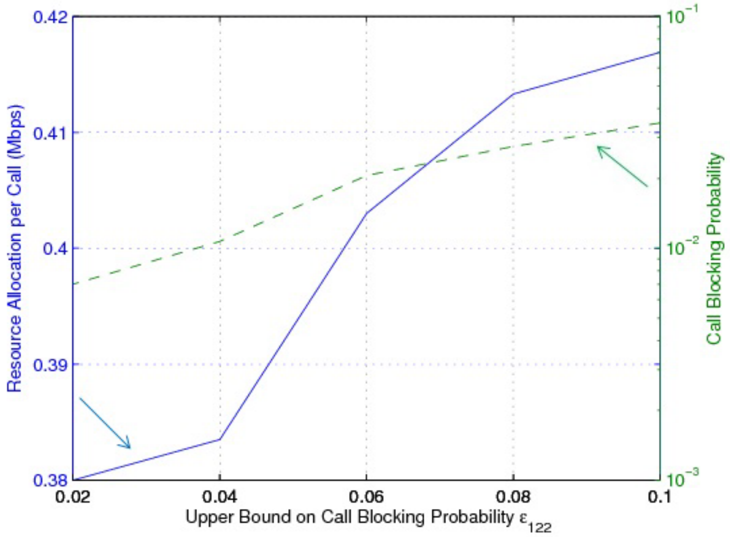
(c)

Figure 4.4: Cont. Performance comparison for single-network service: (a) Bandwidth allocation per call; (b) Number of admitted calls; (c) Call blocking probability.

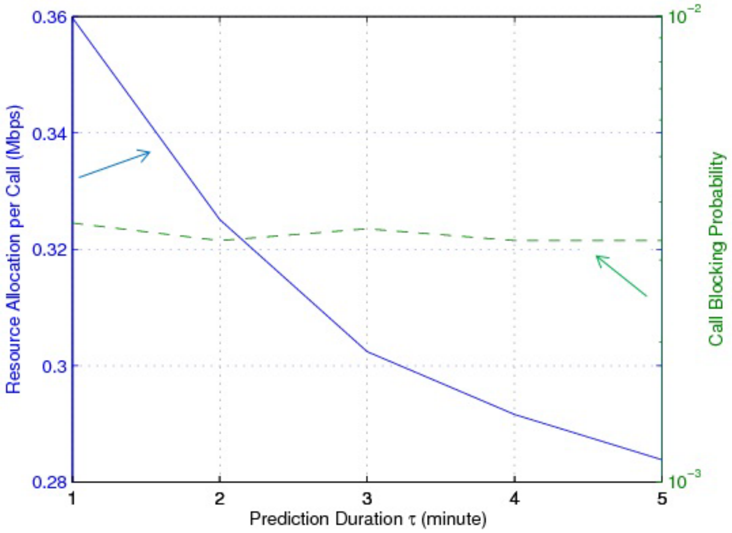
observations hold for single-network service.

Figure 4.5a plots the performance of the DSRA mechanism in terms of the amount of allocated bandwidth per call and call blocking probability versus ϵ_{122}, with call arrival rate $\upsilon_{122} = 1.4$ call/minute and $\tau = 1$ minute. A small value of ϵ_{122} results in a low call blocking probability. However, this corresponds to a large number of predicted calls (and hence large BS/AP link access price values), which results in a small amount of bandwidth allocation per call. On the other hand, a large value of ϵ_{122} results in a high call blocking probability and a large amount of bandwidth allocation per call. Overall, the call blocking probability does not exceed its upper bound $\epsilon_{122} = 1\%$. The upper bound ϵ_{122} should be chosen to balance the trade-off between the allocated bandwidth per call and the call blocking probability.

Figure 4.5b investigates the performance of the DSRA mechanism in terms of the



(a)



(b)

Figure 4.5: The DSRA mechanism performance versus: (a) ϵ_{122}; (b) τ.

amount of allocated bandwidth per call and call blocking probability versus the prediction duration τ, with $\upsilon = 1.4$ call/minute and $\epsilon_{122} = 1\%$. As τ increases, the DSRA mechanism updates the BS/AP link access price less frequently and hence a larger number of simultaneously present calls is predicted. As a result, the allocated bandwidth per call is reduced. Also, simulation results indicate that the call blocking probability does not exceed its target upper bound $\epsilon_{122} = 1\%$.

4.6 Summary

In this chapter, a decentralized resource allocation mechanism is proposed for a heterogeneous wireless access medium to support MTs with single-network and multi-homing services. The mechanism gives MTs an active role in the resource allocation operation, such that an MT with single-network service can select the best wireless network available at its location and asks for its required bandwidth, while an MT with multi-homing service can determine the required bandwidth share from each network in order to satisfy its total required bandwidth. The resource allocation relies on concepts of short-term call traffic prediction and network cooperation in order to perform the decentralized resource allocation in an efficient manner. The mechanism has two design parameters, namely ϵ_{lvk} and τ, which should be properly chosen to strike a balance between the desired performance in terms of the allocated bandwidth per call and the call blocking probability, and between the performance and implementation complexity.

Chapter 5

Energy and Content Aware Multi-homing Video Transmission

Multi-homing video transmission can improve the perceived video quality in many aspects. However, due to the MT battery energy limitation, an energy management mechanism is required in order to efficiently utilize the MT available energy to support video transmission. How to devise such an energy management mechanism while considering the characteristics of video packet-level traffic and the heterogeneous wireless access medium needs investigation. In this chapter, we propose an energy and content aware video transmission mechanism using a multi-homing service in a heterogeneous wireless access medium. The proposed mechanism takes account of the energy limitation of MTs and the required QoS for video streaming applications, and utilizes the available opportunities in the heterogeneous wireless access medium.

5.1 Related Work

In literature, several works have studied how to achieve high video quality with low power consumption. In these works, the main objective is to design energy efficient video packet scheduling mechanisms. Two categories of video packet scheduling mechanisms can be distinguished. The first category includes single path video transmission techniques, while the second category includes video transmission over multiple network paths.

The main objective of single-path video packet scheduling is to schedule packet transmission such that packets do not miss their playback deadlines. Packets whose playback deadlines have passed are dropped so as not to waste network resources. The scheduling policy should incorporate the video packet characteristics (in terms of delay deadline and distortion impact) and the time varying wireless channel condition. In [76], the problem of video packet scheduling is studied for multiple users in the downlink of a wireless communication system. A playout adaptive packet scheduling algorithm is proposed in [77] for video delivery over wireless networks. A cross layer video packet scheduling scheme is presented in [78], which targets downlink transmission. In [19], a Markov decision process (MDP) is used to formulate the video packet scheduling problem and balance the packet distortion impact with the consumed energy. One limitation of extending an MDP formulation to a multi-homing scenario is the curse of dimensionality as the state space and actions will suffer from an exponential growth as a function of the number of the available networks. The energy budget effect is considered in the packet scheduling framework of [20] which aims to maximize the perceived video quality through a joint optimization scheme of modulation and coding, and transmission power allocation. The problem of joint packet scheduling and power allocation is also investigated in [6] in order to minimize video quality distortion for multiple users in the uplink of a CDMA network. As these works target single-path video transmission, they do not benefit from the multi-homing video transmission advantages.

Several works in literature have studied packet scheduling for multi-path video streaming. In [79], a multi-path transmission control scheme is proposed, combining bandwidth aggregation and packet scheduling for real time streaming in a multi-path environment. The streaming policy of [80] consists of a joint selection of the network path and of the video packets to be transmitted along with their sending times. Almost all the multi-path video transmission policies discussed in literature do not target a heterogeneous wireless access medium. Instead, for multi-path video transmission policies in literature, all the used paths belong to the same network such as a mobile ad hoc network. As a result, when energy efficiency is considered, as in [81] and [82], the objective of packet scheduling is to avoid paths along which nodes are suffering from energy depletion. When energy efficiency is considered in a heterogeneous wireless access medium, one objective is to exploit the available bandwidth and channel conditions experienced by different radio interfaces of an MT in order to support a long duration video transmission with acceptable quality subject to the MT battery energy constraint.

Video streaming in a heterogeneous wireless access medium is studied in [83]. The objective is to investigate the heterogeneous networking attributes that may affect the streaming performance, in terms of the trade-off between jitter frequency and buffer delay. Yet, the work in [83] does not target a multi-homing service and the MT connects only to one wireless access network at a time. The work of [84] studies video transmission in a heterogeneous wireless access medium and employs multi-homing service in downlink transmission. Hence, these works do not investigate how to exploit the channel conditions and available bandwidths at different networks to support uplink multi-homing video transmission, while considering the MT battery energy limitation.

In this chapter, we aim to develop an energy and content aware mechanism for multi-homing video transmission in a heterogeneous wireless access medium. The objective is to perform power allocation and packet scheduling to different radio interfaces of an MT, subjected to the MT battery energy limitation, in order to satisfy the packet required QoS in terms of playback deadline and to minimize video quality distortion.

5.2 System Model

For wireless multi-homing video transmission, we focus on single MT. Hence, the subscript m is omitted. In addition, the video call has a target duration T_c.

A. Video Packet Traffic Model

The video sequence is encoded into a bit stream using a layered/scalable video encoder [85]. The layered representation of the video sequence is composed of a base layer and several enhancement layers [86]. The base layer, which can be decoded independently of the enhancement layers, provides a basic level of video quality. The decoding of enhancement layers is based on the base layer and serve to improve the base layer quality. Each video layer is periodically encoded using a group-of-picture (GoP) structure. Time is partitioned into time slots, $\mathcal{T} = \{1, 2, \ldots, T\}$, of equal duration $\widetilde{\tau}$, $T = \lceil \frac{T_c}{\widetilde{\tau}} \rceil$. Every time slot, the MT has a new GoP, from different layers, ready for transmission. Each time slot has $\mathcal{F}$ frames from different layers, $\mathcal{F} = \{1, 2, \ldots, F\}$, and each frame can be of I, P, or B type. I Frames are compressed versions of raw frames independent of other frames. P frames only refer to preceding I/P frames, while B frames can refer to both preceding and succeeding frames. The data within one time slot are encoded interdependently through motion estimation, while data belonging to different time slots are encoded independently [19]. A video frame has the following characteristics [19]:

- Size - Each frame f is encoded into packets and each packet contains data relative to at most one frame [80]. Frame f is fragmented into G_f packets, $G_f \in [1, G_f^{\max}]$, where $G_f^{\max}$ denotes the maximum allowable size for frame f at each GoP. The frame size (in number of video packets, G_f) is represented by an independent identically distributed (i.i.d.) random variable that follows a PMF $f_{G_f}(g_f)$ [19]. The frame size across different GoPs follows the same PMF given the frame type (I, P, or B). The PMF, $f_{G_f}(g_f)$, can be calculated for different video contents and frame types using the approach in [87]. The frame size, G_f, for frames of I, P, or B types is constant within one time slot[1] and varies from one time slot to another. The packet size (in bits) for frame f is denoted by l_f.

- Distortion Impact - Each frame, f, has a distortion impact value per packet, i_f. It represents the amount by which video distortion is reduced if this packet is received, on time, at the decoder side. The packet distortion impact value, i_f, for different video contents and frame types can be calculated as discussed in [88].

- Delay Deadline - It represents the time by which the frame should be decoded at the destination, which is also known as decoding time stamp [6]. Packets that belong to the same frame have the same delay deadline, which is denoted by d_f. Since videos are encoded using a fixed number of frames per second (fps) within the same layer, the difference in the delay deadline between any two consecutive frames within the layer is constant [6]. The delay difference is given by $|d_{f+1} - d_f| = \Delta D_{f+1,f}$. The transmission deadlines of all packets within a given GoP expire by $\widetilde{\tau}$. It is assumed that the delay at the MT dominates the end-to-end delay.

- Dependence - Within each time slot, since some frames are encoded based on the prediction of other frames, there are dependences among these frames. Hence, packet decoding of one frame depends on the successful decoding of packets from

[1] The assumption of constant frame size within the same frame type in one time slot is adopted for clarity of presentation. However, the proposed energy management mechanism is not limited by this assumption and the extension to a general case is straightforward.

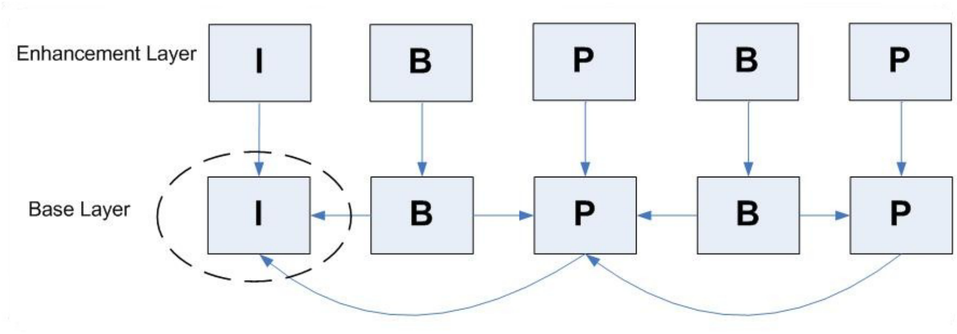


Figure 5.1: GoP structure with frames dependences [80].

other frames. These dependences among packets of different frames, within one time slot, are expressed using a directed acyclic graph (DAG) [19], as shown in Figure 5.1. For instance, the circled I frame, in Figure 5.1, is an ancestor for the first B and P frames in the base layer and the I frame in the enhancement layer. Hence, each video packet z_f is said to have ancestors $\mathcal{Z}_z^f$. Packets which belong to $\mathcal{Z}_z^f$ $\forall f \in \mathcal{F}$ have higher distortion impact and smaller delay deadline than packet z_f.

B. Video Transmission Model

Consider an uplink video transmission from an MT for posting on social network sites [6]. The MT, using its multi-homing capability, establishes communications with multiple wireless networks simultaneously and employes them for video packet transmission. Let the set of MT radio interfaces used for video transmission denoted by $\mathcal{U}$. The uplink bandwidth allocated to the MT for radio interface $u \in \mathcal{U}$ is denoted by b_u, which is assumed to be constant over the call duration. Let γ_u denote the received signal-to-noise ratio (SNR) at the BS/AP communicating with radio interface u. The received SNR value, γ_u, $\forall u \in \mathcal{U}$, is constant within one time slot and varies independently from one time slot to another [6].

For each time slot, let x^f_{zu} denote a video packet scheduling decision, where $x^f_{zu} = 1$ if packet z of frame f is assigned to radio interface u, otherwise $x^f_{zu} = 0$, and p_u is the instantaneous power allocated by the MT to radio interface u. The average power allocated to radio interface u is denoted by $\bar{p}_u$. The MT available energy at the beginning of the call is denoted by E. It is assumed that the MT always has packets ready for transmission and hence we do not put its radio interfaces in the sleep mode for energy saving.

5.3 Problem Formulation

In this section, the problem formulation for energy and content aware multi-homing video transmission is discussed in a heterogeneous wireless access medium. The objective is to minimize the video quality distortion, on a time slot level [6], subject to the MT energy constraint, through optimizing the power allocation to each radio interface and scheduling the most valuable video packets (packets with highest distortion impact) for transmission, while dropping the remaining ones if necessary. First, we formulate the problem as an MINLP which can be computationally intractable for a large-size problem. Then, we employ a piecewise linearization approach and solve the problem using a cutting plane method in order to reduce the associated complexity from MINLP to a series of MIPs.

A. MINLP Problem Formulation

The optimization framework aims to minimize the distortion in the perceived video quality under the MT battery energy limitation. The minimization of video quality distortion can be achieved through scheduling video packets with high distortion impact [19, 80] to the available multiple radio interfaces. This is given by

$$I = \sum_{\mathcal{U}} \sum_{z_f, f \in \mathcal{F}} i_f x^f_{zu}. \tag{5.1}$$

As packets that belong to the same frame have the same delay deadline of the frame, the required minimum data rate to transmit a video packet z_f, $\forall f \in \mathcal{F}$, is given by $r(z_f) = l_f/\Delta D_{f+1,f}$ [6]. The overall required data rate for packet transmission using radio interface $u \in \mathcal{U}$ should satisfy the achieved data rate over this radio interface, which is given by

$$\sum_{z_f, f \in \mathcal{F}} x_{zu}^{f} r(z_f) \leq b_u \log_2(1 + \frac{\Omega_u p_u}{\eta_0 b_u}), \quad \forall u \in \mathcal{U} \tag{5.2}$$

where Ω_u denotes the channel power gain between the MT and the BS/AP communicating with radio interface u and η_0 denotes the one-sided noise power spectral density. The right-hand-side of (5.2) is from Shannon formula. In a case that the required data rate to transmit all the video packets is larger than the overall achieved data rate using all the radio interfaces, given the MT battery energy limitation, video packets with less distortion impact have to be dropped.

The total allocated power to the MT different radio interfaces should satisfy its battery energy limitation expressed by the specified energy budget per time slot E_c. The energy budget per time slot E_c should vary from one time slot to another depending not only on the MT energy limitation but also the current channel conditions for different radio interfaces. However, in this chapter, we let E_c be fixed over $\mathcal{T}$ independent of the channel conditions. Hence, in this work, E_c is determined by dividing the MT available energy at the beginning of video transmission over the T time slots. Hence, we have

$$\sum_{\mathcal{U}} p_u \leq \frac{E_c}{\widetilde{\mathcal{T}}}. \tag{5.3}$$

Video packet scheduling should capture the dependence relationship among different packets. Video packets whose ancestors are not scheduled for transmission should not be transmitted as they will not be successfully decoded at destination and hence waste both the MT and network resources. This can be described using a precedence constraint given by

$$x_{zu}^{f} \leq x_{z'u'}^{f'}, \quad \forall z'_{f'} \in \mathcal{Z}_{z}^{f}, z_f \in \underset{f \in \mathcal{F}}{\cup} z_f, u, u' \in \mathcal{U}. \tag{5.4}$$

In addition, a video packet can be assigned to one and only one radio interface of the MT, which is expressed by

$$\sum_{\mathcal{U}} x_{zu}^{f} \leq 1, \quad \forall z_f, f \in \mathcal{F}. \tag{5.5}$$

Hence, the energy and content aware multi-homing video transmission problem is given by

$$\begin{aligned} \max_{x_{zu}^{f}, p_u} \quad & I \\ s.t. \quad & (5.2) - (5.5) \\ & x_{zu}^{f} \in \{0, 1\} \\ & p_u \geq 0. \end{aligned} \tag{5.6}$$

The optimization problem (5.6) should be solved at the beginning of every time slot $t \in \mathcal{T}$ with a new GoP from different layers. The problem formulation accounts for the video packet characteristics in terms of distortion impact, delay deadlines, and packet dependence relation, the characteristics of the multiple wireless interfaces in terms of the channel conditions and the allocated bandwidth, and the MT battery energy limitation. Problem (5.6) is an MINLP since it involves the optimization over real variables p_u and binary variables x_{zu}^{f}, and hence it is NP-hard [64, 89]. It can be computationally intractable to solve large instances of (5.6) (i.e., large number of video packets) in real-time, and hence in the following we aim to reduce the problem computational complexity.

B. Piecewise Linearization Approach

Let $\Lambda_u = \frac{\Omega_u}{\eta_0 b_u}$. The function $\log_2(1+\Lambda_u p_u)$ on the right-hand-side of (5.2) is a concave and continuous function that can be approximated with a set of piecewise linear functions using a first order Taylor expansion around points p_u^ϑ, $\vartheta \in \Theta$ [90], where Θ denotes a set of all points in the domain of the logarithmic function. Hence,

$$\log(1+\Lambda_u p_u) \approx \min_{\vartheta\in\Theta}\{\log(1+\Lambda_u p_u^\vartheta) + \frac{\Lambda_u(p_u - p_u^\vartheta)}{1+\Lambda_u p_u^\vartheta}\}. \tag{5.7}$$

Thus, (5.2) can be written as

$$\sum_{z_f, f\in\mathcal{F}} x_{zu}^f r(z_f) \leq \frac{b_u}{\log(2)}\{\log(1+\Lambda_u p_u^\vartheta) + \frac{\Lambda_u(p_u - p_u^\vartheta)}{1+\Lambda_u p_u^\vartheta}\}. \tag{5.8}$$

Rearranging (5.8) , we have

$$\sum_{z_f, f\in\mathcal{F}} x_{zu}^f r(z_f) - \frac{b_u\Lambda_u}{\log(2)(1+\Lambda_u p_u^\vartheta)} p_u \leq \frac{b_u}{\log(2)}\log(1+\Lambda_u p_u^\vartheta) - \frac{b_u\Lambda_u p_u^\vartheta}{\log(2)(1+\Lambda_u p_u^\vartheta)},$$
$$\forall u \in \mathcal{U}, \vartheta\in\Theta. \tag{5.9}$$

Hence, problem (5.6) can be re-written as

$$\begin{aligned} \max_{x_{zu}^f, p_u} \quad & I \\ s.t. \quad & (5.3)-(5.5), (5.9) \\ & x_{zu}^f \in \{0,1\} \\ & p_u \geq 0. \end{aligned} \tag{5.10}$$

The non-linearity of (5.6) is eliminated by adding a large number of constraints using (5.9). Hence, the problem complexity is reduced from MINLP to a linear MIP. Ideally, we need all points p_u^ϑ in the domain of $\log(1+\Lambda_u p_u^\vartheta)$, Θ, in order to approximate it. However, in order to find the optimal solution of (5.10), we only need an approximation of $\log(1+\Lambda_u p_u^\vartheta)$ around the optimal solution. Let $\widetilde{\Theta}$ denote a subset of Θ. A cutting plane/constraint generation approach is used to add the necessary constraints through (5.9). We start by an initial set of points p_u^ϑ with $\vartheta \in \widetilde{\Theta}$, and hence an initial set of constraints through (5.9), and the rest of points (constraints) are added as needed using the cutting plane algorithm [90], given in Table 5.1.

Table 5.1: Cutting plane algorithm.

```
**Initialization**: $p_u^{\vartheta}$, $\vartheta \in \widetilde{\Theta}$, $u \in \mathcal{U}$, $y_1 = 1$, $y_2 = 0$;
**while** $y_2 = 0$ **do**
    Solve (5.10), and denote its solution as $(\widetilde{x}_{zu}^{f}(y_1), \widetilde{p}_u(y_1))$;
    **if** $\widetilde{p}_u(y_1) \notin p_u^{\vartheta}\ \forall \vartheta \in \widetilde{\Theta}$ **then**
        Append new cut to (5.10) using $p_u^{\vartheta+1} = \widetilde{p}_u(y_1)$;
        $y_1 = y_1 + 1$;
    **else**
        $y_2 = 1$;
    **end if**
**end while**
**Output**: $\widetilde{x}_{zu}^{f}(y_1)\ \forall z_f$, $\forall f \in \mathcal{F}$, $\widetilde{p}_u(y_1)\ \forall u \in \mathcal{U}$.
```

In the algorithm, the linear approximation is done dynamically, solving for a better approximation at every iteration, until an optimal solution is found. It has been proven in [90] that the cutting plane algorithm is finite and thus converges to the optimal solution in a finite number of iterations. While the cutting plane algorithm significantly reduces the computational complexity of (5.6), especially for a large-size problem, we still need a powerful optimization solver to be available at the MT in order to solve (5.10), such as CPLEX [72], for the optimal power allocation and packet scheduling. As a result, in the next section, we aim to develop a greedy algorithm that has a performance very close to the optimal solution and require simple operations. We will use the cutting plane algorithm, in Table 5.1, to evaluate the performance of the proposed greedy algorithm.

5.4 Energy and Content Aware Multi-homing Video Transmission Mechanism

Intuitively, the video quality distortion is minimized if more video packets are transmitted and less are dropped. The higher the achieved data rates at the MT different radio interfaces, subject to the MT battery energy limitation, the more transmitted packets and thus the better video quality. Hence, we propose to decouple problem (5.6) into two sub-problems. The first sub-problem is to find the transmission power allocation for each radio interface that maximizes the achieved data rate, subject to the MT battery energy limitation. The second sub-problem is to schedule the most valuable video packets to different radio interfaces for transmission and drop the rest if necessary, given the transmission power allocation. The only difference between the exact problem solution using the cutting plane algorithm, in Table 5.1, and the approximate mechanism is that, the original MINLP performs joint power allocation and packet scheduling, while the proposed mechanism performs these two tasks separately. If the total number of used radio interfaces is $|\mathcal{U}|$, the exact solution can insert a maximum of $|\mathcal{U}| - 1$ additional packets more than the approximate mechanism, due to the joint optimization performed by the exact solution. Since it is not expected to use more than 2 to 3 radio interfaces, the number of additional inserted video packets is small as compared to the approximate solution. With a large number of video packets per time slot, the contribution of these additional packets to the achieved video quality is not significant. Hence, both exact and approximate solutions achieve very close results. This issue is further investigated in the numerical results of Section 5.5.

A. Transmission Power Allocation for Each Radio Interface

The power allocation strategy adapts to the channel conditions and available bandwidths at different radio interfaces in order to maximize the achieved data rate for different radio interfaces while satisfying the MT battery energy limitation. Hence, we solve

$$\begin{aligned} \max_{p_u} \quad & \sum_{\mathcal{U}} b_u \log_2(1+\Lambda_u p_u) \\ s.t. \quad & (5.3) \\ & p_u \geq 0. \end{aligned} \tag{5.11}$$

Problem (5.11) has a concave objective function and linear constraint. Hence, problem (5.11) is a convex optimization problem and can be solved efficiently in polynomial time [44]. Thus, strong duality holds for problem (5.11) and a local maximum is a global maximum as well [44]. The Lagrangian function of (5.11) is given as

$$L(p_u, \varphi) = \sum_{\mathcal{U}} b_u \log_2(1+\Lambda_u p_u) + \varphi(\frac{E_c}{\widetilde{\tau}} - \sum_{\mathcal{U}} p_u) \tag{5.12}$$

where φ is a Lagrangian multiplier that corresponds to the constraint of (5.3), with $\varphi \geq 0$. The dual function is given by

$$H(\varphi) = \max_{p_u \geq 0} L(p_u, \varphi) \tag{5.13}$$

and the dual problem of (5.11) is

$$\min_{\varphi \geq 0} H(\varphi). \tag{5.14}$$

The maximization problem of (5.13) can be written as

$$H(\varphi) = \sum_{\mathcal{U}} \max_{p_u \geq 0} \{b_u \log_2(1+\Lambda_u p_u) - \varphi p_u\}. \tag{5.15}$$

Hence, the optimal power allocation for each radio interface is obtained by solving

$$\max_{p_u \geq 0} \{b_u \log_2(1+\Lambda_u p_u) - \varphi p_u\}. \tag{5.16}$$

For a fixed value of φ, the allocated power p_u can be calculated for each radio interface by applying the KKT conditions on (5.16), which results in

$$p_u = \max\{\frac{b_u}{\varphi \ln(2)} - \frac{1}{\Lambda_u}, 0\}. \tag{5.17}$$

The optimal value of φ that results in the optimal allocated power p_u of (5.17) is determined by solving the dual problem of (5.14). The dual problem can be written as

$$\min_{\varphi \geq 0} \varphi(\frac{E_c}{\widetilde{\tau}} - \sum_{\mathcal{U}} p_u). \tag{5.18}$$

A gradient descent method can be used to calculate the optimal value for φ [44], which is given by

$$\varphi(y+1) = \max\{\varphi(y) - \alpha(\frac{E_c}{\widetilde{\tau}} - \sum_{\mathcal{U}} p_u(y)), 0\} \tag{5.19}$$

where y is an iteration index and α is a fixed sufficiently small step size. Since the gradient of (5.18) satisfies the Lipchitz continuity condition, the convergence of (5.19) towards the optimal φ is guaranteed [44]. Hence, the allocated power p_u of (5.17) converges to the optimal solution. The calculation of the optimal power allocation for each radio interface is described in Table 5.2, where ψ is a small tolerance.

B. Video Packet Scheduling for Multi-homing MTs

The achieved data rate for each radio interface is $r_u = b_u \log_2(1 + \Lambda_u p_u)$, given the transmission power allocation p_u. Hence, the optimization problem (5.6) is reduced to

$$\begin{aligned} \max_{x_{zu}} \quad & I \\ s.t. \quad & \sum_{z_f, f \in \mathcal{F}} x_{zu}^f r(z_f) \leq r_u, \quad \forall u \in \mathcal{U} \\ & (5.4), (5.5) \\ & x_{zu}^f \in \{0, 1\}. \end{aligned} \tag{5.20}$$

Problem (5.20) is a binary program. It can be mapped to a new variant of the famous knapsack problem (KP) [91]. In this context, the available items are the video packets, z_f $\forall f \in \mathcal{F}$, the items' weights are the required data rates, $r(z_f)$, and the profit associated with each item is the packet distortion impact, i_f. The problem has multiple knapsacks, since we have multiple radio interfaces, each with transmission capacity r_u.

Table 5.2: Transmission power allocation for each radio interface.

```
Input: $\Lambda_u$, $b_u$ $\forall u \in \mathcal{U}$, $E_c$, $\widetilde{\tau}$, $\alpha$, $\psi$;
Initialization: $\varphi(1) \geq 0$, $y_1 = 1$, $p_u(0) = \{\}$, $y_2 = 0$;
while $y_2 = 0$ do
    for $u \in \mathcal{U}$ do
        $p_u(y_1) = \max\{\frac{b_u}{\varphi(y_1)\ln(2)} - \frac{1}{\Lambda_u}, 0\}$;
    end for
    if $|p_u(y_1) - p_u(y_1 - 1)| > \psi$ then
        $\varphi(y_1 + 1) = \varphi(y_1) - \alpha(\frac{E_c}{\widetilde{\tau}} - \sum_{\mathcal{U}} p_u(y_1))$;
        $y_1 = y_1 + 1$;
    else
        $y_2 = 1$;
    end if
end while
Output: $p_u$ $\forall u \in \mathcal{U}$.
```

Problem (5.20) resembles the multiple knapsack problem (MKP) [91, 92] in the absence of constraint (5.4). The precedence constraint (5.4) is introduced due to the dependences among different video packets. A precedence-constrained knapsack problem (PC-KP) is studied only in literature for the case of single knapsack [91, 93]. To the best of our knowledge, there is no work in literature that studies a multiple knapsack problem with precedence constraints. Hence, in this work we introduce a new variant of the knapsack problem and we refer to it as PC-MKP. Since PC-MKP contains MKP as a special case, and the latter is known to be NP-hard [91], PC-MKP is also NP-hard. Thus, we present a greedy algorithm that can solve the PC-MKP of (5.20) in polynomial time, which is

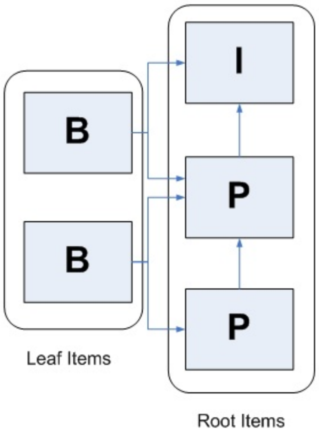


Figure 5.2: Illustration of root and leaf items using base layer frames.

based on the greedy algorithm of [92].

The proposed greedy algorithm consists of two parts. In the first part (A1), we aim to find a feasible solution for the problem through assigning items (video packets) to different knapsacks (radio interfaces) while considering their precedence constraints. Items are first classified into root and leaf items in order to find a feasible solution. This classification is illustrated in Figure 5.2 using video frames from the base layer. In general, root items have higher precedence order than leaf items. For video packet transmission, root items (packets of I and P frames) have higher distortion impact than leaf items (packets of B frames) [19].

The following two steps are used in A1 to find an initial feasible solution:

Step 1: First, root items are packed to different knapsacks as the leaf items cannot be packed without them; then leaf items are packed;

Step 2: Since items are packed in knapsacks in the order of their classification as root and leaf items, some of the early knapsacks may have residual capacity that can be used for packing some of the remaining leaf items whose root items have been packed in Step

1. Hence, the last part of A1 ensures that no residual capacity exists at any knapsack that can be used for packing the remaining leaf items.

In the second part (A2), we aim to improve the obtained feasible solution in A1. This is achieved by considering all pairs of packed items (video packets) and, if possible, interchanges them whenever doing so allows the insertion of an additional item (video packet) from the remaining ones (starting from root items to leaf ones), if all its ancestors are packed, into one of the knapsacks (radio interfaces).

We use the following notations: The feasible packet assignment for each radio interface is given by $\mathcal{G}_u \; \forall u \in \mathcal{U}$. Letting $\mathcal{G} = \underset{\mathcal{U}}{\cup}\mathcal{G}_u$, $\bar{\mathcal{G}} = \underset{f\in\mathcal{F}}{\cup} z_f - \mathcal{G}$ is a set of remaining unassigned video packets. Let R_u be the current used transmission capacity for each radio interface (thus, the remaining capacity is $c_u = r_u - R_u$), and y_{zf} is an index of the radio interface where packet z_f is currently assigned to. The algorithm in Table 5.3 describes video packet scheduling for multi-homing MTs.

It is assumed in the algorithm in Table 5.3 that video packets are sorted according to their classification as root and leaf items. In A2 of Table 5.3, $\mathcal{G}, \bar{\mathcal{G}}, c_u$, and y_{zf} are updated whenever some $\mathcal{G}_u$ is updated. Let the total number of available video packets from the current time slot be $\sum_{f\in\mathcal{F}} G_f$. The complexity of A1 is $O(\sum_{f\in\mathcal{F}} G_f|\mathcal{U}|)$ and A2 is $O(\{\sum_{f\in\mathcal{F}} G_f\}^2)$. Thus, the algorithm, in Table 5.3, has polynomial time complexity.

5.5 Numerical Results and Discussion

This section presents numerical results for the energy and content aware multi-homing video transmission mechanism, for one GoP, in a heterogeneous wireless access medium. Video sequences are compressed at an encoding rate of 30 fps [80, 84], and the GoP structure is composed of 12 frames [94] from one layer (base layer) with one B frame between P frames. Hence, the time slot duration $\widetilde{\tau}$ is set to 400 milli-second. Each

Table 5.3: Video packet scheduling for multi-homing MTs.

```
**A1: Finding a Feasible Solution**
**Initialization**: $\bar{\mathcal{G}} \longleftarrow \bigcup_{f \in \mathcal{F}} z_f$, $R_u \longleftarrow 0$, $\mathcal{G}_u = \{\}\ \forall u \in \mathcal{U}$;
**for** $u \in \mathcal{U}$ **do**
    **for** $z_f \in \bar{\mathcal{G}}$ **do**
        **if** $x^{f'}_{z'u'} = 1\ \forall z'_{f'} \in \mathcal{Z}^f_z, u' \in \mathcal{U}$, $r(z_f) + R_u \leq r_u$ **then**
            $x^f_{zu} = 1$, $R_u = R_u + r(z_f)$;
        **end if**
        $\mathcal{G}_u = \mathcal{G}_u \cup \{z_f\}$;
    **end for**
    $\bar{\mathcal{G}} = \bar{\mathcal{G}} - \mathcal{G}_u$;
**end for**
**for** $u \in \mathcal{U}$ and $c_u > \min\{r(z_f)|z_f \in \bar{\mathcal{G}}\}$ **do**
    **for** $z_f \in \bar{\mathcal{G}}$ **do**
        **if** $x^{f'}_{z'u'} = 1\ \forall z'_{f'} \in \mathcal{Z}^f_z, u' \in \mathcal{U}$, $r(z_f) + R_u \leq r_u$ **then**
            $x^f_{zu} = 1$, $R_u = R_u + r(z_f)$;
        **end if**
        $\mathcal{G}_u = \mathcal{G}_u \cup \{z_f\}$;
    **end for**
    $\bar{\mathcal{G}} = \bar{\mathcal{G}} - \mathcal{G}_u$;
**end for**
**A2: Improving the Feasible Solution**
**for** $z1 \in \{z_f | z_f \in \mathcal{G}, c_{y_{zf}} + \max_{u \neq y_{zf}} c_u \geq \min_{z'_{f'} \in \bar{\mathcal{G}}} r(z'_{f'})\}$ **do**
    **for** $z2 \in \{z_f | z_f \in \mathcal{G}, z_f > z1, y_{zf} \neq y_{z1}, c_{y_{zf}} + c_{y_{z1}} \geq \min_{z'_{f'} \in \bar{\mathcal{G}}} r(z'_{f'})\}$ **do**
        $W(u1) = \max\{r(z1), r(z2)\}$, $W(u2) = \min\{r(z1), r(z2)\}$;
        $g_{u1} = y_{u1}$, $g_{u2} = y_{u2}$, $\delta_{u0} = W(u1) - W(u2)$;
        **if** $\delta_{u0} \leq c_{g_{u2}}$ and $c_{g_{u1}} + \delta_{u0} \geq \min_{z'_{f'} \in \bar{\mathcal{G}}} r(z'_{f'})$ **then**
            $i_{u*} = \max\{i_{z'_{f'}} | z'_{f'} \in \bar{\mathcal{G}}, r(z'_{f'}) \leq c_{g_{u1}} + \delta_{u0}, \mathcal{Z}^{f'}_{z'} \subset \mathcal{G}\}$;
            $\mathcal{G}_{g_{u1}} = (\mathcal{G}_{g_{u1}} - u1) \cup \{u2, u*\}$, $\mathcal{G}_{g_{u2}} = (\mathcal{G}_{g_{u2}} - u2) \cup \{u1\}$;
        **end if**
    **end for**
**end for**
```

encoded video frame has a variable length 6000 - 9600 bits [84]. Specifically, for the GoP under consideration, the frame length is 9600 bits for I frames, 8000 bits for P frames, and 6000 bits for B frames. Each I frame is encoded into 12 packets, while each of B and P frames are encoded into 10 packets. The decoder time stamp difference, ΔD, between two successive frames is 40 milli-second [6]. Thus, each I or P packet requires data rate $r(z_f)$ of 20 Kbps, while an B packet requires a data rate of 15 Kbps. The packet distortion impact values are $i_f = 5$ for I frames, $i_f = 4$ for P frames, and $i_f = 2$ for B frames [80]. Two radio interfaces are used for video transmission ($\mathcal{U} = \{1, 2\}$). The system unit bandwidth is 363 KHz. In the numerical results, the proposed energy and content aware multi-homing video transmission mechanism, the greedy approach (GA), is compared with the exact solution using the cutting plane approach (CPA). The MIPs of the CPA are solved using the CPLEX solver through GAMS [72]. The GA is also compared with two benchmarks. The first benchmark is an energy independent approach (EIA), where problem (5.10) is solved without the MT battery energy constraint of (5.3). The second benchmark is an earliest deadline first approach (EDFA), which is a common benchmark for video packet scheduling [80]. In the EDFA, packets whose deadline is closer are scheduled earlier. Hence, the EDFA is content independent, unlike the GA which first schedules packets with higher distortion impact. In order to determine the power allocation for each radio interface in the EDFA, we employ an equal power allocation approach (EPA) [95], where the energy budget per time slot, E_c, is distributed equally between the two radio interfaces.

Numerical results are studied for multi-homing video transmission of a GoP over one time slot. Two sets of results are presented. In the first set of results, given by Figures 5.3 and 5.4, the energy budget per time slot, E_c, is varied from 10 to 120 milli-joule, which is equivalent to a video transmission duration of 120 to 10 minutes given an MT battery available energy of 180 Joule (a blackberry Lithium Ion battery is 900 mAh and

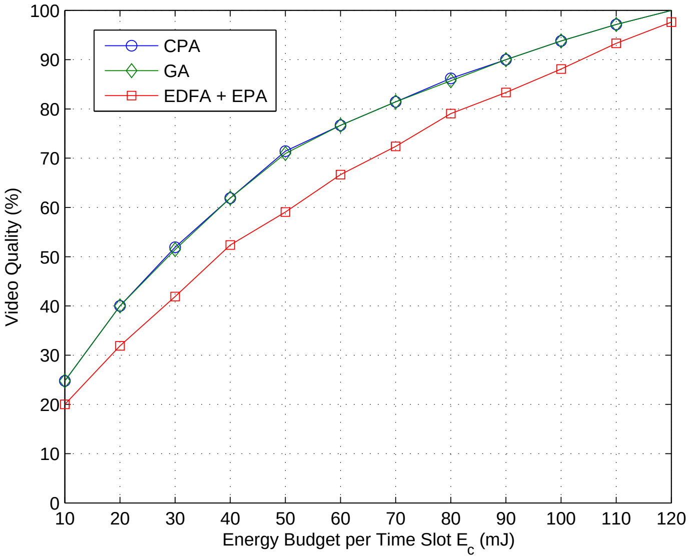


Figure 5.3: The achieved video quality using variable energy budget per time slot E_c.

3.7 Volt, i.e. the battery capacity is 11988 J). For the time slot under consideration, the channel gain is given by $\Omega_1 = 0.5019$ and $\Omega_2 = 0.448$ for the two radio interfaces, and the allocated bandwidth is 1 unit from the first radio interface and 2 units from the second radio interface. The background noise power, $\eta_u = \eta_0 b_u$, is equal to 0.01 watt for the first radio interface [96], and 0.02 watt for the second radio interface. In the second set of results, given by Figures 5.5 and 5.6, the energy budget per time slot is fixed at $E_c = 170$ milli-joule while the channel gain for the first radio interface is varied. For these results, the channel gain for the second radio interface is fixed at $\Omega_2 = 0.448$, the allocated bandwidth is 1 unit from each radio interface, and the background noise power for both radio interfaces is $\eta_u = 0.01$, $u \in \mathcal{U}$. In the numerical results, the video quality metric is defined as the distortion impact ratio of the transmitted packets to the total packets.

Figure 5.3 shows the video quality versus the energy budget per time slot E_c. In general, as expected, as E_c increases, more transmission power can be allocated to both radio interfaces, which results in higher transmission data rates and hence more transmitted packets. The CPA and the GA exhibit very close performance in terms of the perceived video quality. This demonstrates the effectiveness of the GA, whose performance is very close to that of the CPA (the exact solution) but with reduced computational complexity. The main difference between the CPA and the GA is that the CPA jointly optimizes the transmission power allocation and the video packet scheduling. Hence, in the CPA, the transmission capacities of different radio interfaces are determined so as to assign as many valuable video packets as possible in order to minimize the video quality distortion. On the other hand, the GA maximizes the transmission capacity for each radio interface and then performs video packet scheduling. As a result, unlike the CPA, one packet may not fit in any of the radio interfaces although the sum of the residual capacities in both radio interfaces is enough to transmit this packet. This is the reason that the CPA has a slightly higher performance for different E_c values as compared to the GA. However, this is always corresponding to a maximum of one additional packet insertion and its contribution to the total video quality is not significant, as shown in the figure. In general, for $|\mathcal{U}|$ radio interfaces, the CPA can insert a maxmium of $|\mathcal{U}| - 1$ additional packets as compared to the GA. In case that the number of available video packets is small, the CPA and the GA performances will not coincide as in Figure 5.3, but rather the CPA performance will upper bound the GA performance, and the gap between them is due to the distortion impact value corresponding to dropping $|\mathcal{U}| - 1$ packets in the GA as compared to the CPA. However, in practical scenarios, it is expected to have a large number of packets per time slot [84]. Hence, the impact of the additional video packets on the achieved video quality is not significant and the performance of the CPA and GA will almost coincide as in Figure 5.3. The EDFA with EPA achieves lower performance than the content aware approaches (CPA and GA) as it does not schedule packets according

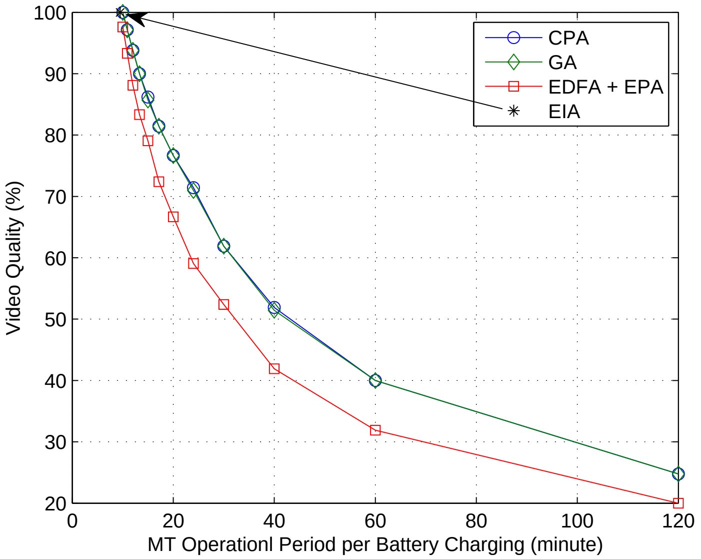


Figure 5.4: The trade-off between the achieved video quality and the MT operational period per battery charging.

to their distortion impact. At a high E_c ($E_c > 100$ milli-joule), both the content aware approaches and the EDFA have sufficient energy budget so that almost all video packets are scheduled for transmission, hence the difference in the scheduling policies (i.e. which packets are dropped) is not significant, which results in the close performance.

Figure 5.4 shows the video quality versus the MT operational period per battery charging. In general, requiring high video quality results in a lower operational period for the MT (less than 20 minutes). However, as shown in figure, the content aware approaches can achieve the same video quality as the EDFA, but at a longer MT operational period per battery charging. For the energy independent approach (EIA), the achieved video quality is always 100%, yet the consumed energy per time slot is always 120 milli-watt. This is equivalent to a video duration of 9.5 minutes given the MT available energy (180 Joule). On the other hand, the GA offers a choice for desirable trade-off between

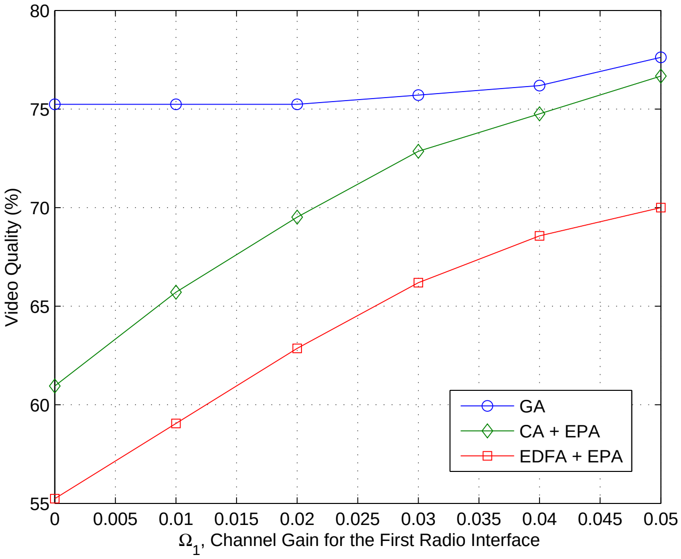


Figure 5.5: Video quality performance for a varying channel gain.

the video quality and the consumed energy per time slot E_c. Hence, while the GA can provide a variable video quality ranging from $25-100\%$ for a total duration of $120-10$ minutes, the energy independent approaches present only a fixed video quality for a short MT operational period.

Figure 5.5 shows the video quality versus the channel gain of the first radio interface Ω_1. The figure gives a comparison among the GA, the content aware (CA) approach based on the algorithm in Table 5.3 using an EPA for transmission power allocation (instead of the algorithm in Table 5.2 as in the GA), and the EDFA (which is content independent) with EPA. In general, since the EPA approach (for both CA and EDFA) allocates transmission power independent of the channel condition, the achieved transmission capacity is lower than that of the GA at a poor channel condition. This results in an improvement in video quality for the GA as compared with the CA and EDFA with EPA at a poor channel condition. As the EDFA is content independent, it achieves a lower video

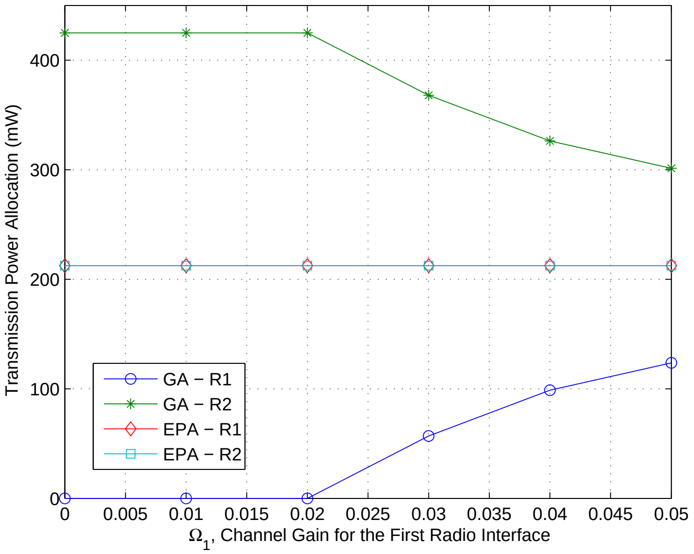


Figure 5.6: Transmission power allocation for varying channel gain.

quality than the CA approach. With an improved channel quality ($\Omega_1 > 0.03$), the CA approach with EPA can achieve performance close to that of the GA. The transmission power allocation for each radio interface (R1 and R2) versus the channel gain of the first radio interface is given in Figure 5.6. The EPA has a fixed power allocation independent of the channel condition. On the other hand, the GA adapts its power allocation for each radio interface based on the channel condition for the interface, hence maximizing the achieved transmission capacity and the achieved video quality.

5.6 Summary

In this chapter, energy and content aware multi-homing video transmission is presented for a heterogeneous wireless access medium. The objective is to perform power allocation and video packet scheduling for different radio interfaces in order to minimize the perceived video quality distortion with an acceptable computational complexity. The newly proposed energy and content aware video transmission mechanism offers a desirable trade-off between the perceived video quality and the MT operational period. The energy and content aware multi-homing video transmission problem formulation is based on an MINLP which can be computational intractable for an expected large number of video packets. A piecewise linearization approach is employed to reduce the problem complexity from MINLP to a series of MIPs, which is very efficient for a large-size problem. For practical implementation in MTs, a greedy approach (GA) is proposed to perform the power allocation and packet scheduling in polynomial time complexity. The GA separates the problem into two stages. The first stage optimizes the allocated power for each radio interface given the interface available bandwidth, channel condition, and the MT battery energy constraint. The second stage performs video packet scheduling to different radio interfaces so as to minimize the resulting video quality distortion. We map the packet scheduling problem for multi-homing video transmission to a new variant of the knapsack problem, namely PC-MKP, and solve it in polynomial time complexity of the problem parameters in terms of the number of radio interfaces and the number of video packets using a greedy algorithm. Numerical results demonstrate that the proposed framework has performance very close to the exact solution yet at a reduced computational complexity. However, the proposed mechanism equally distribute the MT available energy over different time slots. Given the time varying video packet encoding and channel conditions at different radio interfaces, using this uniform energy distribution will lead to inconsistent temporal fluctuations in the video quality. In the next chapter, we present an energy management mechanism that employs the MT energy in a way such that it can support the target call duration with a consistent video quality over time slots, independent of varying packet encoding and channel conditions.

Chapter 6

Mobile Terminal Energy Management for Video Transmission

In Chapter 5, an energy and content aware multi-homing video transmission mechanism is proposed. The developed mechanism uses a fixed energy budget per time slot over the target call duration. However, in wireless fading channels with time varying conditions and with time varying video packet encoding, using a fixed energy budget per time slot will lead to inconsistent temporal fluctuations in the video quality. An appropriate energy management mechanism should adapt the MT energy consumption at each time slot according to the channel conditions and the video packet encoding in order to achieve a consistent video quality over the target call duration. In this chapter, an energy management mechanism is proposed for MTs to support a sustainable multi-homing video transmission, over the target call duration, in a heterogeneous wireless access medium. The energy management mechanism has two stages. In the first stage, through video quality statistical guarantee, the MT can determine a target video quality lower bound that can be supported for the target call duration with a pre-defined success probability. The target video quality lower bound captures the MT available energy at the beginning of the call, the available bandwidth and time varying channel conditions at different radio interfaces, the target call duration, and the video packet characteristics in terms of distortion impact, delay deadlines, precedence constraints, and video packet encoding statistics, which is discussed in Section 6.1. In the second stage, the MT adapts its energy consumption, based on the current conditions of the channels at different radio interfaces, to achieve at least the target video quality lower bound throughout the call, which is discussed in Section 6.2. The same system model presented in Chapter 5 is considered in this chapter. Moreover, each radio interface $u \in \mathcal{U}$ can support a discrete set of data rates r_{u,v_u}, with $v_u \in \mathcal{V} = \{1, 2, \dots, V\}$.

6.1 Statistical QoS Guarantee for Wireless Multi-homing Video Transmission

This section presents the energy management mechanism first stage, namely the call set-up phase. In the call set-up stage, the main objective is to determine the maximum QoS lower bound that can be supported with statistical guarantee for multi-homing video transmission.

Let Q_t denote the video quality metric which is defined as the distortion impact ratio of the transmitted packets to the total available packets in time slot $t \in \mathcal{T}$. Due to channel fading, and hence time varying data rates at different radio interfaces, and packet encoding statistics, the video quality metric Q_t is a discrete random variable. For a stationary and ergodic process of system dynamics (in terms of channel fading and packet encoding), Q_t is i.i.d. with respect to t and therefore the time subscript t can be omitted. Hence, Q is given as

$$Q = \frac{\sum_{\mathcal{U}} \sum_{z_f, f \in \mathcal{F}} x_{zu}^{f} i_f}{\sum_{z_f, f \in \mathcal{F}} i_f}. \tag{6.1}$$

We aim to find the video quality CDF, $F_Q(q)$, given the MT available energy, the available bandwidth and time varying channel conditions at different radio interfaces, the target call duration, and the video packet characteristics in terms of distortion impact, delay deadlines, precedence constraints, and packet encoding statistics. Using the video quality CDF, we can find the video quality lower bound, q_l, that can be supported by the MT for the target call duration such that $\Pr(Q \leq q_l) \leq \varepsilon$, with $\varepsilon \in [0, 1]$. This is achieved following a three-step framework: 1) The probability of employing a given set of data rates at different radio interfaces is calculated; 2) Using a video packet scheduling algorithm, given the frame size and data rate statistics, we find the video quality PMF and hence calculate the video quality CDF; and 3) Through optimal average power allocation to different radio interfaces, we find the maximum video quality lower bound, q_l, that can be supported with a success probability $\varepsilon_s = 1 - \varepsilon$. This is discussed in more details in the following.

A. Data Rate PMF

Radio interface $u \in \mathcal{U}$ can support data rate r_{u,v_u} if the received SNR value, γ_u, at the BS/AP communicating with u exceeds some threshold Γ_{u,v_u}. The set of thresholds Γ_{u,v_u}, $\forall u \in \mathcal{U}$, can be calculated using Shannon formula as

$$\Gamma_{u,v_u} = 2^{\frac{r_{u,v_u}}{b_u}} - 1, \quad u \in \mathcal{U}, v_u \in \mathcal{V} \tag{6.2}$$

and $\Gamma_{u,V+1}$ is assumed to be ∞.

In a fading channel, the received SNR value, γ_u, is larger than a threshold, Γ_{u,v_u}, with probability

$$P_{u,v_u} = \Pr(\gamma_u > \Gamma_{u,v_u}). \tag{6.3}$$

The probability that data rate r_{u,v_u} is used at radio interface u, $v_u \in \mathcal{V}$, is given by

$$\Psi_{u,v_u} = P_{u,v_u} - P_{u,v_u+1}, \quad v_u \in \mathcal{V}. \tag{6.4}$$

For independent fading statistics at different radio interfaces, the probability that data rates $r_{1,v_1}, r_{2,v_2}, \ldots, r_{|\mathcal{U}|,v_{|\mathcal{U}|}}$ are used at radio interfaces $1, 2, \ldots, |\mathcal{U}|$ can be calculated as

$$f_{R_{1,v_1},\cdots,R_{|\mathcal{U}|,v_{|\mathcal{U}|}}}(r_{1,v_1}, \cdots, r_{|\mathcal{U}|,v_{|\mathcal{U}|}}) = \prod_{u=1}^{|\mathcal{U}|} \Psi_{u,v_u}. \tag{6.5}$$

For instance, in a Rayleigh fading channel, γ_u follows an exponential distribution, which is given by

$$f_{\Upsilon_u}(\gamma_u) = \frac{1}{\bar{\gamma}_u} \cdot e^{-\frac{\gamma_u}{\bar{\gamma}_u}}, \quad u \in \mathcal{U} \tag{6.6}$$

where $\bar{\gamma}_u = \frac{\bar{p}_u \bar{\Omega}_u}{b_u \eta_0}$ denotes the average received SNR for radio interface u and $\bar{\Omega}_u$ denotes the average channel power gain for radio interface u. Hence, $f_{R_{1,v_1},\cdots,R_{|\mathcal{U}|,v_{|\mathcal{U}|}}}(r_{1,v_1}, \cdots, r_{|\mathcal{U}|,v_{|\mathcal{U}|}})$ is given by

$$f_{R_{1,v_1},\cdots,R_{|\mathcal{U}|,v_{|\mathcal{U}|}}}(r_{1,v_1}, \cdots, r_{|\mathcal{U}|,v_{|\mathcal{U}|}}) = \prod_{u=1}^{|\mathcal{U}|}(e^{-\frac{\Gamma_{u,v_u}}{\bar{\gamma}_u}} - e^{-\frac{\Gamma_{u,v_u+1}}{\bar{\gamma}_u}}). \tag{6.7}$$

B. Video Quality CDF

In the following, we aim to find the video quality q that can be achieved given the MT data rates r_{u,v_u} at different radio interfaces and frame size g_f with f belongs to I, P, and B types. Using the data rate and packet encoding statistics, we find the video quality CDF, $F_Q(q)$.

From Chapter 5, the multi-homing video packet scheduling, given the available data rates $r_{1,v_1}, r_{2,v_2}, \ldots, r_{|\mathcal{U}|,v_{|\mathcal{U}|}}$ at different radio interfaces and frame size g_f with f belonging to I, P, and B types, should satisfy

$$\begin{aligned} \max_{x_{zu}^f} \quad & q \\ s.t. \quad & \sum_{z_f, f \in \mathcal{F}} x_{zu}^f r(z_f) \leq r_{u,v_n}, \quad \forall u \in \mathcal{U}, v_u \in \mathcal{V} \\ & (5.4), (5.5) \\ & x_{zu}^f \in \{0, 1\}. \end{aligned} \tag{6.8}$$

Using the packet scheduling algorithm in Table 5.3 in solving (6.8), the video quality q that can be achieved using data rates $r_{1,v_1}, r_{2,v_2}, \ldots, r_{|\mathcal{U}|,v_{|\mathcal{U}|}}$ at radio interfaces $1, 2, \ldots, |\mathcal{U}|$ and frame size g_f with f belonging to I, B, and P types can be calculated. The set of different data rates and packet encoding combinations that result in the same video quality q is denoted by $\mathcal{Q}$. We can map the data rate and frame size statistics into a video quality PMF given by

$$f_Q(q) = \sum_{\mathcal{Q}} \{ f_{R_{1,v_1},\cdots,R_{|\mathcal{U}|,v_{|\mathcal{U}|}}}(r_{1,v_1}, \cdots, r_{|\mathcal{U}|,v_{|\mathcal{U}|}}) \cdot f_{G_I,G_B,G_P}(g_I, g_B, g_P) \} \tag{6.9}$$

where $f_{G_I,G_B,G_P}(g_I, g_B, g_P)$ denotes the joint PMF of video packet encoding for I, B, and P frames which is given as the multiplication of the PMFs of I, B, and P frames assuming an i.i.d. frame size statistics [19]. As a result, the video quality CDF, $F_Q(q)$, can be calculated.

A look-up table can be stored at the MT to derive the CDF of the video quality that can be achieved, as given in (6.9). Sample packet encoding PMF can be used according the video type (high motion or low motion). In addition, the discrete set of data rates that can be used at different radio interfaces are already known. Hence, using the packet scheduling algorithm in Table 5.3 and given the packet encoding statistics and allowed data rates at different radio interfaces, a look-up table can be created with two columns, the first column gives the video quality that can be achieved and the second column gives the corresponding probability as a function of the the average received SNR values, $\bar{\gamma}_u$ $\forall u \in \mathcal{U}$. Once $\bar{\gamma}_u$ $\forall u \in \mathcal{U}$ is specified, as will be explained in the next sub-section, an approximate expression of the achievable CDF of the video quality is obtained.

C. Maximum QoS Lower Bound That Can Be Achieved with Statistical Guarantee

From (6.7), the probability that data rates r_{u,v_u} are used at different radio interfaces depends on the average received SNR values, $\bar{\gamma}_u$ $\forall u \in \mathcal{U}$. As a result, the video quality CDF is a function of the average transmission power at different radio interfaces. Hence, the distribution of the average transmission power, $\frac{E}{T_c}$, among different radio interfaces, $\bar{p}_u$, affects the resulting video quality CDF. Assuming an ergodic process of system dynamics (in terms of channel fading and video packet encoding), in order to find the maximum video quality lower bound, q_l, that can be supported for the target call duration, T_c, with some statistical guarantee, ε, we need to solve

$$\begin{aligned} \max_{\bar{p}_u \geq 0} \quad & q_l \\ s.t. \quad & F_Q(q_l) \leq \varepsilon \\ & \sum_{\mathcal{U}} \bar{p}_u \leq \frac{E}{T_c}. \end{aligned} \tag{6.10}$$

The first constraint in (6.10) has an inequality (instead of equality) since the supported data rates at different radio interfaces form a discrete set, and hence the achieved video quality is also discrete. As a result, an equality in the first constraint of (6.10) cannot always be satisfied, unlike the inequality. In (6.10), ε is a design parameter that can be chosen to strike a balance between the desired performance (in terms of the video quality and energy consumption) and success probability of the call delivery. This issue is further investigated in the simulation result section.

Heuristic optimization techniques, e.g. the Genetic Algorithm [97], can be used to solve the optimization problem (6.10). The Genetic Algorithm can be easily implemented in smart phones as it consists of simple iterations. In addition, using the Genetic Algorithm in solving (6.10) is fast due to the small number of variables (the number of radio interfaces can be from 2 to 3). Following (6.10), the MT can support a multi-homing video quality at least equals to q_l for the target call duration, T_c, with success probability ε_s.

6.2 Energy Efficient QoS Provision for Wireless Multi-homing Video Transmission

This section presents the energy management mechanism second stage. During the call, the MT adapts its energy consumption to satisfy at least the video quality lower bound, q_l, calculated in the call set-up. This is performed in three steps: 1) The MT determines the total required data rate, at the current time slot, in order to satisfy at least q_l, given the current time slot video packet encoding; 2) The MT determines the minimum power required at each radio interface, and hence the required data rate at each radio interface, in order to satisfy the total required data rate calculated in 1), given the current time slot channel fading; and 3) The MT performs video packet scheduling given the data rate at each radio interface, calculated in 2). These are discussed in more details in the following.

A. Total Required Data Rate

Due to the time varying video packet encoding (i.e. g_f for f belongs to I, B, and P packets), the total required data rate in order to satisfy at least the video quality lower bound, q_l, varies over time. As a result, at the beginning of each time slot, $t \in \mathcal{T}$, given the available video packets ready for transmission, the MT determines the total required data rate, r, that satisfies at least the video quality lower bound. Let q_t denote the resulting video quality that can be achieved at time slot $t \in \mathcal{T}$ by scheduling a set $\mathcal{G}$ of video packets for transmission. The total required data rate, r, can be calculated using the algorithm in Table 6.1.

In the algorithm in Table 6.1, it is assumed that video packets are sorted according to their classification as root and leaf items. The algorithm finds the total data rate required to satisfy at least the video quality lower bound, q_l, by scheduling video packets with the highest distortion impact for transmission until q_l at least is satisfied.

Table 6.1: Calculation of total required data rate to satisfy QoS lower bound.

```
Input: $g_f \ \forall f \in \mathcal{F}$;
Initialization: $\mathcal{L} \longleftarrow \bigcup_{f \in \mathcal{F}} z_f$, $r \longleftarrow 0$, $\mathcal{G} = \{\}$;
while $q_t < q_l$ do
    if $x^{f'}_{z'u'} = 1 \ \forall z'_{f'} \in \mathcal{Z}^f_z, u' \in \mathcal{U}$ then
        $x^f_{zu} = 1$, $r = r + r(z_f)$;
    end if
    $\mathcal{G} = \mathcal{G} \cup \{z_f\}$;
end while
Output: $r$.
```

B. Minimum Power Allocation

Due to the time varying channel conditions at different radio interfaces, the required power allocation, p_u, to satisfy the total data rate, r, needs to be determined at the beginning of every time slot $t \in \mathcal{T}$. Assuming available perfect channel state information (CSI) [98] and through power allocation, the received SNR value, γ_u, for different radio interfaces can be determined. When γ_u exceeds threshold Γ_{n,v_u}, radio interface u can support data rate r_{u,v_u}. Hence, power allocation affects the resulting data rate at each radio interface, r_{u,v_u}. As a result, the objective is to find the minimum power allocation to different radio interfaces, which is required to satisfy the total data rate r calculated in Table 6.1. Let E_t denote the MT available energy at the beginning of time slot t. The power allocation problem can be described as in (6.11). Similar to (6.10), (6.11) can be solved using the Genetic Algorithm.

$$\begin{aligned} \min_{p_u \geq 0} \quad & \sum_{\mathcal{U}} p_u \widetilde{\tau} \\ s.t. \quad & \sum_{\mathcal{U}} r_{u,v_u} \geq r \\ & \sum_{\mathcal{U}} p_u \widetilde{\tau} \leq E_t. \end{aligned} \tag{6.11}$$

C. Video Packet Scheduling

Using the data rates, r_{u,v_u}, that can be supported through the power allocation, p_u, calculated in (6.11), the packet scheduling algorithm in Table 5.3 is used to schedule the current time slot available video packets for transmission. The resulting video quality satisfies the lower bound q_l calculated in (6.10) with a success probability ε_s.

The energy management sub-system procedure for supporting a sustainable video transmission over a target call duration with consistent video quality is summarized in Figure 6.1.

6.3 Benchmarks

In this section, two benchmarks are presented for comparison. The first benchmark aims to maximize the resulting video quality in the absence of an energy management sub-system, similar to [80]. The second benchmark satisfies an energy budget per time slot for energy management, similar to [20].

A. Multi-homing Video Transmission Without Energy Management

In the absence of an energy management mechanism, the main objective is to maximize the resulting video quality subject to the MT battery energy limitation. Intuitively, the higher the achieved data rates at different radio interfaces, subject to the MT battery

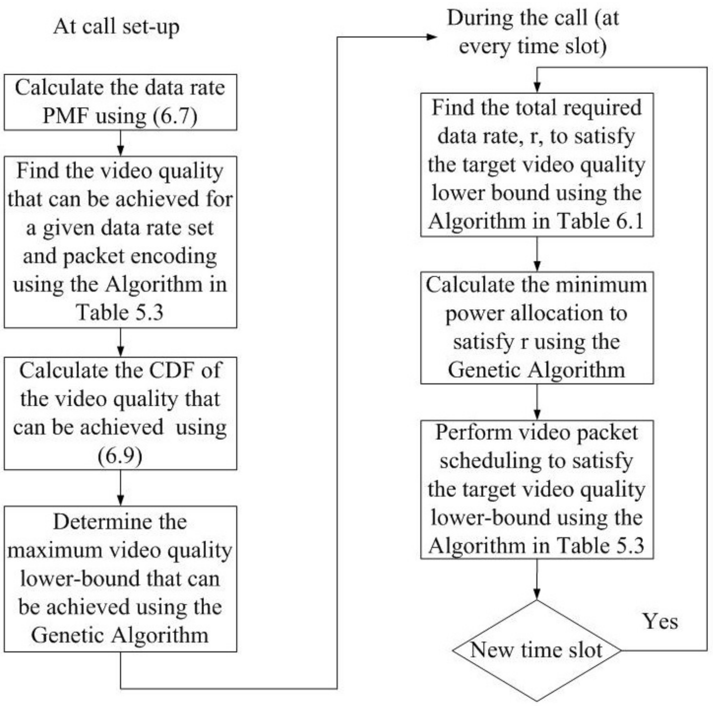

At call set-up
Calculate the data rate PMF using (6.7)
Find the video quality that can be achieved for a given data rate set and packet encoding using the Algorithm in Table 5.3
Calculate the CDF of the video quality that can be achieved using (6.9)
Determine the maximum video quality lower-bound that can be achieved using the Genetic Algorithm
During the call (at every time slot)
Find the total required data rate, r, to satisfy the target video quality lower bound using the Algorithm in Table 6.1
Calculate the minimum power allocation to satisfy r using the Genetic Algorithm
Perform video packet scheduling to satisfy the target video quality lower-bound using the Algorithm in Table 5.3
New time slot
Yes

Figure 6.1: Flow chart of the proposed energy management mechanism procedure.

energy limitation, the more transmitted video packets and thus the better video quality. Hence, at the beginning of every time slot $t \in \mathcal{T}$, the MT performs power allocation at different radio interfaces to maximize the resulting sum data rate. This is given by

$$\begin{aligned} \max_{p_u \geq 0} \quad & \sum_{\mathcal{U}} r_{u,v_u} \\ s.t. \quad & \sum_{\mathcal{U}} p_u \widetilde{\tau} \leq E_t. \end{aligned} \tag{6.12}$$

Problem (6.12) is solved using the Genetic Algorithm. Given the power allocation, p_u, and hence the data rates r_{u,v_u} $\forall u \in \mathcal{U}$, the packet scheduling algorithm in Table 5.3 is used to schedule the current time slot available video packets for transmission.

B. Multi-homing Video Transmission With Uniform Energy Management

In this case a uniform energy budget per time slot is considered. Hence, the MT available energy at time slot t is uniformly distributed over the remaining time slots. The energy budget per time slot, starting from time slot t, is given by $E_{ct} = \frac{E_t}{T-t}$. At the beginning of time slot t, the MT determines the maximum data rate that can be supported at each radio interface through power allocation subject to the energy budget constraint. This is achieved by solving (6.12) while replacing E_t in the problem constraint by E_{ct}. Given the resulting data rates r_{u,v_u} $\forall u \in \mathcal{U}$, the packet scheduling algorithm in Table 5.3 is used to schedule the available video packets for transmission in the current time slot.

6.4 Simulation Results and Discussion

This section presents simulation results for the proposed energy management mechanism. Video sequences are compressed at an encoding rate of 30 fps [80, 84]. The GoP structure consists of 13 frames with one layer (base layer) and one B frame between P frames [94]. As a result, the time slot duration $\widetilde{\tau}$ is 433 milli-seconds. In practice, the PMFs of the I, B, and P frame sizes can be generated using the video trace as in [87]. For simplicity, sample PMFs of the I, B, and P frame sizes are arbitrary generated as shown in Figure 6.2. The decoder time stamp difference between two successive frames, ΔD, is 40 milli-seconds [6]. Each video packet requires a transmission data rate of 2 Kbps. The video packet distortion impact values are $i_f = 5$ for I frames, $i_f = 4$ for P frames, and $i_f = 2$ for B frames [80]. Two radio interfaces are used for video transmission ($|\mathcal{U}| = 2$). The allocated bandwidth is 1 unit for the first radio interface and 2 units for the second radio interface, where the unit bandwidth is 250 KHz. The set of data rates that can be supported on each radio interface is $\mathcal{R} = \{0, 0.256, 0.512, 1, 1.5, 2, 2.5\}$ Mbps. Using (6.2), $\mathcal{R}$ is supported with different thresholds at the two different radio interfaces. The background noise power, $\eta_u = \eta_0 b_u$, is 0.01 watts for the first radio interface, and 0.02

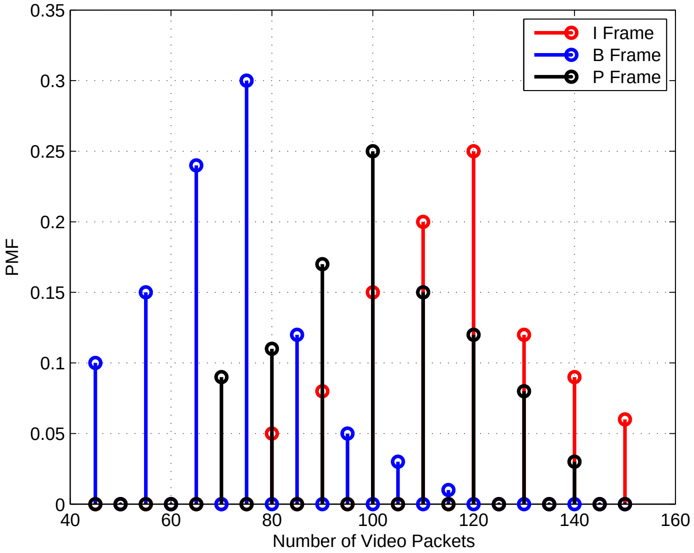


Figure 6.2: The probability mass function (PMF) of I, B, and P frame sizes.

watts for the second radio interface. Each radio interface suffers from a Rayleigh fading channel with average channel power gain of $\bar{\Omega}_1 = 0.2031$ and $\bar{\Omega}_2 = 0.1852$.

A. Performance of the Proposed Energy Management Mechanism

In the following, the performance of the proposed energy management mechanism is investigated versus MT available energy E, target call duration T_c, and call success probability ε_s. Different performance trade-offs are demonstrated.

Figure 6.3 shows the complementary cumulative distribution function (CCDF), $\Pr(Q > q)$, of the video quality (q) for $E \in [3, 11]$ KJ, $T_c = 20$ minutes, and $\varepsilon_s = 0.9$. The more the available energy at the MT, for the given target call duration, the better the video quality that can be achieved with $\varepsilon_s = 0.9$. For instance, with $E = 11$ KJ, a video quality of 95% can be guaranteed with probability 0.9, while a video quality of only 60% can be guaranteed with probability 0.9 for $E = 3$ KJ.

Figure 6.4 plots the video quality lower bound, q_l, that can be achieved with different

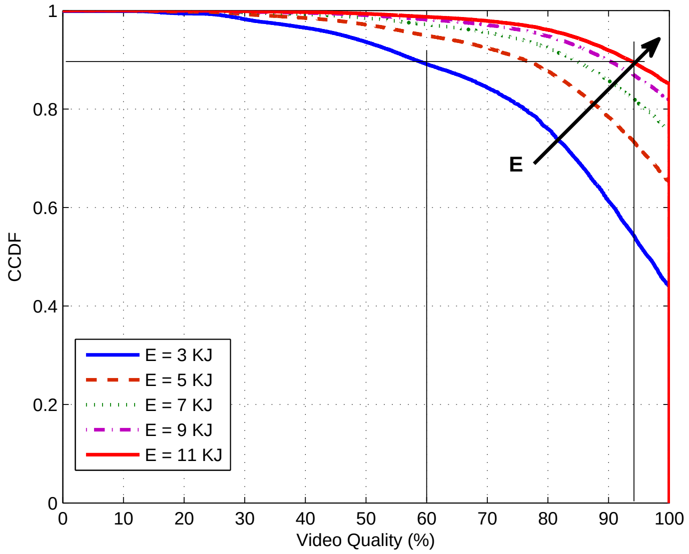


Figure 6.3: The complementary cumulative distribution function (CCDF) of the achieved video quality (q) for different values of MT available energy.

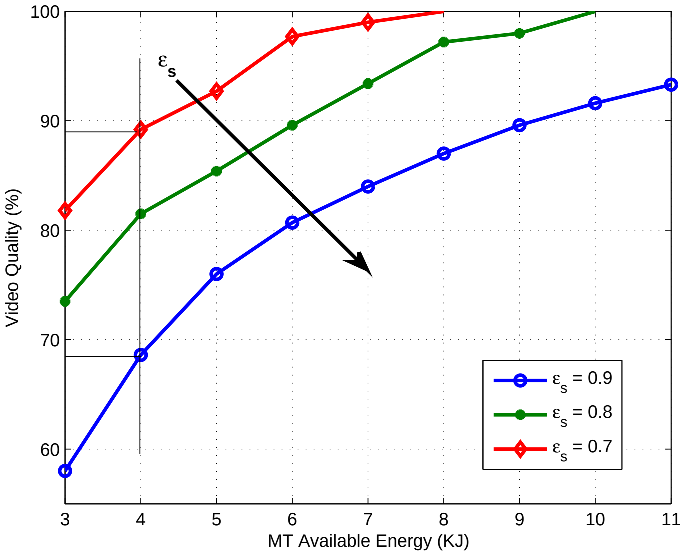


Figure 6.4: The video quality lower bound that can be supported (q_l) versus MT available energy (E) for different success probability ε_s.

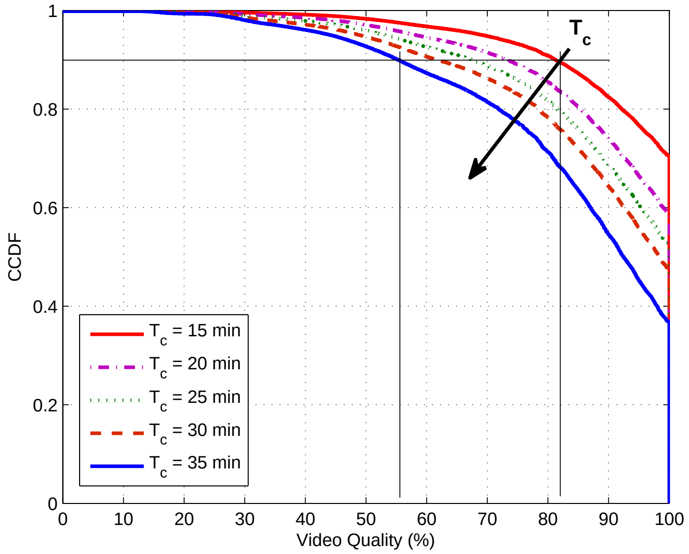


Figure 6.5: The complementary cumulative distribution function (CCDF) of the achieved video quality (q) for different target call duration.

success probability ε_s values, versus the MT available energy. Higher video quality can be supported with a lower success probability, for a given MT available energy. For instance, with $E = 4$ KJ, a video quality of 89% can be achieved with probability 0.7, while a video quality of 68% can be guaranteed with probability 0.9 at the same E.

Figure 6.5 plots the CCDF of the video quality (q) for $T_c \in [15, 35]$ minutes, $E = 5$ KJ, and $\varepsilon_s = 0.9$. The longer the target call duration, for the given MT available energy, the lower the video quality that can be supported with $\varepsilon_s = 0.9$. For example, with $T_c = 35$ minutes, video quality of 55% can be achieved with probability 0.9, while for $T_c = 15$ minutes, a video quality of 81% can be guaranteed with the same probability.

B. Performance Comparison

In the following, the performance of the proposed energy management mechanism is compared with that of the two benchmarks in Section 6.3. The proposed energy

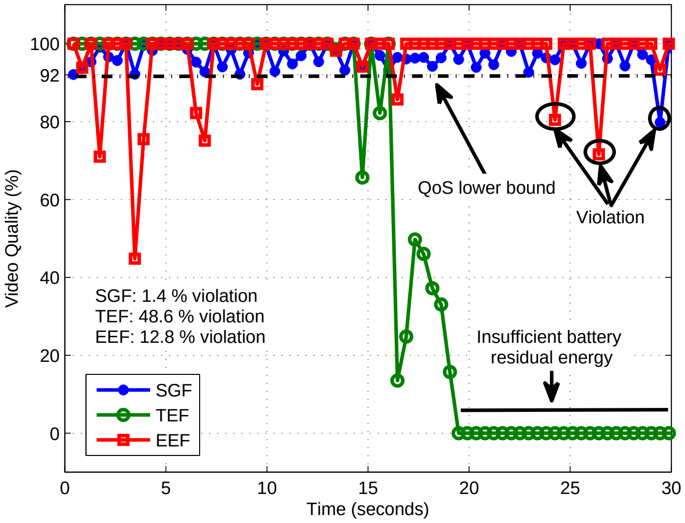


Figure 6.6: The achieved video quality versus time.

management mechanism is referred to as statistical guarantee framework (SGF), while the first benchmark is referred to as total energy framework (TEF), and the second benchmark is referred to as equal energy framework (EEF). A 30-second video call is established using the three frameworks. The available energy at the beginning of the call for the three frameworks is 250 J. For the SGF, the video quality lower bound, q_l, is calculated in the call set-up, and equals to 92%, with $\varepsilon_s = 0.9$.

Figure 6.6 plots the achieved video quality over the entire call duration. The TEF uses up all the MT available energy and hence drain its battery before call completion. This is because the TEF main objective is to maximize the video quality in the current time slot, without considering the impact of the consumed energy on the video quality in the remaining time slots. The EEF takes into consideration the call duration by equally distributing the MT available energy over the remaining time slots. However, due to the time-varying video packet encoding and channel conditions at the different radio interfaces, using this uniform energy budgets leads to inconsistent temporal fluctuations

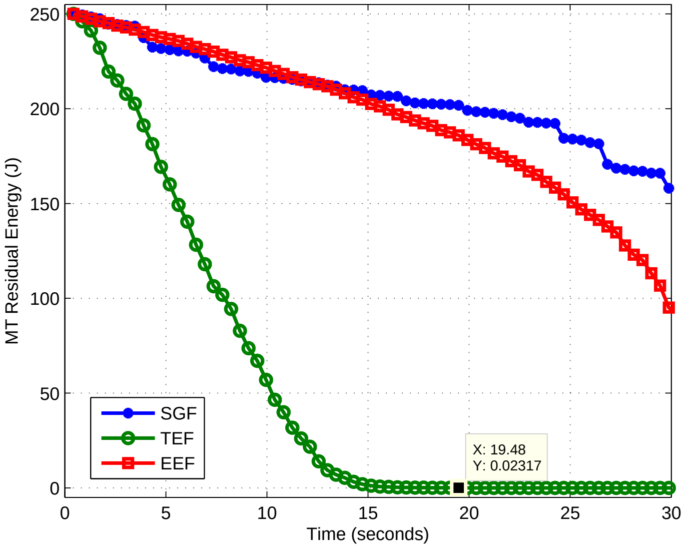


Figure 6.7: The MT residual energy versus time.

in the video quality. The resulting video quality for some time slots can be lower than 50% as shown in the figure. On the other hand, the SGF can adapt the MT consumed energy at every time slot according to the packet encoding and channel conditions at the two radio interfaces. As a result, the SGF can support a consistent video quality over different time slots, which is at least equals to the target lower bound (92%). As shown in figure, unlike the two benchmarks, the SGF can control the resulting QoS lower bound violation probability to be smaller than $\varepsilon = 0.1$.

Figure 6.7 plots the MT residual energy over the entire call duration. The MT residual energy using the TEF in the last third period of the call is insufficient to support video transmission. Since the EEF uses a uniform energy budget for different time slots regardless of the channel fading, the slope of the consumed energy is almost constant over the first two thirds of the call period. In the last third period of the call, a larger energy budget per time slot is used due to the accumulated energy that is unused over previous time slots. For the SGF, the MT consumed energy does not have an equal slope as the MT adapts its energy consumption based on video packet encoding and channel conditions at the different radio interfaces over the time slot. Overall, the SGF has the largest residual energy among the three strategies, as it uses the minimum energy consumption that satisfies the target video quality lower bound.

The advantages of the SGF over the two benchmarks can be summarized as follows: 1) The SGF guarantees a sustainable multi-homing video transmission over the target call duration, unlike the TEF; 2) The SGF supports a consistent video quality over different time slots through adapting its energy consumption according to the video packet encoding and channel conditions at different radio interfaces, and as a result, the SGF can control the QoS lower bound violation probability; and 3) The SGF consumes the least energy to provide the target video quality.

6.5 Summary

In this chapter, an energy management mechanism is proposed to support a sustainable multi-homing video transmission, over the target call duration, in a heterogeneous wireless access medium. The proposed mechanism aims to satisfy a target video quality lower bound that is calculated in the call set-up, given the MT available energy at the beginning of the call, the available bandwidth and time varying channel conditions at different radio interfaces, the target call duration, and the video packet characteristics in terms of distortion impact, delay deadlines, and video packet encoding statistics. Hence, the proposed mechanism enables the MT to support a consistent video quality over the target call duration with a certain success probability ε_s, via adapting its energy consumption according to the video packet encoding and channel conditions at different radio interfaces.

Chapter 7

Conclusions and Further Research

In this chapter, we summarize the main ideas and concepts presented in this thesis and highlight future research directions.

7.1 Conclusions

In this thesis, we have investigated radio resource management for bandwidth allocation, CAC, and MT energy management, in a heterogeneous wireless access medium. For bandwidth allocation and CAC, based on the analysis and discussion provided throughout this thesis, we have the following remarks:

- The heterogeneous wireless access medium creates various opportunities that can enhance the perceived QoS for mobile users. However, it is necessary to develop new radio resource management mechanisms for bandwidth allocation and CAC in order to satisfy the required QoS of different calls, while at the same time making efficient utilization of the available resources from different networks;

- One important aspect of radio resource management mechanisms for bandwidth allocation and CAC is the need to operate in a decentralized manner (i.e. without a central resource manager). This adds a desirable flexibility to the radio resource management and avoids many complications associated with the centralized solutions (e.g., creating a single point of failure);

- The radio resource management mechanisms for bandwidth allocation and CAC should give each network a higher priority in allocating its resources to its own subscribers as compared to other users. In this sense, network users can enjoy their maximum QoS but not at the expense of the network subscribers;

- Co-existence of single-network and multi-homing services in the heterogeneous wireless access medium should be considered. Hence, a radio resource management mechanism is to find the network assignment for MTs with single-network calls and determine the corresponding bandwidth allocation for MTs with single-network and multi-homing calls;

- The stochastic user mobility and call traffic models are necessary for designing the decentralized radio resource management mechanisms for bandwidth allocation and CAC, so as to investigate their associated impact on the system in terms of signalling overhead and processing time complexity;

- Short-term call traffic load prediction and network cooperation can help to reduce the amount of signalling overhead that is expected in a decentralized architecture;

- There are two performance metrics in radio resource management for bandwidth allocation and CAC, namely the amount of allocated bandwidth per call and the corresponding call blocking probability. In this thesis, we focus on the existing trade-off between these two metrics and present two design parameters that, when appropriately chosen, can strike a balance between the amount of allocated bandwidth per call and the target call blocking probability;

- MTs should play an active role in the resource management operation, instead of being passive service recipients in the networking environment. The mechanisms presented in this thesis for bandwidth allocation and CAC enable an MT with single-network service to select the best wireless access network available at its location and ask for its required bandwidth from that network. In addition, an MT with multi-homing service can determine a required bandwidth share from each available network so as to satisfy its total required bandwidth.

For MT energy management to support multi-homing video transmission, we have the following remarks:

- Multi-homing video transmission can improve the perceived video quality in many aspects. However, due to the MT battery energy limitation, an energy management mechanism is required in order to efficiently utilize the MT available energy to support video transmission;
- The MT energy management mechanism should take account of the MT energy limitation and the required QoS for video streaming applications, and utilizes the multi-homing capability in the heterogeneous wireless access medium. Specifically, such a mechanism should account for the video packet characteristics in terms of distortion impact, delay deadlines, and packet dependence relation, the characteristics of the multiple wireless interfaces in terms of the channel conditions and the allocated bandwidth, and the MT battery energy limitation;
- In the presence of time varying wireless channel conditions and time varying video packet encoding, using a fixed energy budget per time slot will lead to inconsistent

temporal fluctuations in the video quality, although it can support video transmission for the target call duration;

- An appropriate energy management mechanism should adapt the MT energy consumption at each time slot according to the channel conditions and the video packet encoding in order to support a consistent video quality over the target call duration.

7.2 Future Research Directions

This research has developed different mechanisms for decentralized radio resource management for bandwidth allocation and CAC, and for MT energy management in multi-homing video transmission. There are open issues that need further investigation in the mechanisms. These issues can be summarized as follows:

- For the PBRA and DSRA mechanisms presented in Chapters 3 and 4, it is assumed that the BSs/APs of different networks are connected by a backbone to exchange their signalling information. This backbone can be provided through the public network (the Internet). In this case, there is no guarantee that the signalling information will reach the BSs/APs on time. Hence, further investigation is required on the effect of signalling information delivery delay on the performance of the PBRA and DSRA mechanisms in terms of the allocated bandwidth per call and the call blocking probability;

- In the radio resource management mechanisms for multi-homing service, it is assumed that the number of MT radio interfaces is equal to the number of available BSs/APs from different networks. This may not be the case. If the number of BSs/APs is larger than the number of MT radio interfaces, the MT should get its service using a subset of the available BSs/APs, with cardinality that is equal to the number of MT radio interfaces. In this case, the MT should first select a subset of BSs/APs, then run the PBRA mechanism on this subset to determine the required bandwidth share from each network. The selection of the subset of BSs/APs can be based on RSS, available bandwidth, monetary cost, or using a utility function that combines several metrics. Hence, further investigation is required on the performace of the PBRA mechanism in this scenario;

- For the MT energy management mechanism in Chapter 6, it is assumed that the available bandwidth from each network is constant during the call. Yet, with call arrivals and departures, the offered bandwidth from each network fluctuates over time. Hence, the statistics of the offered bandwidth from each network should be considered in finding the video quality lower bound to be supported for the target call duration. In this case, two time scales should be considered in the study. The first is a fast time scale for the channel fading. The second is a slow time scale for the bandwidth dynamics;

- In Chapter 6, it is assumed that the MT completes its call within the same service area and no handoff is considered among adjacent service areas. Hence, an extension of the system model is required to include the user mobility among adjacent service areas. Using user mobility models and the statistics of the offered bandwidth from different networks in these service areas, more accurate results can be obtained for the video quality lower bound that can be supported for the target call duration. More complex analysis is expected in this case, as three time scales should be considered for the user mobility, bandwidth dynamics, and channel fading.

In this thesis, we mainly focus on exploiting cooperative networking in a heterogeneous wireless access medium to enhance service quality to mobile users in terms of bandwidth allocation, CAC, and MT energy management. Cooperative networking can also help to improve the overall network performance. One research direction in the context of cooperative networking is related to green radio communications [9]. This research direction is motivated by the increasing BS energy consumption of the wireless networks, which affects the annual profits of the service providers and has a significant impact on the environment due to the associated CO_2 emissions [9], [99]-[101]. Cooperative networking can help to improve the networks' energy efficiency. This can be achieved through dynamic planning where networks with overlapped coverage area can save energy by alternately switching on and off their radio resources according to call traffic load fluctuations [9]. Also, multi-homing radio resource allocation can be used to achieve energy saving by allocating the energy optimal transmission rate [102] from each network to the MT, while satisfying the MT required total transmission rate. Cooperative networking concepts need to be developed in order to enable a decentralized implementation of the aforementioned mechanisms.

Bibliography

[1] D. Cavalcanti, D. P. Agrawal, C. Cordeiro, B. Xie, and A. Kumar, “Issues in integrating cellular networks, WLANs, and MANETs: A futuristic heterogeneous wireless network,” *IEEE Wireless Commun.*, vol. 12, no. 3, pp. 30-41, June 2005.

[2] M. Ismail and W. Zhuang, ”Cooperative networking in a heterogeneous wireless medium,” *Springer Briefs in Computer Science*, Springer, New York, April 2013.

[3] W. Zhuang and M. Ismail, “Cooperation in wireless communication networks,” *IEEE Wireless Commun.*, vol. 19, no. 2, pp. 10-20, April 2012.

[4] I. F. Akylidiz, J. Xie, and S. Mohanty, “A survey of mobility management in next-generation all-IP based wireless systems,” *IEEE Wireless Commun.*, vol. 11, no. 4, pp. 16-28, June 2004.

[5] X. Yan, Y. A. Sekercioglu, and S. Narayanan, “A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks” *Computer Networks*, vol. 54, no. 11, pp. 1848-1863, August 2, 2010.

[6] K. Pandit, A. Ghosh, D. Ghosal, and M. Chiang, “Content-aware optimization for video delivery over WCDMA,” *EURASIP J. Wireless Commun. and Networking*, July 2012.

[7] L. Golubchik, J. C. S. Lui, T. F. Tung, A. L. H. Chow, W. J. Lee, G. Franceschinis and C. Anglano, “Multi-path continuous media streaming: what are the benefits?” *Performance Evaluation*, vol. 49, no. 1, pp. 429-449, Sept. 2002.

[8] M. D. Trott, “Path diversity for enhanced media streaming,” *IEEE Commun. Magazine*, vol. 42, no. 8, pp. 80-87, Aug. 2004.

[9] M. Ismail and W. Zhuang, “Network cooperation for energy saving in green radio communications,” *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 76-81, Oct. 2011.

[10] K. Piamrat, C. Viho, A. Ksentini, and J. M. Bonnin, “Resource management in mobile heterogeneous networks: state of the art and challenges,” *Institute national de recherche en informatique et en automatique*, no. 6459, Feb. 2008.

[11] M. Ismail, W. Zhuang, and M. Yu, “Radio resource allocation for single-network and multi-homing services in heterogeneous wireless access medium,” *IEEE VTC’12*, pp. 1-5, Sept. 2012.

[12] M. Ismail and W. Zhuang, “Decentralized radio resource allocation for single-network and multi-homing services in cooperative heterogeneous wireless access medium,” *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 4085-4095, Nov. 2012.

[13] M. Ismail and W. Zhuang, “A distributed multi-service resource allocation algorithm in heterogeneous wireless access medium,” *IEEE J. Select. Areas Commun.*, vol. 30, no. 2, pp. 425-432, Feb. 2012.

[14] M. Ismail and W. Zhuang, “A distributed resource allocation algorithm in heterogeneous wireless access medium,” *Proc. IEEE ICC’11*, June 2011.

[15] M. Ismail, A. Abdrabou, and W. Zhuang, "Cooperative decentralized resource allocation in heterogeneous wireless access medium," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 714-724, Feb. 2013.

[16] G. Miao, N. Himayat, Y. Li, and A. Swami, "Cross-Layer optimization for energy-efficient wireless communications: a survey," *Wiley J. Wireless Commun. and Mobile Computing*, vol. 9, pp. 529-542, March 2009.

[17] G. P. Perrucci , F. H.P. Fitzek, and J. Widmer, "Survey on energy consumption entities on the smartphone platform," in *Proc. IEEE VTC'11*, pp. 1-6, May 2011.

[18] E. Rantalai, A. Karppanen, S. Granlund, and P. Sarolahti, "Modeling energy efficiency in wireless internet communication," in *Proc. MobiHeld '09. ACM*, pp. 67-72, Aug. 2009.

[19] F. Fu and M. van der Schaar, "Structural solutions for dynamic scheduling in wireless multimedia transmission," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, no. 5, pp. 727-739, May 2012.

[20] S. P. Chuah, Z. Chen, and Y. P. Tan, "Energy-efficient resource allocation and scheduling for multi-cast of scalable video over wireless networks," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1324-1336, April 2012.

[21] Q. Du and X. Zhang, "Statistical QoS provisionings for wireless unicast/multicast of multi-layer video streams," *IEEE J. Select. Areas Commun.*, vol. 28, no. 3, pp. 420-433, April 2010.

[22] M. Ismail, W. Zhuang, and S. Elhedhli, "Energy and content aware multi-homing video transmission in heterogeneous networks," *IEEE Trans. Wireless Commun.*, to appear.

[23] M. Ismail and W. Zhuang, "Statistical QoS guarantee for wireless multi-homing video transmission," *IEEE GLOBECOM'13*, to appear.

[24] M. Ismail and W. Zhuang, "Mobile terminal energy management for sustainable multi-homing video transmission," *IEEE Trans. Wireless Commun.*, under review.

[25] 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 Trans. Vehicular Technology*, vol. 59, no. 6, pp. 3047-3056, July 2010.

[26] W. Shen, and Q. Zeng, "Resource management schemes for multiple traffic in integrated heterogeneous wireless and mobile networks," in *Proc. 17th Int. conf. ICCCN*, pp. 105-110, August 2008.

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

[28] K. Chebrolu and R. Rao, "Max-min fairness based radio resource management in fourth generation heterogeneous networks," in *Proc. 9th International Symposium on Communications and Information Technology*, pp. 208-213, 2009.

[29] I. Blau, G. Wunder, I. Karla, and R. Sigle, "Decentralized utility maximization in heterogeneous multicell scenarios with interference limited and orthogonal air interfaces," *EURASIP Journal on Wireless Communications and Networking*, 2009.

[30] C. Luo, H. Ji, and Y. Li, "Utility-based multi-service bandwidth allocation in the 4g heterogeneous wireless access networks," in *Proc. IEEE WCNC'09*, 2009.

[31] D. Niyato and E. Hossain, “Bandwidth allocation in 4G heterogeneous wireless access networks: A noncooperative game theoretical approach,” in *Proc. IEEE GLOBECOM’06*, 2008.

[32] D. Niyato and E. Hossain, “A noncooperative game-theoritic framework for radio resource management in 4G heterogeneous wireless access networks,” *IEEE Trans. Mobile Computing*, vol. 7, no. 3, pp. 332-345, March 2008.

[33] D. Niyato and E.Hossain , “A cooperative game framework for bandwidth allocation in 4G heterogeneous wireless networks,” in *Proc. IEEE ICC’06*, 4357-4362, 2006.

[34] C. Truong, T. Geithner, F. Sivrikaya and S. Albayrak, “Network level cooperation for resource allocation in future wireless networks,” in *Proc. 1st IFIP Wireless Days*, 2008.

[35] S. Mohanty and I. F. Akylidiz, “A cross-layer (layer 2 + 3) handoff management protocol for next generation wireless systems,” *IEEE Trans. Mobile Computing*, vol. 5, no. 10, pp. 1347-1360, 2006.

[36] C. Chi, X. Cai, R. Hao, and F. Liu, “Modeling and analysis of handover algorithms,” in *Proc. IEEE GLOBECOM’07*, pp. 4473-4477, USA,

[37] W. Shen and Q. Zeng, “Resource management schemes for multiple traffic in integrated heterogeneous wireless and mobile networks,” *Proc. 17th Int. Conf. ICCCN*, pp. 105-110, Aug. 2008.

[38] E. S. Navarro, Y. Lin, and W. S. Wong, “An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks,” *IEEE Trans. Vehicular Technology*, vol. 57, no. 2, pp. 1243-1254, March 2008.

[39] Q. T. N. Vuong, Y .G . Doudane, and N. Aqoulmine, "On utility models for access network selection in wireless heterogeneous networks," in *Proc. IEEE/IFIP Network Operations and Management Symposium*, pp. 144-151, 2008.

[40] K. Chebrolu and R. Rao, "Bandwidth aggregation for real time applications in heterogeneous wireless networks," *IEEE Trans. Mobile Computing*, vol. 5, no. 4, pp. 388-402, April 2006.

[41] R. Litjens, H. van den Berg, and R. J. Boucherie, "Throughputs in processor sharing models for integrated stream and elastic traffic," *Performance Evaluation*, vol. 65, no. 2, pp. 152-180, Feb. 2008.

[42] W. Song, Y. Cheng, and W. Zhuang, "Improving voice and data services in cellular/WLAN integrated network by admission control," *IEEE Trans. Wireless Commun.*, vol. 6, no. 11, pp. 4025-4037, Nov. 2007.

[43] H. Shen and T. Basar, "Differentiated Internet pricing using a hierarchical network game model," in *Proc. 2004 American Control Conference*, pp. 2322-2327 vol.3, 2004.

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

[45] L. S. Lasdon, *Optimization theory for large systems*, Macmillan series in operations research, 1970.

[46] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as optimization decomposition: a mathematical theory of network architectures," *Proc. IEEE*, vol. 95, no. 1, pp. 255-312, January 2007.

[47] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Select. Areas Commun.*, vol. 24, no. 8, pp. 1439-1451, August 2006.

[48] D. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: framework and applications," *IEEE Trans. Automatic Control*, vol. 52, no. 12, pp. 2254-2269, December 2007.

[49] L. Koutsopoulos and G. Losifidis, "A framework for distributed bandwidth allocation in peer-to-peer networks," *Performance Evaluation*, vol. 67, no. 4, pp. 285-298, April 2010.

[50] M. H. Ahmed, "Call admission control in wireless networks: a comprehensive survey," *IEEE Commun. Surveys*, vol. 7, no. 1, pp. 50-69, 2005.

[51] E. S. Navarro, V. S. Mansouri, and V. W. S. Wong, "Handoff management and admission control using virtual partitioning with preemption in 3G cellular/802.16e interworking," *IEEE Trans. Vehicular Technology*, vol. 59, no. 1, Jan. 2010.

[52] E. S. Navarro, V. S. Mansouri, and V. W. S. Wong, "Resource sharing in an integrated wireless cellular/WLAN system," in *Proc. Canadian conference on electrical and computer engineering*, pp. 631-634, 2007.

[53] W. Song and W. Zhuang, "QoS provisioning via admission control in cellular/wireless LAN interworking," in *Proc. 2nd International Conference on Broadband Networks*, pp. 585-592, 2005.

[54] Y. Fang and Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks," *IEEE Trans. Vehicular Technology*, vol. 51, no. 2, March 2002.

[55] D. Hong and S. S. Rappaport, “Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures,” *IEEE Trans. Vehicular Technology*, vol. VT-35, no. 3, August 1986.

[56] S. J. Gerasenko, A. Rayaprolu, S. Ponnavaikko and D. K. Agrawal, “Beacon signals: what, why, how, and where,” *Computer*, vol. 34, pp. 108-110, 2001.

[57] M. Li, M. Claypool, R. Kinicki, and J. Nichols, “Characteristics of streaming media stored on the Web,” *IEEE/ACM Trans. Networking*, vol. 5, no. 4, pp. 601-626, Nov. 2005.

[58] W. Song and W. Zhuang, “Resource allocation for conversational, streaming, and interactive services in cellular/WLAN interworking,” in *Proc. IEEE GLOBECOM’07*, pp. 4785-4789, Dec. 2007.

[59] N. Benameur, S. B. Fredj F. Delcoigne, S. Oueslati-Boulahia, and J. W. Roberts,, “Integrated admission control for streaming and elastic traffic,” in *Proc. 2nd Intl. Workshop on Quality of Future Internet Services*, pp. 69-81, Sept. 2001.

[60] A. Feldmann and W. Whitt, “Fitting mixtures of exponentials to longtail distributions to analyze network performance models,” *Performance Evaluation*, vol. 31, no. 3-4, pp. 245-279, Jan. 1998.

[61] D. Gross, J. F. Shortle, J. M. Thompson and C. .M. Harris, *Fundamentals of Queueing Theory*, Wiley series in probability and statistics, Aug. 2008.

[62] M. R. H. Mandjes and P. Zuraniewski, “M/G/infinity transience, and its applications to overload detection,” *Performance Evaluation*, vol. 68, no. 6, pp. 507-527, Feb. 2011.

[63] M. Soleimanipour, W. Zhuang, and G. H. Freeman, "Optimal resource management in wireless multimedia wideband CDMA systems," *IEEE Trans. Mobile Computing*, vol. 1, no. 2, pp. 143-160, April-June 2002.

[64] P. Bonami, M. Kilinc, and J. Linderoth, "Algorithms and software for convex mixed integer nonlinear programs," *Technical Report 1664*, Computer Sciences Department, University of Wisconsin-Madison, 2009.

[65] I. E. Grossmann, "Review of nonlinear mixed-integer and disjunctive programming techniques," *Optimization and Engineering*, vol. 3, no. 3, pp. 227-252, Sept. 2002.

[66] M. Tawarmalani and N. .V. Sahinidis, "Global optimization of mixed integer nonlinear programs: a theoritical and computational study," *Math. Program.*, vol. 99, no. 3, April 2004.

[67] M. Schluter, J. A. Egea, and J. R. Banga, "Extended ant colony optimization for non-convex mixed integer nonlinear programming," *Comput. Oper. Res.*, vol. 36, no. 7, pp. 2217-2229, 2009.

[68] M. R. Bussieck and S. Vigerske, "MINLP Solver Software," *Wiley Encyclopedia of Operations Research and Management Science*, Apr. 2011.

[69] M. Schluter, M. Gerdts, and J. J. Ruckmann, "MIDACO: new global optimization software for MINLP," Sept. 2011.

[70] N. V. Sahinidis, "BARON: users manual," version 4, Jun. 2000.

[71] N. V. Sahinidis and M. Tawarmalani, "BARON: GAMS solver manual," May 2011.

[72] www.gams.com.

[73] A. Neumaier, O. Shcherbina, W. Huyer, and T. Vinko, "A comparison of complete global optimization solvers," *Math. Program.*, vol. 103, pp. 335-356, 2005.

[74] M. C. Ferris, R. Jain, and S. Dirkse, "GDXMRW: interfacing GAMS and MATLAB," Feb. 2011.

[75] W. Song, "Resource allocation for cellular/WLAN integrated networks," in *PhD thesis*, University of Waterloo, 2007.

[76] A. Dua, C. W. Chan, N. Bambos, and J. Apostolopoulos, "Channel, deadline, and distortion (CD^2) aware scheduling for video streams over wireless," *IEEE Trans. Wireless Commun.*, vol. 9, no. 3, pp. 1001-1011, March 2010.

[77] T. Y. Hung, Z. Chen, and Y. P. Tan, "Packet scheduling with playout adaptation for scalable video delivery over wireless networks," *J. Vis. Commun. Image R.*, vol. 22, pp. 491-503, June 2011.

[78] P. Pahalawatta, R. Berry, T. Pappas, and A. Katsaggelos, "Content-aware resource allocation and packet scheduling for video transmission over wireless networks," *IEEE J. Select. Areas Communications*, vol. 25, no. 4, pp. 749-759, May 2007.

[79] M. F. Tsai, N. Chilamkurti, J. H. Park, and C. K. Shieh, "Multi-path transmission control scheme combining bandwidth aggregation and packet scheduling for real-time streaming in multi-path environment," *IET Commun.*, vol. 4, no. 8, pp. 937-945, 2010.

[80] D. Jurca and P. Frossard, "Video packet selection and scheduling for multipath streaming," *IEEE Trans. Multimedia*, vol. 9, no. 3, pp. 629-641, April 2007.

[81] W. Wang and S. Shin, "A new green-scheduling approach to maximize wireless multimedia networking lifetime via packet and path diversity," *Relaible and Autonomous Computational Science*, Part 2, pp. 167-180, 2010.

[82] I. Politis, M. Tsagkaropoulos, T. Dagiuklas, and S. Kotsopoulos, "Power efficient video multipath transmission over wireless sensor networks," *Mobile Netw Appl*, vol. 13, pp. 274-284, 2008.

[83] G. Ji, B. Liang, and A. Saleh, "Buffer schemes for VBR video streaming over heterogeneous wireless networks," in Proc. *IEEE ICC'09*, pp. 1-6, June 2009.

[84] W. Song and W. Zhuang, "Performance analysis of probabilistic multipath transmission over video streaming traffic over multi-radio wireless devices," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1554-1564, 2012.

[85] M. Ghareeb, "About multiple paths video-streaming: state of the art," *Institute de recherche en informatique et systemes aleatoires*, no. 1905, Oct. 2008.

[86] J. Chakareski, S. Han, and B. Girod, "Layered coding vs. multiple description for video streaming over multiple paths," *Proc. eleventh ACM Intl. Conf. on Multimedia*, pp. 422-431, 2003.

[87] D. Fiems, B. Steyaert, and H. Bruneel, "A genetic approach to Markovian characterisation of H.264 scalable video," *Multimedia Tools Appl.*, vol. 58, no. 1, pp. 125-146, 2012.

[88] M. van der Schaar and D. Turaga, "Cross-layer packetization and retransmission strategies for delay-sensitive wireless multimedia transmission," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 185-197, Jan. 2007.

[89] M. R. Garey and D. S. Johnson, *Computers and intractability: a guide to the theory of NP-completeness*, W. H. Freeman and Company, New York, 1979.

[90] S. Elhedhli, "Exact solution of class of nonlinear knapsack problems," *Operations Research Letters*, vol. 33, pp. 615-624, 2005.

[91] H. Kellerer, U. Pferschy, and D. Pisinger, *Knapsack problems*, Springer, 2004.

[92] S. Martello and P. Toth, "Heuristic algorithms for the multiple knapsack problem," *Computing*, vol. 27, no. 2, pp. 93-112, 1981.

[93] N. Samphaiboon and T. Yamada, "Heuristic and exact algorithms for the precedence-constrained knapsack problem," *J. Optimization Theory and Applications*, vol. 105, no. 3, pp. 659-676, June 2000.

[94] W. Simpson, *Video over IP*, Elsevier, 2008.

[95] W. Dang, M. Tao, H. Mu, and J. Huang, "Subcarrier-pair based resource allocation for cooperative multi-relay OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1640-1649, May 2010.

[96] H. T. Cheng and W. Zhuang, "An optimization framework for balancing throughput and fairness in wireless networks with QoS support," *IEEE Trans. Wireless Commun.*, vol. 7, no. 2, pp. 584-593, June 2008.

[97] M. Affenzeller, S. Wagner, and S. Winkler, *Genetic algorithms and genetic programming: modern concepts and practical applications*, vol. 6. Chapman and Hall/CRC, 2009.

[98] D. W. K. Ng and R. Schober, "Resource allocation and scheduling in multi-cell OFDMA systems with decode-and-forward relaying," *IEEE Trans. Wireless Commun.*, vol. 10, no. 7, pp. 2246-2258, July 2011.

[99] Z. Zheng, L. X. Cai, R. Zhang, and X. Shen, "RNP-SA: joint relay placement and sub-carrier allocation in wireless communication networks with sustainable energy," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3818-3828, Oct. 2012.

[100] L. X. Cai, Y. Liu, H. Luan, X. Shen, J. W. Mark, and H. V. Poor, "Dimensioning network deployment and resource management in green mesh networks," *IEEE Wireless Commun. Magazine*, vol. 18, no. 5, pp. 58-65, Oct. 2011.

[101] R. Lu, X. Li, X. Liang, X. Lin, and X. Shen, "GRS: the green, reliability, and security of emerging machine to machine communications," *IEEE Wireless Commun.*, vol. 49, no. 4, pp. 28-35, Apr. 2011.

[102] W. Wang, X. Wang, and A. A. Nilsson, "Energy-efficient bandwidth allocation in wireless networks: algorithms, analysis, and simulations," *IEEE Trans. Wireless Commun.*, vol. 5, no. 5, pp. 1103 1114, May 2006.