A Generalized Cooperative and Cognitive Radio Resource Management Architecture for Future Wireless Networks

Eng Hwee Ong

B.E. (Elect.) Hons. I

A thesis submitted in partial fulfilment of
the requirements for the degree of

Doctor of Philosophy

School of Electrical Engineering
and Computer Science

THE UNIVERSITY OF
NEWCASTLE
AUSTRALIA

August 2010

STATEMENT OF ORIGINALITY

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying subject to the provisions of the Copyright Act 1968.

ACKNOWLEDGEMENT OF AUTHORSHIP

I hereby certify that the work embodied in this thesis is the result of original research and contains several published papers/scholarly work which have been co-authored with none other than my academic supervisors based on normal candidate-supervisor practice.

Eng Hwee Ong

To my parents,

Ong Chon Seng and Nyoe Ah Moi

ACKNOWLEDGMENTS

My deepest gratitude goes to my supervisor Dr. Jamil Khan for his invaluable guidance throughout my doctoral studies. I sincerely thank him for giving me an opportunity to work together in exploring new frontiers of future wireless world. He has given me the latitude to embark on intriguing ideas, steering me when necessary. I am appreciative of his continuous encouragement, wise advices, and long hours of intense discussions. His dedicated supervision is directly reflected in the outcome of this thesis.

I wish to acknowledge the Commonwealth Department of Education, Science and Training, and The University of Newcastle for providing the EIPRS and UNRS-C Scholarships to fund and fulfill my wish in pursuing this doctoral research.

I am grateful to Dr. David Allingham for his teachings and insights on Bayesian data analysis beyond his academic duties. I would like to thank Dr. Xiangwei Zhuo for many fruitful conversations which have given me valuable inspirations to the work of Chapter 3. I am especially thankful to Dr. Kaushik Mahata for numerous enlightening discussions. Without his immense knowledge, patience, and guidance, the work in Chapter 6 of this thesis would not have been possible.

Thanks to Meng Wang, Jinchuan Zheng, Nigel Loh, Kim Siang Ang, Kexiu Liu, Yuen Yong, Xinzhi Yan, Yinghui Liu, and the fellow committee members of the Singapore Students Society for their friendships during my years in Newcastle. Your presence have spiced up my life apart from the regular hours in the laboratories.

I am also grateful to my fiancee, Lee Hui Koh, for her unconditional support during this long, undulating journey and the countless hours spent in reading the chapters. I am deeply indebted to my parents who have inculcated me values that are precious, irrespec-

tive of time and place. Special thanks to my sister, Li Ying Ong, for taking care of our family during my absence. Their love, patience, tolerance, understanding, and encouragement in times of difficult moments have spurred me to front new challenges.

ABSTRACT

Heterogeneity and *convergence* are two distinctive connotations of future wireless networks. Multiple access networks are expected to converge in a manner where heterogeneity can be exploited as an enabler to realize the *Optimally Connected, Anywhere, Anytime* vision of the International Telecommunication Union (ITU). This stimulates current trends toward the convergence of complementary heterogeneous access networks in an all-internet protocol (IP) core network and raises the importance of *cooperation* in such a multiple radio access technology (multi-RAT) environment.

This thesis defines, develops, implements, and analyzes a novel generalized cooperative and cognitive RRM (CCRRM) architecture, anchored on the key principle of *technology agnostic approach*, to optimize radio resources usage, maximize system capacity, and improve quality of service (QoS) in future wireless networks. A novel measurement-based network selection technique, formulated based on mathematical framework, and terminal-oriented network-assisted (TONA) handover architecture are the main actors of this technology agnostic approach. In particular, QoS parameters estimation is a *cornerstone* of the generalized CCRRM architecture to facilitate technology abstraction and provide link layer cognition in an effort to realize seamless mobility in future wireless networks.

By leveraging on the cooperative exchange of QoS context information over the converged all-IP core and novel concept of *reactive QoS balancing (RQB)* to achieve the end-to-end goal of promoting a *QoS-balanced system*, three RQB algorithms augmented with multi-domain cooperation techniques are developed to exploit the heterogeneity of access networks and distribute load opportunistically. Additionally, the radio resource management (RRM) design of the generalized CCRRM architecture is based on a network-terminal distributed decision making process, *similar and compliant* to the recent IEEE 1900.4 standard.

Performance evaluation is conducted with comprehensive discrete event based simulation studies to gain insights of the promising intrinsic benefits associated with RQB under realistic, pragmatic scenarios. Furthermore, an elegant unified analytical model is developed to obtain the key performance metrics for the IEEE 802.11 distributed coordination function (DCF) infrastructure basic service set (BSS), under *non-homogeneous* conditions, by integrating a Markov chain model in conjunction with a finite queueing model. These performance metrics serve as bounds for reliable capacity analysis from which a model-based predictive QoS balancing (PQB) algorithm is developed as a benchmark for comparative performance studies with the proposed measurement-based RQB algorithm.

The contributions of this thesis are not restricted to multiple access point (multi-AP) wireless local area network (WLAN), and the proof of concept is validated based on a heterogeneous multi-AP WLAN where appropriate. Moreover, conditions under which the generalized CCRRM architecture provides abstraction from underlying technologies and stays relevant to future IP-based multi-RAT environment have been established.

LIST OF ABBREVIATIONS AND ACRONYMS

i.i.d.	independent and identically distributed
w.r.t.	with respect to
3GPP	Third Generation Partnership Project
4G	fourth generation
ABC	always best connected
ACS	ambient control space
ACK	acknowledgment
AIFS	arbitration interframe space
AIFSN	arbitration interframe space number
ANI	ambient network interface
AP	access point
APC	access point controller
ARI	ambient resource interface
ASI	ambient service interface
AWGN	additive white Gaussian noise
B3G	beyond third generation
BER	bit error rate
BPSK	binary phase shift keying
BSS	basic service set
CBR	constant bit rate
CCRRM	cooperative and cognitive RRM
CDF	cumulative distribution function
CLM	cooperative load metric
COGNITION	cognition incorporating cooperation
CoRe	COGNITION relationships
CSMA/CA	carrier sense multiple access with collision avoidance
CTS	clear to send
CU	channel utilization
CUSUM	cumulative sum
CW	contention window

CWN	composite wireless network
DANS	dynamic access network selection
DAPU	data acquisition and processing unit
DCF	distributed coordination function
DIFS	distributed (coordination function) interframe space
DL	downlink
DSSS	direct sequence spread spectrum
EDCA	enhanced distributed channel access
EDCAF	enhanced distributed channel access function
ERP	extended rate PHY
ERP-OFDM	extended rate PHY using orthogonal frequency division multiplexing modulation
EWMA	exponential weighted moving average
FER	frame error rate
FIFO	first in first out
FTP	file transfer protocol
GSM	global system for mobile communications
HCUFO	highest channel utilization first out
HHO	horizontal handover
HOL	head-of-line
HR/DSSS	high rate direct sequence spread spectrum using the long preamble and header
IBSS	independent basic service set
iLB	integrated load balancing
IMT	International Mobile Telecommunications
IP	internet protocol
ITU	International Telecommunication Union
KPI	key performance indicator
LAP	load adaptation policy
LAS	load adaptation strategy
LBI	load balance index
LBM	load balancing mechanism
LTE	long term evolution
MAC	medium access control
MACK	missed acknowledgment
MADM	multi-attribute decision making
MAHO	mobile-assisted handover
MCHO	mobile-controlled handover

MIH	media independent handover
MIP	mobile IP
MIRAI	multimedia integrated network by radio access innovation
MPEG	motion picture experts group
MSDU	MAC service data unit
multi-AP	multiple access point
multi-RAT	multiple radio access technology
NACK	negative acknowledgment
NAHO	network-assisted handover
NCHO	network-controlled handover
NLOS	non-line-of-sight
NP	non-deterministic polynomial-time
NRM	network reconfiguration manager
NRT	non-real-time
OFDM	orthogonal frequency division multiplexing
OSM	operator spectrum manager
PD	packet delay
PER	packet error rate
PHY	physical layer
PLB	predictive load balancing
PLCP	physical layer convergence procedure
PLR	packet loss rate
PQB	predictive QoS balancing
QAM	quadrature amplitude modulation
QBI	QoS balance index
QLO	QoS-inspired load optimization
QoS	quality of service
QPSK	quadrature phase shift keying
QSF	QoS satisfaction factor
QSTA	QoS station
RAN	radio access network
RATs	radio access technologies
RE	radio enabler
RMC	RAN measurement collector
RQB	reactive QoS balancing
RRC	RAN reconfiguration controller
RRM	radio resource management
RSSI	received signal strength indicator

RT	real-time
RTP	real-time transport protocol
RTS	request to send
SAW	simple additive weighting
SDR	software defined radio
SIFS	short interframe space
SINR	signal-to-interference and noise ratio
SNR	signal-to-noise ratio
STA	station
TCP	transmission control protocol
TFI	throughput fairness index
TONA	terminal-oriented network-assisted
TOPSIS	technique for order preference by similarity to ideal solution
TMC	terminal measurement collector
TRC	terminal reconfiguration controller
TRM	terminal reconfiguration manager
TXOP	transmission opportunity
UDP	user datagram protocol
UL	uplink
UMTS	universal mobile telecommunications system
VBR	variable bit rate
VHO	vertical handover
VoIP	voice over internet protocol
VoWLAN	voice over WLAN
WiMAX	worldwide interoperability for microwave access
WLAN	wireless local area network

Table of Contents

CHAPTER 1

INTRODUCTION

Future wireless networks would be radically different from today's independent radio access technologies (RATs) through the widely accepted notion of convergence in heterogeneity. According to the ITU's vision of *Optimally Connected, Anywhere, Anytime* [1], it aims at the integration of existing and evolving RATs to support data rates up to 100 Mbps for high mobility applications and 1 Gbps or more for nomadic mobility access. This stimulates trends toward the integration of new RATs with different characteristics to a multitude of existing independent RATs, each supporting distinct coverage, mobility, data rates, and QoS, in a supplementary way. Consequently, future wireless networks have been envisaged as a convergence platform, in which the congregation of complementary heterogeneous RATs leverages on a converged all-IP core network to create an adaptive and self-resilient network, such that multimedia services could be provisioned optimally through the most efficient access network[1] to anyone at anywhere, anytime.

Heterogeneity and *convergence* are the two distinctive connotations of future wireless networks which include heterogeneous access network convergence, heterogeneous terminal convergence, and heterogeneous service convergence. The key driver for the convergence of heterogeneous access networks is attributed to the explosive success of internet and the exponential growth of IP-based applications. The convergence of heterogeneous terminals is seen as an intrinsic byproduct of heterogeneous access network convergence which would allow end-users to have either an individual or concurrent access, known

---

[1] The terms access network and RAT are used synonymously throughout this thesis to refer to radio access network (RAN).

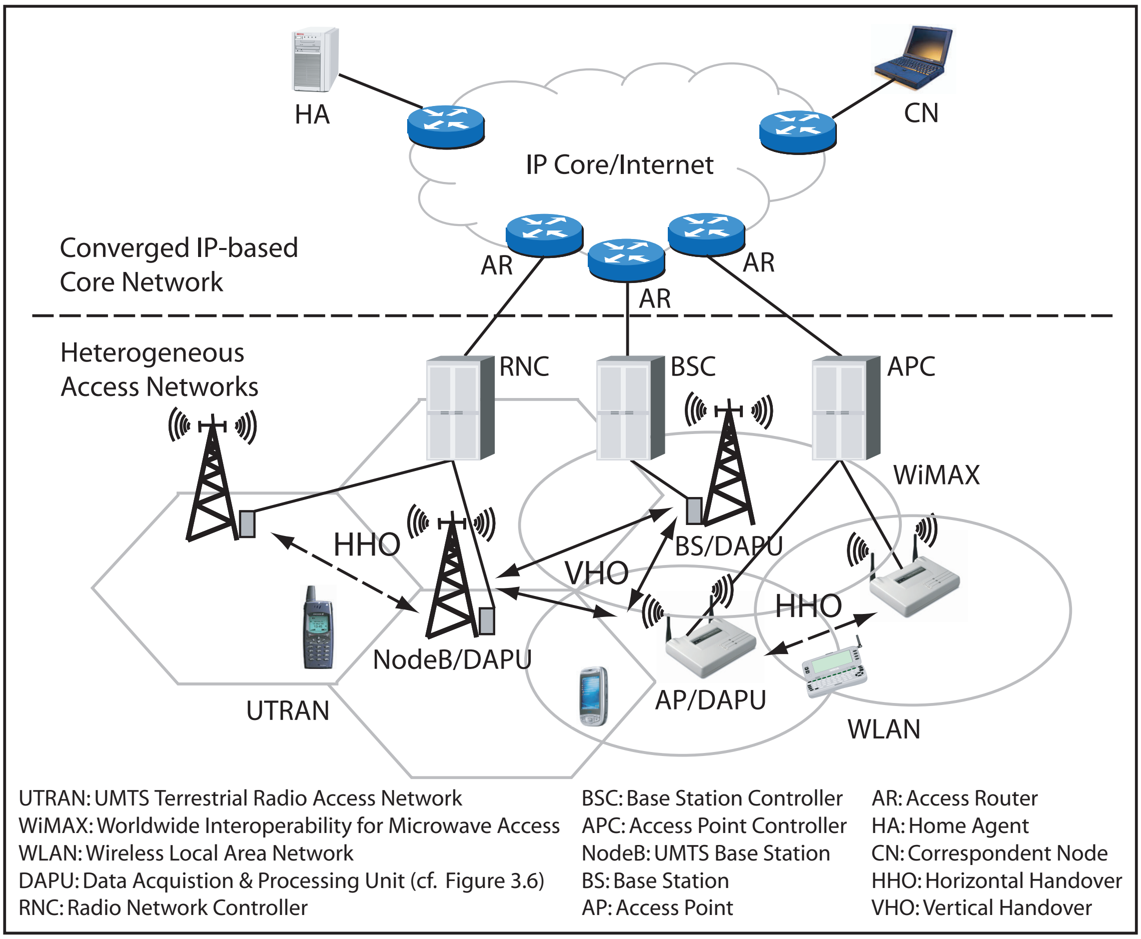

HA
CN
IP Core/Internet
Converged IP-based Core Network
AR
AR
AR
Heterogeneous Access Networks
RNC
BSC
APC
WiMAX
HHO
BS/DAPU
VHO
HHO
NodeB/DAPU
AP/DAPU
UTRAN
WLAN
UTRAN: UMTS Terrestrial Radio Access Network
WiMAX: Worldwide Interoperability for Microwave Access
WLAN: Wireless Local Area Network
DAPU: Data Acquistion & Processing Unit (cf. Figure 3.6)
RNC: Radio Network Controller
BSC: Base Station Controller
APC: Access Point Controller
NodeB: UMTS Base Station
BS: Base Station
AP: Access Point
AR: Access Router
HA: Home Agent
CN: Correspondent Node
HHO: Horizontal Handover
VHO: Vertical Handover

Figure 1.1: Future wireless networks: IP-based multi-RAT environment.

as multi-homing, to different networks within a single mobile terminal. The convergence of heterogeneous service complements both heterogeneous access networks and terminals, as well as places a strong emphasis on user-centric design so that end-users can be always best connected [2] while remaining technology agnostic. Figure 1.1 illustrates an example of future wireless networks envisioned as an IP-based multi-RAT environment comprising of universal mobile telecommunications system (UMTS), WLAN, and worldwide interoperability for microwave access (WiMAX) where anyone (end-users) can enjoy ubiquitous connectivity via the ‘best’ available access networks at anywhere, anytime to achieve seamless mobility with QoS transparency.

1.1 Trends of Future Wireless Networks

Perhaps one of the earliest known works that envisioned future wireless networks as a converged IP-based core network where different RATs congregate is reported in Seshan's doctoral thesis [3]. Following that, Stemm and Katz [4] coin the term vertical handover (VHO) and implement a VHO system based on mobile IP (MIP) to enable handover in wireless overlay networks which compose of a hierarchy of complementary access networks with differing characteristics, such as bandwidth and coverage area. In a similar vein, Wang *et al.* [5] introduce a policy-based handover system, which is an improvement over the single handover policy in [4], to consider more general policies in the context of a MIP environment.

In November 2002, the Federal Communications Commission's Spectrum Policy Task Force published a report [6] which identifies that the scarcity of spectrum is a result of legacy spectrum allocation scheme rather than the the physical lack of spectrum. This essentially means that from a frequency scan over time, there exist high possibilities that some frequency bands are largely unoccupied most of the time and others are partially utilized while the remaining frequency bands are congested. Those unoccupied frequency bands can be viewed as spectrum holes defined as a band of frequencies assigned to a primary user but not utilized by that user at a particular time and specific geographic location. This unveils an avenue to exploit the largely unoccupied and partially unused part of the radio spectrum opportunistically by the means of cognitive radio. The concept of cognitive radio is coined by Mitola [7] in his doctoral thesis as a means to improve spectrum efficiency by the exploitation of these spectrum holes through an appropriate radio etiquette. Cognitive radio, which leverages on the flexibility of software defined radio (SDR), is a radio that uses model-based reasoning to derive a goal-driven framework where it could autonomously monitor the radio environment, infer context, assess alternatives, generate plans, and learn from past experiences through the cognition cycle. The potential of cognitive radio is later reinforced by Haykin [8]. He gives a detailed exposition of signal processing and adaptive procedures based on the cognitive cycle which focuses on the interaction of the radio frequency environment and three fundamental cognitive tasks.

By far, the most prominent example of concerted efforts to provide end-users with seamless multi-access connectivity is the integrated Wireless World Initiative's ambient network project [9] of the 6th Framework Programme. Since future networks are envisioned to be heterogeneous, these networks are expected to be composed dynamically in response to the changing conditions. This creates a problem space too large for any single project. Therefore, ambient network concentrates on networking aspects while the four other sister projects, viz., (i) mobile life [10] deals with applications and services from a user-centric perspective; (ii) service platform for innovative communication environment [11] defines a unified service platform which enables cross-domain service access with service roaming support; (iii) wireless world initiative new radio [12] envisions a single ubiquitous radio access system concept whereby parameters can be adapted to a comprehensive set of mobile communication scenarios; and (iv) end-to-end reconfigurability [13] aims to achieve efficient support of ubiquitous access, pervasive services, and dynamic resources management in the heterogeneous mobile radio systems through reconfigurations.

The key challenges of the ITU's *Optimally Connected, Anywhere, Anytime* vision are: (i) *seamless mobility* for end-users roaming between different environments and RATs; and (ii) *QoS transparency* support for demanding multimedia traffic consisting of real-time (RT) and non-real-time (NRT) applications. To realize these, the heterogeneity of access networks, terminals, and services should be exploited, whenever possible during convergence, to enable better utilization of radio resources in order to improve the overall system capacity and QoS of end-users. Particularly, the exploitation of heterogeneity within the complementary future wireless networks is a natural starting point. An IP-based core network convergence would enable easy exploitation of existing MIP techniques to achieve seamless handover. However, the access network heterogeneity demands an efficient network selection scheme such that end-users can remain 'best' connected through multi-mode terminals. In addition, the possibility of moving user sessions between different RATs demands an efficient handover control to account for the QoS requirements of RT and NRT applications. Such handover control subsumes QoS-based VHO which introduces more dimensions such as QoS, load balancing, QoS balancing, user preference, and cost to the decision space as compared to radio-related horizontal handover (HHO). This postulates that end-users should remain ‘best’ connected *during* the initial network access and also *throughout* the entire duration of their connection. Such always best connected (ABC) concept [2] could be addressed by performing VHO to the next ‘best’ network that would satisfy the end-user QoS profile, delineating the need for adaptation to prevailing network conditions.

Without loss of generality, although an all-IP network makes it possible to support seamless mobility, maintaining end-user’s QoS transparency, regardless of access method and network being used, demands QoS support in order to meet end-user’s expectations in different scenarios. Moreover, the provisioning of QoS guarantee is becoming extremely important in future wireless networks as bandwidth-intensive and QoS-demanding multimedia services are expected to prevail. This leads to challenging research opportunities in terms of developing an advanced RRM architecture where techniques of cooperation and cognition could be augmented to provision seamless multimedia services delivery over heterogeneous access networks, for which the following definition is offered:

Timely delivery of differentiated services with temporal and spatial continuity to anyone, anywhere, at any time according to user preferences and prevailing network conditions in an always best connected manner, while providing statistical QoS guarantee for end-users, irrespective of radio access technologies.

1.2 Research Focus

In recent times, the research community is motivating the union of cooperative and cognitive networks in order to exploit their highly complementary characteristics. According to [14], cooperation and cognition in wireless networks have significant inter-dependencies which intensify with the increasing heterogeneity of networks, terminals, and services. In essence, this elicits that cognition depends on the data acquisition of performance metrics through cooperation while cooperation depends on the awareness of surroundings through cognition. To be more specific, cooperation could be seen as the first step toward the optimization of network performance in future wireless networks with heterogeneous access networks, terminals, and services. Thereafter, cognition becomes crucial for any beneficial interactions. Thus, cooperation in wireless networks lays foundation for cognitive functionality to fully unleash their potential benefits in future wireless networks. Although significant progress has been achieved in cooperative and cognitive networks, their advancements are mainly autonomous. The *cross-fertilization* of cooperation and cognition will become imperative as the heterogeneity of access networks, terminals, and services escalates. This creates a new set of problems and research opportunities. In particular, the relationships between cooperation and cognition in future wireless networks are still in the stage of infancy, and it is an area where this thesis makes a number of novel contributions.

This thesis defines, develops, implements, and analyzes a novel generalized CCRRM architecture as illustrated in Figure 1.2. The generalized CCRRM architecture is envisioned to harmonize various RRM functional blocks by leveraging on the *technology agnostic approach* to support access network heterogeneity and provide link layer cognition in Layer 2.5. Additionally, network context information and policies could be shared between different layers and network entities through various domains of cooperation to provide guidelines for orchestrating efficient radio resource usage. This will serve to maximize overall composite capacity[2] and improve the perceived QoS of end-users. Henceforth, the focus of this thesis is aimed toward:

- Technology agnostic approach to support access network heterogeneity and provide link layer cognition to facilitate handover initiation and network selection for

---

[2] The terms overall composite capacity and overall system capacity are used synonymously throughout this thesis to refer to the aggregate capacity of a multi-AP WLAN or multi-RAT environment.

the coordination of informed VHO[3] in future wireless networks, envisaged as a multi-RAT environment.

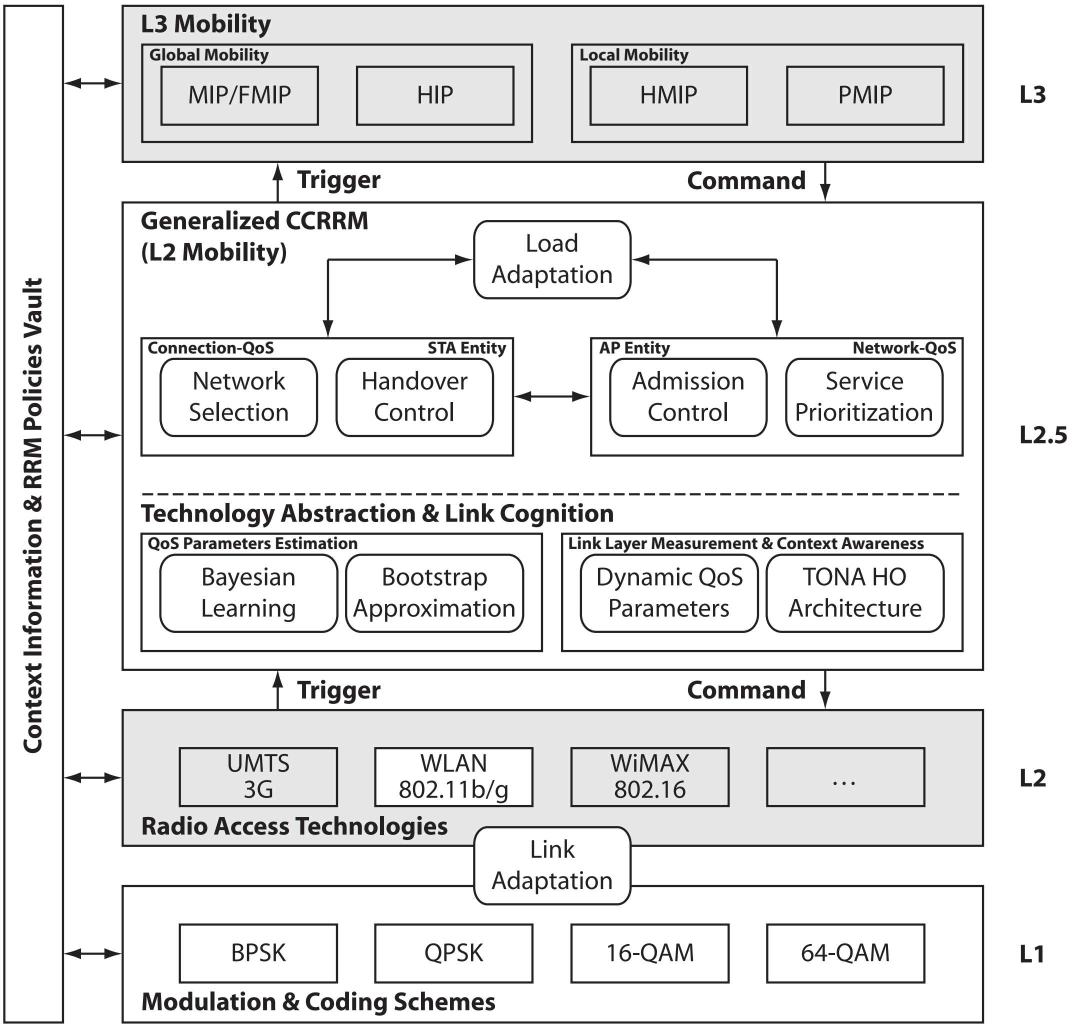
- Layer 2 handover latency reduction to achieve seamless mobility and QoS transparency support, regardless of underlying technologies.
- Improvement of overall composite capacity and QoS of end-users through dynamic load distribution based on network-terminal distributed RRM decision to optimize radio resource usage in which VHO is used as a vehicle to exploit heterogeneity opportunistically.

It is important to note that this thesis focuses on optimizing Layer 2 (link layer) mobility by reducing handover latency associated with the detection of changes in link states and scanning procedures, and incorporating distributed RRM decision making functions between network-terminal entities. Although Layer 3 (network layer) mobility is not explicitly considered, the technology agnostic approach adopted in this thesis can be readily exploited to trigger handover preparation procedures at the network layer for facilitating movement detection and expediting IP reconfiguration process [17], [18], as well as performance enhancement of mobility management protocols [19]. Such cooperation between link layer events to trigger network layer mobility functions is currently being addressed by the handover for ubiquitous and optimal broadband connectivity among cooperative networking environments project [20] of the 7th Framework Programme to provide seamless inter-technology mobility.

On the other hand, proxy MIP [21] is the latest mobility solution proposed by Internet Engineering Task Force to provide network-based localized mobility management to meet the demands of future wireless networks. Unlike MIP, fast MIP, and hierarchical MIP,

---

[3] The widely accepted notion of VHO involves handover between different technologies, e.g., the IEEE 802 wireless networks and UMTS cellular networks. In the more general case, VHO can be defined as an asymmetric handover process between networks of differing characteristics (see, e.g., Chapter 1 of [15]) in which it also encompasses handover between heterogeneous IEEE 802.11 wireless networks as specified in the recent IEEE 802.21 standard [16]. Such QoS-based handover is the context of VHO considered and addressed throughout this thesis.



MIP: Mobile IP
FMIP: Fast Mobile IP
HIP: Host Identity Protocol
HMIP: Hierarchical Mobile IP
PMIP: Proxy Mobile IP
QoS: Quality of Service
HO: Handover

UMTS: Universal Mobile Telecommunications System
WLAN: Wireless Local Area Network
WiMAX: Worldwide Interoperability for Microwave Access
BPSK: Binary Phase Shift Keying
QPSK: Quadrature Phase Shift Keying
QAM: Quadrature Amplitude Modulation
TONA: Terminal-Oriented Network Assisted

Figure 1.2: A generalized CCRRM architecture.

which are host-based, proxy MIP does not involve any IP reconfigurations in the end-host, i.e., the end-host is able to maintain a fixed IP address as it roams within the same proxy MIP domain. In fact, proxy MIP solution is inspired partially by the huge success of WLAN switches which perform localized management without modifications to the end-host's IP stack. Another advantage of such network-based solution is the independence from global mobility protocols [22] such as MIP or host identity protocol, and it has been adopted in both WiMAX and long term evolution (LTE) networks. However, such localized mobility management at the IP layer will still require an equally optimized Layer 2 mobility solution, such as the one embodied in this thesis, to fully optimize handover in a complementary fashion and achieve seamless inter-technology mobility in future wireless networks [23].

The concepts embodied in this thesis serves to identify and address several real world issues associated with the convergence of heterogeneity in future wireless networks. In fact, the key concept of technology agnostic approach rooted in the generalized CCRRM architecture could be complemented with the ambient network project (cf. Section 2.1.2) which focuses in establishing inter-networking roaming agreements on the fly. Although some of the concepts developed in this thesis apply to a more general system, the proof of concept is exemplified from the standpoint of a heterogeneous multi-AP WLAN. It is argued in this thesis that the technology abstraction and link cognition module of the generalized CCRRM architecture is readily applicable to future IP-based multi-RAT environment seeking unification via cooperation and radio resource usage optimization through cognition. This is possible as QoS parameters estimation, which is a cornerstone of the technology agnostic approach, provides a generic way to characterize the quality of wireless network and its channel, and provide link layer triggers. This provides a portal to facilitate QoS transparency by relating the QoS requirements of end-user to underlying QoS of system. Furthermore, the measurement-based nature of the technology agnostic approach implies that it is applicable to any given wireless networks, irrespective of their access heterogeneity.

The IEEE 802.11[4] WLAN is chosen as a platform for the proof of concept since majority of the current state of the art wireless technologies are based on the unlicensed industrial, scientific, and medical radio bands. Moreover, the lack of cooperation and cognition mechanisms in existing WLAN provides a good benchmark for evaluating any performance gains. Accordingly, the shaded blocks in Figure 1.2 are outside the scope of this thesis. Further descriptions of the IEEE 802.11 WLAN model is included in Appendix A-2.

1.3 Contributions of this Thesis

The specific contributions of each chapter are listed in their respective chapters throughout the thesis. The key contributions of this thesis are accentuated in the following.

- The *COGNITION relationships (CoRe) methodology* is developed as a concrete framework for lateral and in-depth investigations in the establishment of the relationships between cooperation and cognition from a user-centric perspective. In particular, the CoRe methodology serves to provide guidelines for innovative solutions toward the design and creation of the generalized CCRRM architecture.
- The development and implementation of the TONA handover architecture which enables *inter-network cooperation* by taking advantage of the converged IP core network to facilitate the cooperative exchange of QoS context information and dissemination of RRM policy. In addition, the TONA handover architecture is inspired by the concepts of *network-assisted discovery* and *terminal-oriented decision*, which further facilitate *inter-entity cooperation* between network-terminal entities to support distributed decision making process. The key advantages of the TONA handover architecture lie in the support of fast handover and power saving features for terminals. These will become important for future multi-mode, SDR-

---

[4] In March 2007, the IEEE 802.11 Task Group-ma created a single document that merged 8 amendments (802.11a, b, d, e, g, h, i, j) with the base standard to form the IEEE 802.11-2007 [24] which is the current base standard used throughout this thesis.

based devices since it is unlikely to perform service discovery by scanning all available networks and simultaneously operating multiple air interfaces due to delay and power constraints, respectively. To ensure interoperability with existing standard, an efficient implementation of QoS broadcast mechanism with beacon frame over the TONA handover architecture is proposed. Furthermore, an evaluation of the system cost associated with the generalized CCRRM architecture, and the tradeoffs between QoS performance and QoS broadcast intervals are also investigated.

- A low complexity, measurement-based dynamic access network selection (DANS) algorithm capable of providing a pragmatic approach to estimate dynamic QoS parameters is developed to coordinate VHO in a multi-RAT environment. In particular, the dual-stage *QoS parameters estimation* process is a *cornerstone* of the *technology agnostic approach* which provides abstraction from underlying technologies. Furthermore, it facilitates as link layer cognition to filter unnecessary handovers and achieve an optimal network selection outcome even in the presence of dynamic QoS parameters. This marks a significant step closer to ABC services for realizing seamless mobility in future wireless networks as compared to existing cost function approach which will result in system instability as a consequence of 'ping-pong' handovers when evaluating dynamic QoS parameters. Such a measurement-based network selection technique will complement the recent IEEE 802.21 media independent handover (MIH) standard [16] which does not specify any handover controls, policies, and algorithms involved in VHO decision mechanisms.

- The notion of *RQB* is defined and developed to achieve the end-to-end goal of the generalized CCRRM architecture, which aims to promote a long-term *QoS-balanced system*. A suite of RQB algorithms which leverages on various domains of cooperation is developed to exploit the heterogeneity of access networks and perform dynamic load distribution in an opportunistic yet altruistic manner. To be more specific, the integrated load balancing (iLB) scheme based on bi-domain co-operation forms the *baseline design* of the generalized CCRRM architecture. The

QoS-inspired load optimization (QLO) framework based on tri-domain cooperation features an additional *intra-layer cooperation* to optimize load distribution in a single rate WLAN. The load adaptation strategy (LAS) framework based on quad-domain cooperation introduces a supplementary *inter-layer cooperation* to mitigate rate anomaly in a multirate WLAN-based cognitive network. Extensive simulation studies with comprehensive pragmatic scenarios have revealed several favorable intrinsic properties of RQB, viz., (i) providing statistical QoS guarantee; (ii) maximizing overall system capacity by maintaining QoS and throughput fairness; (iii) precluding unnecessary handovers; (iv) mitigating the rate anomaly problem; and (v) preserving baseline QoS. Consequently, the notion of *QoS balance* is advocated as the criterion to quantify the state of balance in future wireless networks where dynamic network conditions are commonplace.

- In order to benchmark the performance of RQB algorithm to gain further engineering insights, a unified analytical model is developed to obtain the key performance metrics of medium access control (MAC) delay, packet loss rate (PLR), and throughput efficiency for the IEEE 802.11 DCF infrastructure BSS. The analytical model incorporates both Markov chain model and finite queueing model to capture *non-saturation* operating conditions. Moreover, it considers *non-homogeneous* conditions by modeling asymmetric traffic load between an access point (AP) and its associated stations (STAs)[5], heterogeneous flows between STAs, and diverse wireless channel conditions between BSSs. These performance metrics serve as bounds for reliable capacity analysis from which the model-based PQB algorithm is developed to serve as a basis for conducting comparative studies with the measurement-based predictive load balancing (PLB) and RQB algorithms. Extensive analyses and simulations also uncover that backoff freezing for an infrastructure BSS should

---

[5] The terms STAs and terminals are used synonymously throughout this thesis to refer to end-user mobile devices. Specifically, STA is used to refer to any device that contains an IEEE 802.11 compliant MAC and physical layer (PHY) interface to the wireless medium, whereas terminal is used to refer to any IEEE 1900.4 compliant radio node and also in the more general case.

be properly modeled and considered in order to derive accurate performance metrics.

- Intricate *similarities* between the TONA handover architecture and the recent IEEE 1900.4 system architecture are established. Additionally, the pertinence of the LAS framework to the IEEE 1900.4 functional architecture is exemplified in the implementation of load adaptation policy (LAP) based on the distributed radio resource usage optimization use case. Comprehensive simulation studies have shown that the IEEE 1900.4 RRM, based on the LAP, can effectively exploit the cooperative exchanges of context information between network-terminal entities to facilitate the coordinated use of radio resources which harnesses overall composite capacity and QoS improvements. Remarkably, these results are consistent to the studies in [25] (see, e.g., §3.2.4 – 3.2.5) which also show that the IEEE 1900.4 RRM yields better performances for network capacity and user perceived QoS. Furthermore, the comparative performance evaluation between the three dynamic load distribution algorithms reveals that the measurement-based RQB algorithm outperforms both the PLB and PQB algorithms to achieve higher QoS fairness and end-user throughput. It is worth noting that while the RQB algorithm is found to preserve the highly desirable *baseline QoS* property, the same property does not exist in the PLB algorithm which also relies on the measurement-based approach. Collectively, these simulation results function as an early investigation to provide insights on the performance benefits of the baseline IEEE 1900.4 standard [26] and contribute toward the emerging IEEE 1900.4.a and IEEE 1900.4.1 standards.

1.4 Thesis Outline

The remainder of this thesis is organized into a chapter of overview, five chapters of contributions, and a chapter of conclusions. Several thematic connections can be established between the chapters of contributions as illustrated in Figure 1.3. Chapter 3 in the center of Figure 1.3 forms a *cornerstone* of the generalized CCRRM architecture which

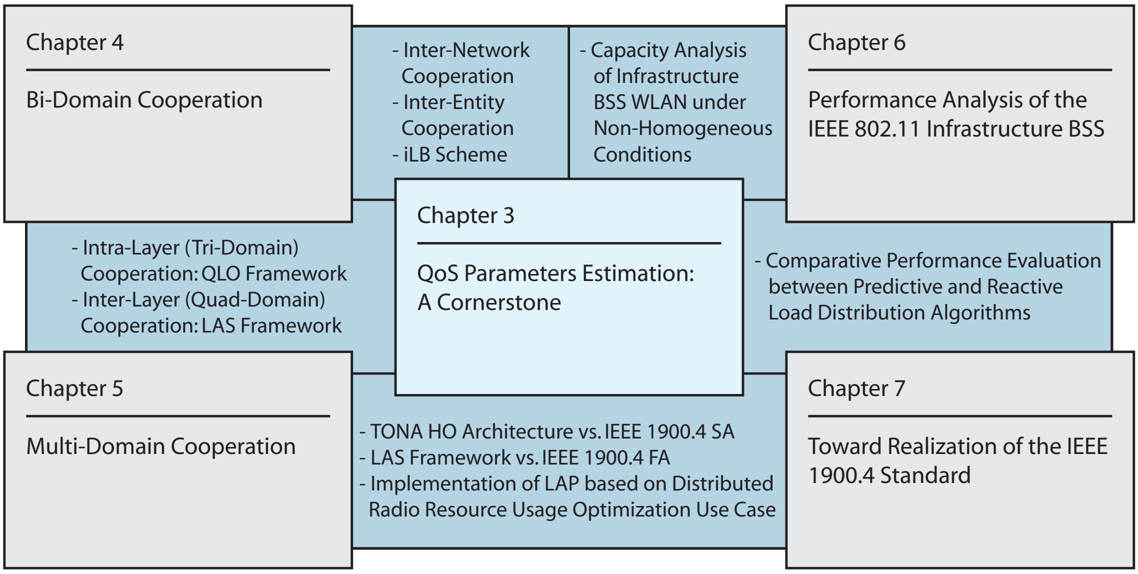

Chapter 4
Bi-Domain Cooperation
- Inter-Network Cooperation
- Inter-Entity Cooperation
- iLB Scheme
- Capacity Analysis of Infrastructure BSS WLAN under Non-Homogeneous Conditions
Chapter 6
Performance Analysis of the IEEE 802.11 Infrastructure BSS
- Intra-Layer (Tri-Domain) Cooperation: QLO Framework
- Inter-Layer (Quad-Domain) Cooperation: LAS Framework
Chapter 3
QoS Parameters Estimation: A Cornerstone
- Comparative Performance Evaluation between Predictive and Reactive Load Distribution Algorithms
Chapter 5
Multi-Domain Cooperation
- TONA HO Architecture vs. IEEE 1900.4 SA
- LAS Framework vs. IEEE 1900.4 FA
- Implementation of LAP based on Distributed Radio Resource Usage Optimization Use Case
Chapter 7
Toward Realization of the IEEE 1900.4 Standard

Figure 1.3: A Generalized Cooperative and Cognitive RRM Architecture for Future Wireless Networks: Thematic connections between five chapters of contributions.

provides technology abstraction and link layer cognition to support lateral and in-depth investigations into the novel concept of RQB over a wide range of realistic scenarios and operating conditions. Chapter 3 and Chapter 4 focus on the development and evaluation of the iLB scheme, which is the baseline design of the generalized CCRRM architecture, based on bi-domain cooperation. Chapter 3, Chapter 4, and Chapter 5 concentrate on the development and evaluation of both the QLO and LAS frameworks which extend bi-domain cooperation to multi-domain cooperation. Chapter 3, Chapter 5, and Chapter 7 elicit that the TONA handover architecture and the LAS framework are compliant to the IEEE 1900.4 standard [26] through the implementation of the LAP based on the distributed radio resource usage optimization use case. Chapter 3 and Chapter 6 are both devoted to derive the key performance metrics of an infrastructure BSS WLAN under non-homogeneous conditions, which are crucial for proper admission control and provide insights into capacity analysis of an AP. The main difference is that Chapter 3 employs the measurement-based approach, whereas Chapter 6 utilizes the model-based approach. Finally, Chapter 3, Chapter 6, and Chapter 7 examine the comparative analysis between different dynamic load distribution algorithms which are important for maximizing overall composite capacity and QoS in future wireless networks.

The relevant literature reviews associated with these chapters are organized and reported in each of the corresponding chapter. In addition, the excerpts of each chapter are elucidated as follows:

Chapter 2 presents an overview of the beyond third generation (B3G)/fourth generation (4G) research activities, research methodology, and taxonomy of the generalized CCRRM architecture.

Chapter 3 is devoted to the development and implementation of a technology abstraction and link cognition module of the generalized CCRRM architecture. This module features a novel TONA handover architecture to enable inter-network cooperation and a novel DANS algorithm to support inter-entity cooperation, collectively known as bi-domain cooperation. The DANS algorithm is a dual-stage QoS parameters estimation process, which forms a cornerstone of the technology agnostic approach, consisting of bootstrap approximation and Bayesian learning. Bootstrap approximation is a generic measurement-based technique to support universal usability and abstraction from underlying technologies while Bayesian learning provides link layer cognition to facilitate handover initiation and network selection. This QoS parameters estimation process can be used to augment existing handover decision mechanisms, which are typically formulated as a multi-attribute decision making (MADM) problem, when evaluating dynamic QoS parameters. It is shown to achieve an optimal outcome and consequently improve system stability, QoS performance, and system capacity.

Chapter 4 investigates the benefits of novel RQB philosophy based on bi-domain cooperation through the iLB scheme which forms the baseline design of the generalized CCRRM architecture. The iLB scheme constitutes a synergy between fast handover and soft admission control designed to protect the QoS of RT traffic from network overloading by exploiting heterogeneity in a multi-AP WLAN and performing load distribution in an opportunistic yet altruistic manner. The key advantages of the iLB scheme are its lightweight design to perform RQB based only on PLR and packet delay (PD), as well as its adaptability to dynamic network conditions, attributed to these QoS parameters which will be influenced by traffic conditions explicitly and wireless channel conditions implicitly. RQB has intrinsic properties of providing statistical QoS guarantee. In addition, it provides QoS and throughput fairness, which jointly maximize overall system capacity. The iLB scheme will be useful for the QoS provisioning of RT services over the legacy DCF, thanks to the fast handover design which incurs only an average Layer 2 handover latency of less than 20 ms and PLR of less than 2%. This implies than RT connections can be opportunistically redistributed while still meeting their stringent QoS requirements.

Chapter 5 extends the novel concept of RQB from bi-domain cooperation to multi-domain cooperation in seek for further exploitation of heterogeneity in a multi-AP WLAN. Through additional service prioritization and intra-layer cooperation between different RRM functional blocks, the QLO framework is presented to optimize load distribution in a single rate WLAN under dynamic network conditions. Finally, the QLO framework evolves to the LAS framework which incorporates inter-layer cooperation to exploit the benefits of both link adaptation and load adaptation on-demand to enhance multimedia service delivery in a multirate WLAN-based cognitive network, also under dynamic network conditions. Particularly, it is shown that RQB has additional intrinsic properties of mitigating the rate anomaly problem in a multirate WLAN-based cognitive network and precluding unnecessary handovers. Both the QLO and LAS frameworks are compatible with the DCF and enhanced distributed channel access (EDCA) channel access mechanisms to fully provision QoS for multimedia service delivery in a single unifying generalized CCRRM architecture.

Chapter 6 begins by analyzing the key performance metric for the IEEE 802.11 DCF infrastructure BSS using a unified analytical model that is developed by integrating a Markov chain model in conjunction with a finite queueing model. The performance metrics of an AP are of particular interest as the AP relays all traffic to and from the WLAN, and consequently is the capacity bottleneck of an infrastructure BSS. The performance analysis is carried out with a comprehensive model which captures the transition from the non-saturation to saturation mode of operation under various non-homogeneous conditions. Subsequently, the developed analytical model is used for the implementation of the PQB algorithm to serve as baseline for comparative performance analysis with the PLB and RQB algorithms where the latter is advocated in this thesis.

Chapter 7 first gives an overview of the recently approved IEEE 1900.4 standard [26] which is similar to the TONA handover architecture in many aspects. In particular, the key idea of distributed decision making process between network-terminal entities is advocated by the standard. Additionally, the applicability of the LAS framework anchored on the underlying principle of RQB to the IEEE 1900.4 standard is exemplified through an implementation of the LAP based on the distributed radio resource usage optimization use case. Next, a performance comparison between the PQB algorithm based on the unified analytical model, the PLB algorithm based on the load balancing mechanism (LBM), and the RQB algorithm based on the iLB scheme is evaluated in a voice over WLAN (VoWLAN) scenario under diverse conditions. The RQB algorithm reveals an additional intrinsic property of preserving baseline QoS, which is not found in the PLB algorithm, when comparing with the PQB algorithm.

Chapter 8 summarizes the main contributions and results presented throughout this thesis, followed by some thoughts on future research directions.

1.5 Prior Publications

Earlier versions of some proposed techniques, algorithms, and results embodied in this thesis have been published in various refereed conference proceedings and journal articles. The complete list of prior publications is given below.

I. Refereed journal articles

- E. H. Ong and J. Y. Khan. Cooperative radio resource management framework for future IP-based multiple radio access technologies environment, *Comput. Netw.*, 54(7):1083–1107, May 2010.
- E. H. Ong and J. Y. Khan. On optimal network selection in a dynamic multi-RAT environment, *IEEE Communications Letters*, 14(3):217–219, March 2010.

II. International refereed conference proceedings

- E. H. Ong, J. Y. Khan and K. Mahata. A simple model for non-homogeneous and non-saturated IEEE 802.11 DCF infrastructure BSS, To appear in *Proc. IEEE 12th International Conference on Communication Systems. ICCS 2010*, pages 1–6, November 2010.
- E. H. Ong, J. Y. Khan and K. Mahata. On dynamic load distribution algorithms for multi-AP WLAN under diverse conditions, In *Proc. IEEE Wireless Communications and Networking Conference. WCNC 2010*, pages 1–6, April 2010.
- E. H. Ong, J. Y. Khan and K. Mahata. Comparative performance analysis of dynamic load distribution algorithms in a multi-AP wireless network, In *Proc. Annual IEEE India Conference, 2009. INDICON 2009*, pages 1–4, December 2009.
- E. H. Ong and J. Y. Khan. Distributed radio resource usage optimization of WLANs based on IEEE 1900.4 architecture, In *Proc. IFIP Wireless Days Conference, 2009. WD 2009*, pages 1–6, December 2009.

- E. H. Ong and J. Y. Khan. On load adaptation for multirate multi-AP multimedia WLAN-based cognitive networks, In *Proc. IFIP Wireless Days Conference, 2009. WD 2009*, pages 1–6, December 2009.

- E. H. Ong and J. Y. Khan. A unified QoS-inspired load optimization framework for multiple access points based wireless LANs, In *Proc. IEEE Wireless Communications and Networking Conference, 2009. WCNC 2009*, pages 1–6, April 2009.

- E. H. Ong and J. Y. Khan. An integrated load balancing scheme for future wireless networks, In *Proc. IEEE Global Communications Conference Workshops, 2008. GLOBECOM Workshops '08*, pages 1–6, November/December 2008.

- E. H. Ong and J. Y. Khan. QoS provisioning for VoIP over wireless local area networks, In *Proc. IEEE 11th International Conference on Communication Systems, 2008. ICCS 2008*, pages 906–911, November 2008.

- E. H. Ong and J. Y. Khan. Dynamic access network selection with QoS parameters estimation: A step closer to ABC, In *Proc. IEEE 67th Vehicular Technology Conference, 2008. VTC 2008-Spring*, pages 2671–2676, May 2008.

CHAPTER 2

OVERVIEW OF THE GENERALIZED CCRRM ARCHITECTURE

The two distinctive features of future wireless networks have been identified as heterogeneity and convergence in the introduction. In addition, the augmentation of cooperation and cognition techniques via cross-fertilization to provision seamless multimedia services delivery over heterogeneous access networks through an advanced RRM architecture has been defined in Section 1.2. Accordingly, future wireless networks will consist of highly heterogeneous access networks where fruitful interactions require an effective acquisition of knowledge which in turn requires some forms of efficient cooperation to provide such awareness. Hence, the four fundamental building blocks of future wireless networks can be defined as follows:

- An all-IP core network which facilitates the integration of heterogeneous access networks and provides a flat, common platform to support access network heterogeneity.
- Cooperation between network-terminal entities and/or layers of protocol stack to allow effective cognition.
- Cognitive functionality in network-terminal entities and/or layers of protocol stack to enable efficient cooperation.

- Multi-mode devices that are powered by SDR to offer dynamic end-to-end reconfigurability.

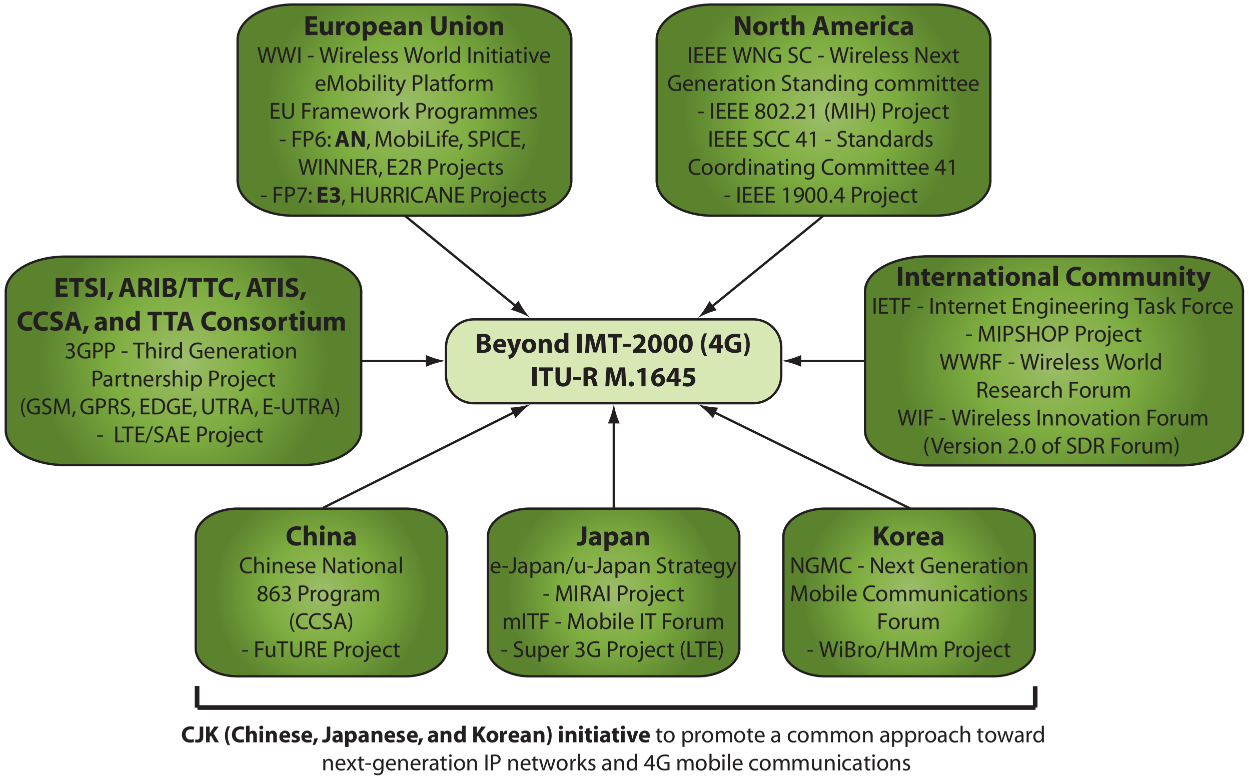
The generalized CCRRM architecture proposed in this thesis addresses the first three building blocks to provide a concrete baseline for incorporating reconfigurable multi-mode devices in the future. An overview of the generalized CCRRM architecture will follow after a review of the major B3G/4G research activities and a detailed exposition of the associated research methodology employed in its development and implementation.

This chapter is outlined as follows. Section 2.1 presents a review of the major B3G/4G research activities. Section 2.2 describes the research methodology for establishing relationships between cooperation and cognition from a user-centric perspective. Section 2.3 gives an overview of the generalized CCRRM architecture by illustrating its taxonomy, and Section 2.4 concludes this chapter.

2.1 Major B3G/4G Research Activities: An International Perspective

The inception of the ITU-R M.1645 recommendation [1] in the beginning of June 2003 has provided the future framework for International Mobile Telecommunications (IMT)-2000 and beyond which aims to develop global consensus in shaping the future wireless networks, commonly known as B3G or 4G. Since then, numerous research initiatives among the Standards Developing Organizations around the world have blossomed and evolved into some of the representative B3G/4G research activities illustrated in Figure 2.1.

In North America, the IEEE 802 LAN/MAN Standards Committee develops local area network and metropolitan area network standards mainly for the lowest two layers of the Open Systems Interconnection Reference Model. There are several IEEE 802 standards, e.g., the IEEE 802.11 WLAN, IEEE 802.16 broadband wireless access, and IEEE


European Union
WWI - Wireless World Initiative
eMobility Platform
EU Framework Programmes
- FP6: **AN**, MobiLife, SPICE, WINNER, E2R Projects
- FP7: **E3**, HURRICANE Projects
North America
IEEE WNG SC - Wireless Next Generation Standing committee
- IEEE 802.21 (MIH) Project
IEEE SCC 41 - Standards Coordinating Committee 41
- IEEE 1900.4 Project
ETSI, ARIB/TTC, ATIS, CCSA, and TTA Consortium
3GPP - Third Generation Partnership Project
(GSM, GPRS, EDGE, UTRA, E-UTRA)
- LTE/SAE Project
Beyond IMT-2000 (4G)
ITU-R M.1645
International Community
IETF - Internet Engineering Task Force
- MIPSHOP Project
WWRF - Wireless World Research Forum
WIF - Wireless Innovation Forum
(Version 2.0 of SDR Forum)
China
Chinese National 863 Program
(CCSA)
- FuTURE Project
Japan
e-Japan/u-Japan Strategy
- MIRAI Project
mITF - Mobile IT Forum
- Super 3G Project (LTE)
Korea
NGMC - Next Generation Mobile Communications Forum
- WiBro/HMm Project
CJK (Chinese, Japanese, and Korean) initiative to promote a common approach toward next-generation IP networks and 4G mobile communications

SDOs
ETSI - European Telecommunications Standards Institute (EU)
ARIB - Association of Radio Industries and Businesses (Japan)
TTC - Telecommunication Technology Committee (Japan)
ATIS - Alliance for Telecommunications Industry Solutions (North America)
CCSA - China Communication Standards Association (China)
TTA - Telecommunications Technology Association (South Korea)

Projects
AN/E3 - Ambient Networks/End-to-End Efficiency
LTE/SAE - Long Term Evolution/System Architecture Evolution
MIRAI - Multimedia Integrated Network by Radio Access Innovation
FuTURE - Future Technologies for a Universal Radio Environment
MIH - Media Independent Handover
WiBro/HMm - Wireless Broadband/High Speed Mobile Multimedia
MIPSHOP - Mobility for IP: Performance, Signaling, and Handoff Optimization

Figure 2.1: An Overview of major B3G/4G research activities around the world.

802.21 MIH that will play an integral part in future wireless networks. Additionally, the IEEE 802.11 WLAN Working Group has a Wireless Next Generation Standing Committee which concentrates on the continued evolution of WLAN technology, e.g., the IEEE 802.11n and coexistence with other technologies, e.g., the IEEE 802.19. Recently, the IEEE Standards Coordinating Committee 41 is formed to develop supporting standards related to dynamic spectrum access and advanced spectrum management. One of the recently published standards is the IEEE 1900.4 which provides the architectural building blocks for radio resource optimization in heterogeneous wireless access networks. In the European Union, the Framework Programme is the key platform for collaborative research funded by the European Commission to reinforce Europe's leadership in mobile and wireless communications, e.g., the Wireless World Initiative's ambient network (AN) project in the 6th Framework Programme and eMobility's end-to-end efficiency (E3) project in the 7th Framework Programme.

In China, the 863 program is funded and administered by the government to stimulate the development of advanced technologies, e.g., the future technologies for a universal radio environment (FuTURE) project is an important part of the wireless communication branch. In Japan, the e-Japan strategy which aims to position itself as the world's most advanced information communication technology nation by 2005, e.g., the multimedia integrated network by radio access innovation (MIRAI) project has evolved to u-Japan strategy which aims to realize a ubiquitous network society by 2010. The keyword 'u' in u-Japan represents ubiquitous, universal, user-oriented, and unique, which reflect the vision of the ITU. On the other hand, the Mobile IT Forum is created to realize an early implementation of future mobile communication systems including systems beyond IMT-2000, e.g., the super 3G project. In Korea, the Next Generation Mobile Communications Forum, which is a rebirth of earlier initiatives by the Electronics and Telecommunications Research Institute, leads the industry by analyzing the trend of mobile communication industry to establish the strategy for technology development and standardization, e.g., the wireless broadband/high speed mobile multimedia (WiBro/HMm) project. Together, China, Japan, and Korea have formed the Chinese, Japanese, and Korean initiative to promote a common approach toward next-generation IP networks and 4G mobile communications.

The Third Generation Partnership Project (3GPP) is a collaborative agreement between a consortium of telecommunications associations, which includes the European Telecommunications Standards Institute, Telecommunications Technology Association, Association of Radio Industries and Businesses/Telecommunication Technology Committee, Alliance for Telecommunications Industry Solutions, and China Communications Standards Association, established in 1998 to standardize the development of universal terrestrial radio access (UTRA) and evolved universal terrestrial radio access (E-UTRA), e.g., the LTE/system architecture evolution (SAE) project. The refined scope of 3GPP also includes the maintenance and development of global system for mobile communications (GSM), technical specifications and technical reports for evolved RATs, e.g., the general packet radio service (GPRS) and enhanced data rates for GSM evolution (EDGE).

The Internet Engineering Task Force is a large, open international community of network designers, operators, vendors, and researchers concerned with the operation and evolution of the internet architecture. The developments in the Internet Engineering Task Force, e.g., the mobility for IP: performance, signaling, and handoff optimization (MIPSHOP) project is particularly relevant to future wireless networks as IP is envisioned to become the '*lingua franca*' of mobile and wireless communications. The Wireless World Research Forum is established as a global, open initiative of manufacturers, network operators, small and medium-size enterprises, research and development centers, and academic institutions in pursuit of a common global vision for future wireless world. The Wireless Innovation Forum, also known as the version 2.0 of SDR Forum, is devoted to SDR, cognitive radio, and dynamic spectrum access technologies to address the emerging wireless communications requirements.

In the remainder of this section, the pertinent B3G/4G research activities associated with IP convergence, cooperative networks, cognitive networks, and SDR will be reviewed.

2.1.1 Toward IP Convergence

One of the earliest large scale national research project which embarks on the development of new technologies for seamless integration of heterogeneous wireless network is the MIRAI [27], a Japanese word for future, funded under the e-Japan strategy. The main requirements for provisioning seamless services over the heterogeneous networks have been identified as follows:

- The need for RAN discovery, in assisting end-users to find the available RAN in a particular geographic location for initial access, and RAN selection which allows delivery of services by the most efficient network.
- Common core network to ensure seamless HHO within the same RAN and VHO among different RANs with guaranteed QoS, as well as a resource manager to coordinate traffic distribution across RANs.
- Multi-service user terminal, which is SDR-based, capable of operating in multiple modes such as multiple air interface standards, multiple modulation techniques, or multiple access methods in order to access different RANs.

The concept of MIRAI is built around the above requirements focusing on the developments in three core areas comprising of IP-based common core network, basic access network, and SDR-based multi-service user terminal. These three basic entities of MIRAI result in a simple architecture, and therefore low cost of implementation. In addition, the concepts of a common core network and a separate basic access network enable wireless service providers to set up an infrastructure without huge investment cost as they need to roll out only base stations and access mechanism for terminals. In other words, they need not worry about technical issues such as interconnecting, routing, handoffs, and business issues such as billing and managing customer profiles, which are already provided by the MIRAI architecture.

The keystone of the MIRAI architecture is the basic access network which supports heterogeneous, energy efficient paging to terminals in a mobile environment. The basic access network is a wireless system that provides basic access signaling between the network and end-users. It could be deployed by using the existing wireless system with basic access signaling as an overlay or a wireless system dedicated to basic access signaling from which the MIRAI has adopted the latter. The purpose of basic access signaling is to provide a set of functions specific to heterogeneous wireless networks including RAN discovery, RAN selection, VHO, location update, paging, as well as authentication, authorization, and accounting. Another important role of the basic access network is to facilitate IP mobility functions of the heterogeneous network in a seamless fashion. The mobility management model of the MIRAI is based on the implementation of two-level IP mobility using macro mobility over micro mobility. Specifically, it relies on MIP in [28] and [29] for IP mobility between different common core networks or administrative domains. On the other hand, micro mobility protocol such as cellular IP in [30] or handoff-aware wireless access internet infrastructure (HAWAII) in [31] is used within a common core network.

2.1.2 Cooperative Networks

Cooperation between networks is a rapidly emerging research area. The vision of the cooperative network ad hoc working group in the Wireless World Research Forum [32] is to enable seamless communication on mobile devices operating in heterogeneous network technologies, thus paving way for cooperative networks. The architectural principle of cooperative network is based on a layered approach with the concept that any layered models comprise of at least an application layer, a connectivity layer, and an access layer. This layered approach allows easy adaptation of heterogeneous access networks, changes in technologies, and support for fast service innovation by the virtue of maintaining a well-defined interface with independent functionality. Hence, cooperative network promotes a modular architecture in which different functional blocks could be combined to form customized or complex systems.

Recently, there have been research works in [33] and [34] that attempt to define the upcoming 4G technology from a user-centric perspective by introducing the social dimension of cooperative services. In particular, cooperation in the society is defined as a coordinated effort to reach mutual goals. An example of cooperation in heterogeneous networks between cellular and short range networks is discussed in [34]. The cooperation here refers to the capability of mobile terminals to connect to both cellular network and other mobile terminals simultaneously. The fundamental idea of such collaborative architecture is to leverage on the synergy of cellular network and short range communication such as WLAN with highly complementary characteristics. E.g., the cellular network operates in licensed band with high power, wide coverage, and low data rates while the WLAN operates in unlicensed band with low power, local coverage, and higher data rates. In particular, short range communications technique could bring about enhancements such as reduced power consumption, higher link reliability, and higher spectral efficiency by enabling direct communications between mobile terminals in proximity while maintaining connection to the cellular communication link. This hybrid network that combines centralized networks with distributed networks is referred as cellular controlled short range communications.

As previously mentioned, the ambient network [9] is one of the five Wireless World Initiative integrated projects in the 6th Framework Programme funded by the European Commission. The ambient network project aims to provide seamless inter-working between heterogeneous networks where the cornerstone of ambient network is based on the dynamic composition of networks. This could provide access to any networks through the establishment of inter-network roaming agreements on the fly. The key concept of ambient network as illustrated in Figure 2.2 lies in the sharing of common control plane called ambient control space (ACS) by a set of nodes.

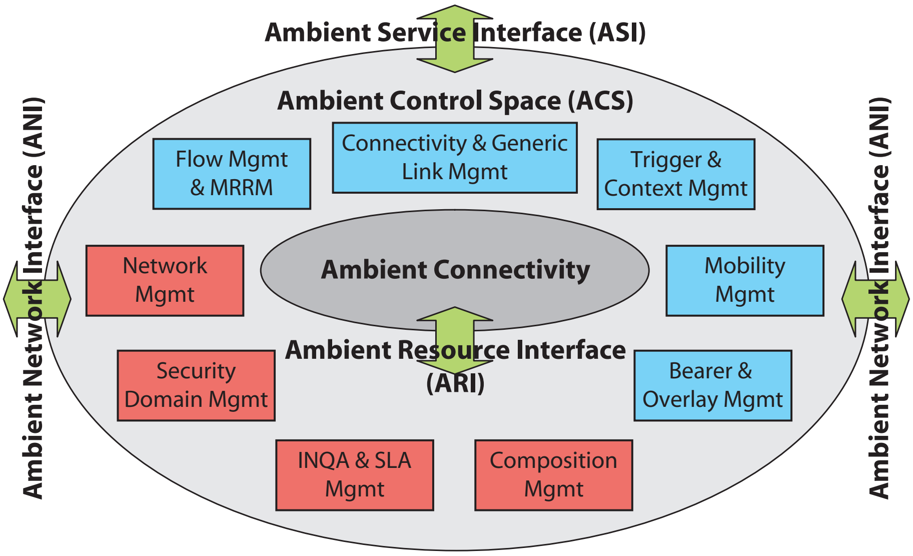

Ambient Service Interface (ASI)
Ambient Control Space (ACS)
Flow Mgmt & MRRM
Connectivity & Generic Link Mgmt
Trigger & Context Mgmt
Network Mgmt
Ambient Connectivity
Mobility Mgmt
Ambient Network Interface (ANI)
Ambient Network Interface (ANI)
Security Domain Mgmt
Ambient Resource Interface (ARI)
Bearer & Overlay Mgmt
INQA & SLA Mgmt
Composition Mgmt

Figure 2.2: Architecture and components of the ACS.

The ACS controls the underlying ambient connectivity which is responsible for the abstraction of the existing network technologies. Essentially, the ACS is composed of two discrete parts, viz., one of which supports and manages the user plane connectivity (in blue) while the other manages the administrative and security domains, as well as policies and composition (in red). It is worth noting that a minimum number of control functions must be present in any ACS before it can be considered as ‘ambient’. The services and functions of the ACS are accessible through three defined interfaces, which are also utilized to fulfill the control tasks of ACS, viz., ambient network interface (ANI), ambient service interface (ASI), and ambient resource interface (ARI). The ANI enables horizontal interactions, i.e., inter-ACS composition to cooperate with other ambient networks by connecting to their ACSs. The ASI allows vertical interactions with the application layers such that applications could issue requests to ACS pertaining to the establishment, maintenance, and termination of end-to-end connectivity while also having the capabilities to provide network context information back to the applications. The concept of bearer abstraction is provided in the ASI to hide the implementation of connectivity from application services. The ARI provides ACS with the capability to manage the resources in the ambient connectivity plane. The concept of flow abstraction is provided in the ARI to present a common view of underlying networks in terms of nodes and links. Collectively, these two levels of abstraction allow the control functions of ACS to interact with applications and resources in a generic and consistent manner, irrespective of underlying technologies, to maintain seamless inter-networking.

2.1.3 Cognitive Networks

In recent years, the context of cognition has been dilated to cognitive networks [14] which postulate the ability to optimize user and network performances, as well as to bring about better utilization of radio resources by adopting cognitive functionality, akin to cognitive radio. The concept of cognitive networks augmented by a knowledge plane that embodies cognitive tools and learning ability is first introduced by Clark *et al.* [35]. However, Thomas *et al.* [36] are the first to define cognitive networks and examine their functionality. Subsequently, Thomas *et al.* [37] propose a three-layer framework for cognitive networks comprising of software adaptable network, cognitive process, and application/user/resource layer where end-to-end goals are defined to drive the behavior of the system. This leads to cognitive networks which resemble cognitive radios. Accordingly, the cognitive process remains as the pivotal mechanism to facilitate learning from past experience for determining future response, whereas software adaptable network is analogous to SDR that offers flexibility required by the cognition process. The only exception is the introduction of end-to-end goals which represent network-wide scope. According to the authors, it is this network-wide scope that differentiates a cognitive network from cognitive radio or cognitive layer which has only a single, local scope. In fact, the recent IEEE 1900.1 standard [38] states that nodes in cognitive networks do not have to be cognitive radios. Rather, cognitive networks are seen as a composition of radio nodes subjected to cognitive functionality where the cognition could be in the radio, in the higher layers, or both.

One of the large scale collaborative integrating projects from the eMobility European Technology Platform, funded by the European Commission under the 7th Framework Programme, is the end-to-end efficiency project that aims to ambitiously integrate cognitive wireless systems into the future heterogeneous wireless system infrastructure. The end-to-end efficiency project can be seen as an extension of the end-to-end reconfigurability project to realize the vision of true end-to-end connectivity based on four top level objectives: (i) exploitation of diversity; (ii) maximization of network management efficiency; (iii) maximization of access network efficiency; and (iv) non-disruptive evolution. Out of the six defined work packages, work packages 3 and 4 are the main drivers to realize these objectives. Accordingly, work package 3 focuses on the development of collaborative cognitive RRM comprising of joint RRM, dynamic spectrum management, and self-organizing networks to optimize radio resources usage in a heterogeneous radio ecosystem. On the other hand, work package 4 concentrates on the solutions related to dynamic and opportunistic spectrum access, and autonomous functionalities in the context of cognitive radio systems.

2.1.4 Software Defined Radio

The SDR [39], also known as software based radio or simply software radio, is a promising approach toward multi-mode devices. Software based radio can be broadly defined as a radio that uses software techniques on digitized radio signals. This marks the paradigm shift from traditional radio implementation based on hardware and specific application to employing software application in performing the radio tasks. In this respect, SDR is defined as a radio where the digitization is performed at the radio back-end, i.e., baseband signal processing and represents a short-term evolution toward software radio. With technology advancements, SDR could evolve to software radio by shifting the digitization of radio signal from the baseband to the radio front-end, i.e., very near to the antenna where all the processing for the radio is carried out by the software residing in high-speed digital signal processing elements. Although the analogue to digital converter should be placed right at the antenna for an ideal software radio, this is not realizable in practice as low noise amplifier in the receive path and power amplifier in the transmit chain would be required. Nevertheless, this would be a constant driver and the eventual goal for future developments.

The SDR can be viewed from the multi-dimensional aspects by classifying it into four planes as described in [39]. The radio implementers' plane can be seen as the implementation techniques of SDR to replace existing methods of implementing transmitter and receivers. The network operator plane defines two primary roles of SDR which are flexible base stations/mobile terminals and the development of adaptive networks. Note that the creation of multiple overlay networks for each new standard in today's approach may no longer be necessary as both SDR-based base stations and mobile terminals could potentially adapt to new standards. The service providers plane offers the flexibility to perform upgrades and configurations of network and terminals through software downloads instead of costly hardware replacements. The user applications plane provides the advantages of SDR from a user perspective to increase functionality while reducing cost by obviating dedicated processing system. These multi-dimensional aspects provide insights which reveal that the full benefits of SDR are not a matter of just modifying the transmitter-receiver pair, but rather it can be attained only by also modifying the network level of the wireless communications system.

To achieve the expectations of SDR in providing seamless service across heterogeneous wireless networks, one has to consider not only the fundamental differences of air interface between different access technologies, but also the different protocol stacks implementation in each of the corresponding access networks. Hence, seamless connectivity could be realized only if reconfigurability of the protocol stacks is addressed together with the SDR-capable terminals and the ability of the network to support reconfiguration management. There are various methods to provide adaptation between different protocols. Readers are referred to Chapter 12 of [39] for a comprehensive treatment to the protocol and network aspects of the SDR.

2.2 Research Methodology

The underlying goal of this thesis is to develop a methodology for establishing the relationships between cooperation and cognition from a user-centric perspective in the design of the generalized CCRRM architecture for future wireless networks. The generalized CCRRM architecture is conceived by integrating four domains of cooperation into a four-stage *cognition incorporating cooperation (COGNITION) process* to create awareness of the surroundings whilst harmonizing four circles of relationship to encompass the factor of *user-centric design*. In a way, the generalized CCRRM architecture amalgamates the benefits of IP convergence and the core advances in cooperative and cognitive networks into a unifying generalized architecture to deliver end-to-end goals by augmenting on IP connectivity, multi-domain cooperation, and Bayesian learning process.

2.2.1 COGNITION Process

Cognition cycle plays a crucial role in the development of any cognitive techniques. It is worth mentioning that the behavioral model for cognitive radio in the form of cognition cycle proposed by Mitola [7] and the cognitive cycle presented by Haykin [8] are key examples of a communication system where feedback is pivotal to support learning from past interactions and maintain harmonious relationships between different network entities or functional blocks to achieve end goals. Figure 2.3 illustrates the four-stage COGNITION process, which draws on these behavioral models of the cognitive radio with an *intrinsic feedback* mechanism, to unify cooperation and cognition in the generalized CCRRM architecture. Note that the experience repository is not explicitly considered as a stage. Rather, it is portrayed here as an intrinsic feedback since it is reasonable to assume that the COGNITION process with machine learning capability such as Bayesian learning would inherently adapt its decision based on past experiences. Also note that the descriptions in parentheses are indications of potential tasks that could be associated with a particular stage but not considered in this thesis.

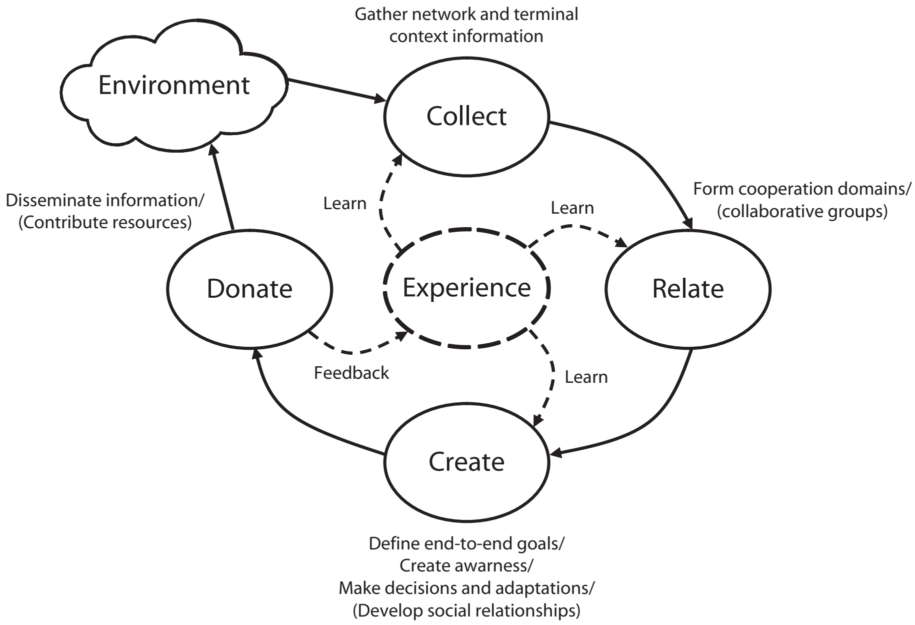

Gather network and terminal context information
Environment
Collect
Disseminate information/ (Contribute resources)
Learn
Learn
Form cooperation domains/ (collaborative groups)
Donate
Experience
Relate
Feedback
Learn
Create
Define end-to-end goals/
Create awarness/
Make decisions and adaptations/
(Develop social relationships)

Figure 2.3: Four-stage COGNITION process.

2.2.2 User-Centric Design

To some extent, the importance of user-centric design is sparkled by attempts to define the upcoming 4G technology where end-users are considered as a cornerstone in the design. Particularly, the human perspective working group (WG1) within the Wireless World Research Forum envisions that the future service architecture would be I-centric [32]. Furthermore, the key technological vision from the Wireless World Research Forum is that 7 trillion wireless devices will be serving 7 billion people by 2017. This implies that every single terminal will be surrounded by a myriad of wireless devices, and hence the potential to cooperate will be apparent. Such a paradigm shift introduces the social dimension of cooperation in future wireless networks. By understanding the end-user's needs and fostering collaborative relationships, it is more likely to accelerate evolutionary development of useful technology that will harness more potential benefits to the greater good of the society. These provide compelling reasons to adopt the user-centric philosophy in the generalized CCRRM architecture by introducing the four circles of relationship as

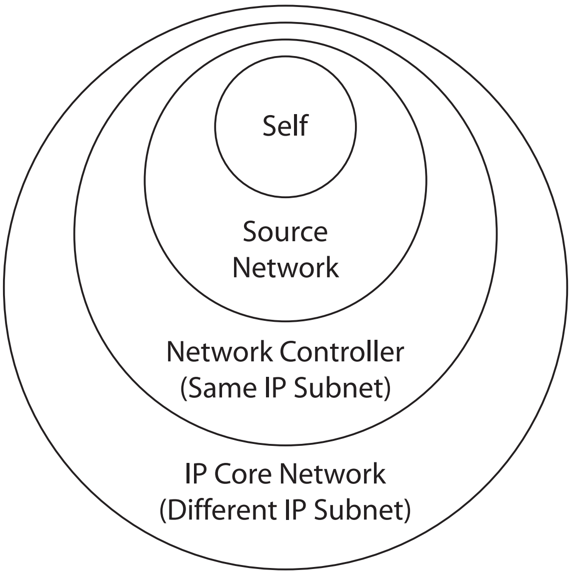

Self
Source Network
Network Controller (Same IP Subnet)
IP Core Network (Different IP Subnet)

Figure 2.4: Four circles of relationship.

depicted in Figure 2.4. These relationship circles are characterized by different sizes to represent different degrees of inter-dependence, shared knowledge, and trust within the generalized CCRRM architecture.

2.2.3 COGNITION Relationships

To this end, it is clear that the user-centric design can now be incorporated by integrating the four circles of relationship with the four-stage COGNITION process to form a two-dimensional grid known as *CoRe* which is shown in Table 2.1. The CoRe, inspired by the work of Shneiderman [40], provides guidelines for innovative solutions toward the development of the generalized CCRRM architecture. To be more specific, it identifies the activities that could be accomplished with the members within each circle of relationship, residing in one of the stages of the COGNITION process. The rows of the table cover activities such as gathering information, communicating with other networks, creating awareness, and sharing information with others through feedback mechanisms. On the other hand, the columns of the table reflect the different ranges of relationship from the peers to the source networks, to the network controllers, and to the broader communities of IP core network. Essentially, this is helpful for stimulating innovative ways to promote the collaborative use of existing information through cooperation and creating awareness of the environment through cognition. As a result, the COGNITION process is a function of cooperative tasks where the mappings of different cooperative tasks to different stages of the COGNITION process are established. In a nutshell, the CoRe constitutes the research methodology to form a concrete framework for lateral and in-depth investigations into the mechanics of cognition within a cooperative environment.

In what follows, the CoRe methodology in Table 2.1 is exemplified through the development of the generalized CCRRM architecture. A detailed treatment of the COGNITION process, which is the heart of the generalized CCRRM architecture, and the key concepts of harmonizing cooperation and cognition in a user-centric manner will be described in the respective stages of the COGNITION process.

Collect Stage

The first stage of the COGNITION process deals with the gathering of information from the environment. In broad sense, environment refers to both prevailing network conditions and terminal context information such as user preferences or QoS requirements, which can be exploited to provide ABC services [2]. Hence, the terminal should be able to discover prospective networks for initial access or handover and acquire terminal context information. In the generalized CCRRM architecture, service discovery is implemented by using the information ‘pushing’ approach where network context information, which includes QoS context information and RRM policy, is appended to the beacon frame and transmitted periodically by the source network. The basic idea is that the source network broadcasts QoS context information of neighboring networks together with its own and the RRM policy known as *network-assisted discovery*. As a result, new terminals or

Table 2.1: CoRe methodology.

CoRe	Self (Terminal)	Source Network	Network Controller (same subnet)	IP core network (different subnet)
Collect	• Terminal context information • Network context information ○ Listen to broadcast	• QoS context information ○ CLM estimations & measurements ○ Bootstrap approximation	• Gather measurement reports from other networks	• Gather measurement reports from other subnets
Relate	• Intra-layer cooperation ○ Connection-QoS entity • Inter-entity cooperation ○ Terminal-oriented decision	• Inter-entity cooperation ○ Network-assisted discovery • Intra-layer cooperation ○ Network-QoS entity	• Inter-network cooperation ○ Homogeneous networks ○ Network controllers • Inter-layer cooperation ○ PHY & MAC layer	• Inter-network cooperation ○ Homogeneous or heterogeneous networks ○ IP core network
Create	• Awareness of environment ○ Optimal target network • Adaptation to environment ○ Link adaptation ○ Handover decision	• Awareness of environment ○ Network context information • Adaptation to environment ○ Link adaptation ○ Distributed load adaptation (RRM policy)	• End-to-end goal ○ QoS-balanced system • Adaptation to environment ○ Intra-technology or QoS-based handover ○ Intra-domain mobility ○ Centralized Load adaptation (RRM policy)	• End-to-end goal ○ QoS-balanced system • Adaptation to environment ○ Intra-technology handover or inter-technology handover ○ Inter-domain mobility
Donate	• Terminal context information ○ QoS requirements ○ User preferences	• Network context information ○ QoS context information ○ RRM policy ▷ Send measurement reports to network controller ▷ Cluster-based broadcast	• Update posterior distribution ○ Bayesian learning • Network context information ○ Exchange network context information within subnet ○ Disseminate to source network	• Network context information ○ Exchange network context information across subnets

terminals with impending handover would be able to gather this information by listening to the broadcast.

Relate Stage

The second stage of the COGNITION process is a crucial step toward unifying cooperation and cognition in future wireless networks. It entails the four domains of cooperation, viz., *inter-network cooperation*, *inter-entity cooperation*, *intra-layer cooperation*, and *inter-layer cooperation*, which may occur at anytime throughout the cooperative environment. The key concept of the generalized CCRRM architecture is the synthesis of these four domains of cooperation within the COGNITION process to take advantage of the synergetic interactions which induce collaborative use of radio resources. Ultimately, this enables the joint optimization of network-terminal distributed decision making in which terminal makes RRM decisions known as *terminal-oriented decision*. This will act to optimize radio resources usage in an efficient and coordinated manner so that a high level of QoS and system capacity could be achieved.

Create Stage

The third stage of the COGNITION process is driven by the keyword *creativity*, i.e., the ability to create end-to-end goals which cover network-wide scope. The definition of end-to-end goals is an important step forward in the development of cognitive networks [37]. In a way, cognitive networks share similar trait with cross-layer design such that additional external information is shared between layers in order to perform adaptation at the layer which receives the information. However, the main difference is that cognitive networks with network-wide scope will reason about tradeoffs between multiple goals while cross-layer design with local scope tends to perform independent optimizations which may lead to suboptimal performance due to unintended consequences from adaptation loops [41]. E.g., the goal of the terminal is to maximize its perceived QoS while the goal of the source network (or network operator) is to maximize the network capacity. Clearly, this is an example of conflicting goals which will produce suboptimal results when optimized in their local scopes without end-to-end considerations.

In order to make decisions based on these end-to-end goals, two important tasks in this stage are identified as creating *awareness* of the environment and subsequently creating *adaptation* to the environment. These tasks occur through collaborative means after the formation of different cooperation domains while adhering to end-to-end goals. In essence, this stage integrates the information emanating from various cooperation domains to derive at a decision after the exploration of alternatives and reasoning about tradeoffs, and then performs adaptations according to that decision. Collectively, this would create a converged all-IP core network with a highly adaptive and self-resilient infrastructure to promote a *QoS-balanced system* which is the chief *end-to-end goal* advocated in this thesis.

Donate Stage

The last stage is concerned with the critical tasks of *disseminating* information and *feedback*, which underpin the COGNITION process. Recall from the four circles of relationship that the discussion of this context can refer to the terminal itself, its source network, its network controller, and the IP core network. The dissemination of information from the terminal perspective is related to its QoS requirements and user preferences. On the other hand, the dissemination of information from a wider scope, other than the terminal, is concerned with the exchange of network context information and dissemination from the network controller for *cluster-based* broadcast via the source network. The exchange of network context information may include QoS context information and/or RRM policy either within or between subnets. In terms of feedback mechanism, the Bayesian learning module in the network controller is a *generalized form* of the Kalman filter which fits the definition of machine learning [36] and forms the basis for cognitive behavior. In fact, Bayesian learning is employed in conjunction with cumulative sum (CUSUM) monitoring to estimate and track dynamic QoS parameters, which provide link layer cognition for various RRM functions. From the COGNITION process shown in Figure 2.3, the feedback from the donate stage is kept in the experience repository so that when similar situations are encountered in the future, it will have some idea of where to start or what to avoid. In other words, the COGNITION process could learn from the past experiences of its interactions with various network entities.

2.3 Taxonomy of the Generalized CCRRM Architecture

The generalized CCRRM architecture shown in Figure 2.5 features the cross-fertilization of cooperative and cognitive principles. Fundamentally, it is conceived based on the integration of the four domains of cooperation with the four-stage COGNITION process from a user-centric perspective, augmented by the four circles of relationship, as explained in Section 2.2.

The generalized CCRRM architecture adopts the technology agnostic approach which provides (i) *technology abstraction* for supporting access network heterogeneity; and (ii) *link layer cognition* for facilitating the joint optimization of network-terminal distributed decision making to coordinate informed VHO and dynamic load distribution. Accordingly, the TONA handover architecture and DANS algorithm are the key *enablers* of this technology agnostic approach. To be more specific, the former takes advantage of the IP-based core network to enable the cooperative exchange of QoS context information between access networks while the latter leverages on the notions of network-assisted discovery and terminal-oriented decision to enable distributed decision making process between network-terminal entities. The *cornerstone* of this technology agnostic approach is attributed to the *QoS parameters estimation* process which provides a *generic approach* to characterize the quality of wireless network and its channel, and provide link layer triggers. E.g., the imminent failure of a link is detected by observing the network qual-

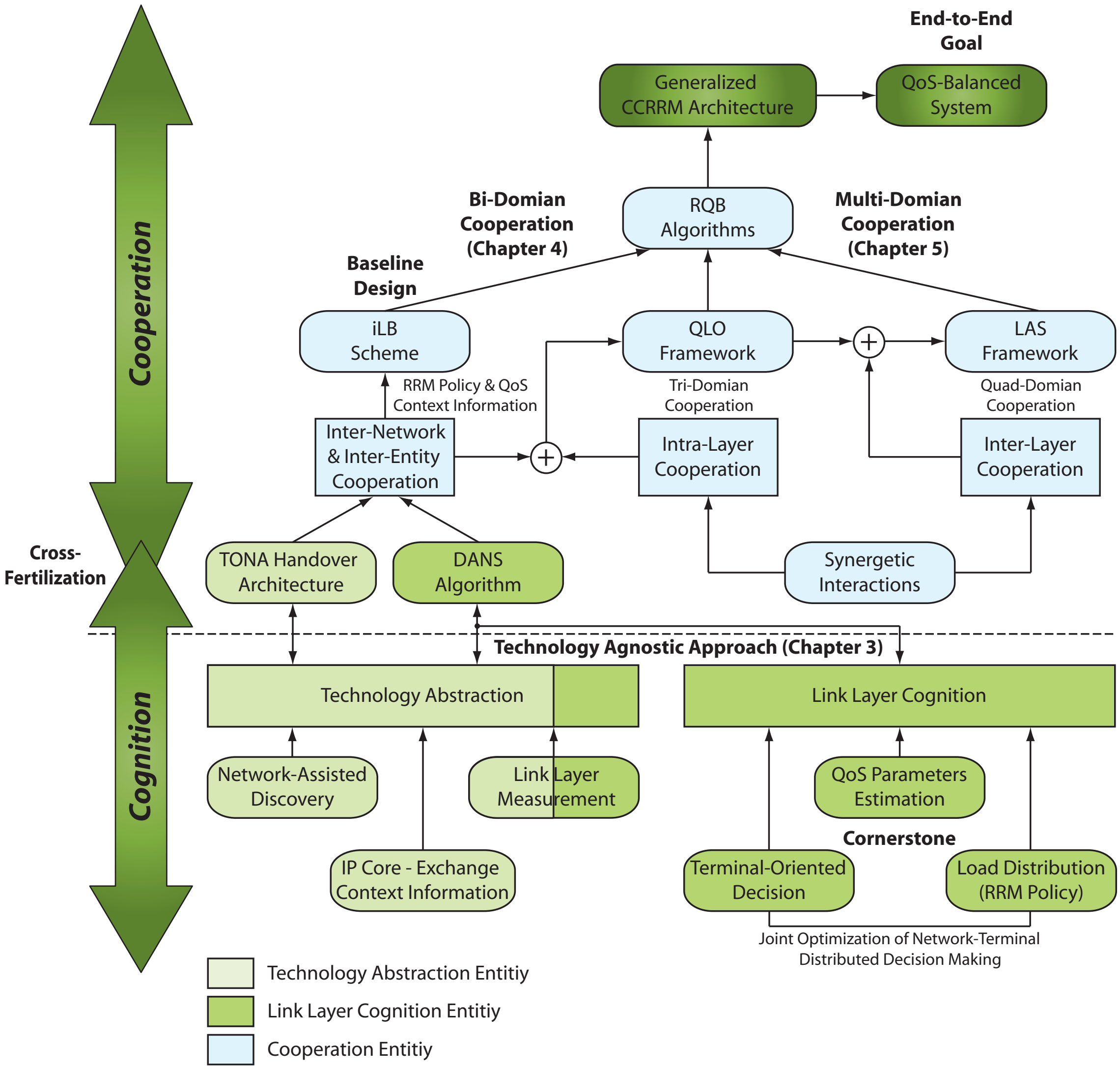

End-to-End Goal
Generalized CCRRM Architecture
QoS-Balanced System
Cooperation
Bi-Domian Cooperation (Chapter 4)
RQB Algorithms
Multi-Domian Cooperation (Chapter 5)
Baseline Design
iLB Scheme
QLO Framework
LAS Framework
RRM Policy & QoS Context Information
Tri-Domian Cooperation
Quad-Domian Cooperation
Inter-Network & Inter-Entity Cooperation
Intra-Layer Cooperation
Inter-Layer Cooperation
Cross-Fertilization
TONA Handover Architecture
DANS Algorithm
Synergetic Interactions
Technology Agnostic Approach (Chapter 3)
Technology Abstraction
Link Layer Cognition
Cognition
Network-Assisted Discovery
Link Layer Measurement
QoS Parameters Estimation
Cornerstone
IP Core - Exchange Context Information
Terminal-Oriented Decision
Load Distribution (RRM Policy)
Joint Optimization of Network-Terminal Distributed Decision Making
Technology Abstraction Entitiy
Link Layer Cognition Entitiy
Cooperation Entitiy

Figure 2.5: Taxonomy of the generalized CCRRM architecture.

ity probability of handover initiation module and VHO is triggered when a better quality AP becomes available by monitoring the outcome of network selection module. In fact, link layer measurements are readily obtainable through a *measurement process* such as the DANS algorithm, and thus it is applicable to any QoS parameters and for any given wireless networks as long as they can be measured from the system. Furthermore, QoS parameters such as PD and PLR are generally *independent* of underlying technologies and can be utilized *without* further normalization. The details of the technology agnostic approach are reported in Chapter 3.

On the other hand, the synthesis of the four domains of cooperation, viz., inter-network cooperation, inter-entity cooperation, intra-layer cooperation, and inter-layer cooperation will take the advantage of context information awareness to induce synergetic interactions between different cooperation domains in order to enable the collaborative use of radio resources. The generalized CCRRM architecture is guided by the *end-to-end goal* of maintaining a *QoS-balanced system*, which aims at adapting network load to the dynamically changing environment, through orchestrated load distribution in an *opportunistic yet altruistic* manner. In order to achieve this end-to-end goal, the generalized CCRRM architecture leverages on the novel concept of RQB. In particular, a suite of RQB algorithms based on multi-domain cooperation techniques is developed to effectuate a QoS-balanced system in which the iLB scheme based on *bi-domain cooperation* forms the *baseline design* of the generalized CCRRM architecture. The benefits of bi-domain cooperation which does not consider service prioritization and link adaptation issues can be found in Chapter 4. Building on the concept of bi-domain cooperation, Chapter 5 investigates two different flavors of multi-domain cooperation. The QLO framework based on tri-domain cooperation deals with service prioritization while the LAS framework based on quad-domain cooperation accounts for both service prioritization and link adaptation. Collectively, the studies of both Chapter 4 and Chapter 5 unveil that RQB has several intrinsic properties which are favorable for the optimization of radio resource usage in future wireless networks.

2.4 Chapter Summary

This chapter has begun by reviewing the pertinent B3G/4G research activities associated with the generalized CCRRM architecture. Alongside with the detailed exposition of the research methodology undertaken in the development and implementation of the generalized CCRRM architecture, a taxonomy of the generalized CCRRM architecture has also been presented to highlight its key components and serve as a precursor to subsequent chapters.

CHAPTER 3

QOS PARAMETERS ESTIMATION: A CORNERSTONE

In recent years, there is a multitude of existing independent RATs which offers complementary coverage, mobility, costs, data rates, and QoS. In the future, it is expected that these independent RATs would congregate in a converged all-IP core network to form a multi-RAT environment. A key challenge to enable multimedia service delivery in future wireless networks envisaged as an IP-based multi-RAT environment is the coordination of VHO between different RATs. Hence, it is important that the handover architecture is highly scalable, adaptive, and self-resilient so that multimedia services could be provisioned optimally through the most efficient access network to anyone at anywhere, anytime. In addition, the handover architecture should be efficient and flexible to support different combinations of existing, evolved, and possibly new RATs which will be present in the future. This chapter concentrates on the development and implementation of a technology abstraction and link cognition module of the generalized CCRRM architecture. The module features a technology agnostic approach to support access network heterogeneity and provide link layer cognition to orchestrate informed RRM decisions, e.g., the coordination of VHO in future wireless networks.

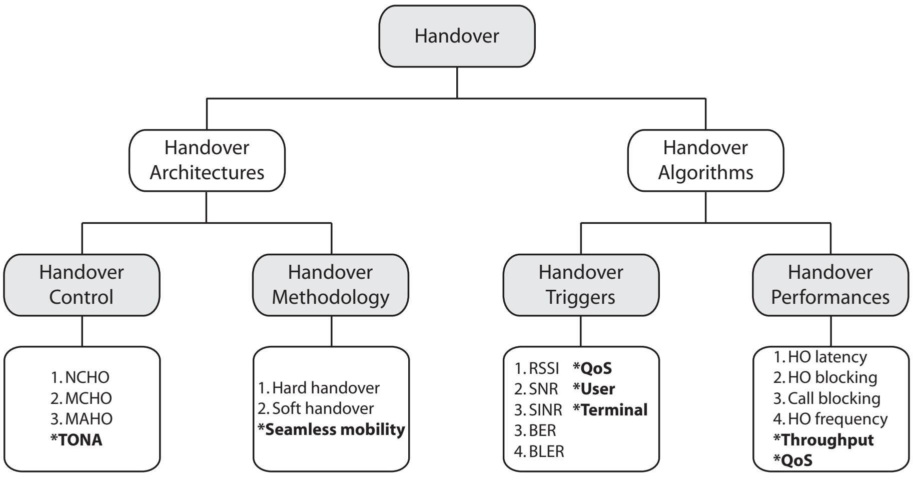
The most challenging part of the generalized CCRRM architecture is embedded in the decision making process which depends on the amount of network state information available. On the other hand, there is often a high cost to communicate this information unnecessarily to network entities which do not require it. Therefore, tradeoffs exist between the cost effectiveness of sharing network state information and the accuracy of the network state information. Hence, the challenge is that the generalized CCRRM architecture must continue to operate successfully in the presence of imprecise information, where concepts from machine learning such as Bayesian learning can be employed, to provide reliable inference from incomplete network state information. In addition, the generalized CCRRM architecture must have the capability to reason about these tradeoffs when using data from different scopes to make RRM decisions while adhering to end-to-end goals. E.g., the network which adopts QoS-balanced system as its end-to-end goal will not insist on absolute load balancing when QoS requirements are met since handovers are costly. On one hand, the generalized CCRRM architecture should remain generic for unanimous applicability to any combination of RATs. On the other hand, the generalized CCRRM architecture should deliver high performance by leveraging on sufficiently elaborated, i.e., *adequate* information to make RRM decisions. Thus, the generalized CCRRM architecture must be able to collect, filter, and channel network state information from various parts of the networks to entities where they are most useful in an efficient and not overly complex manner.

This chapter is outlined as follows. Section 3.1 presents an overview of the key aspects in handover. Section 3.2 introduces a novel distributed TONA handover architecture, as a vital part of the generalized CCRRM architecture, to support informed VHO in future wireless networks. Section 3.3 presents the motivations for QoS parameters estimation that forms a cornerstone of the generalized CCRRM architecture and advocates a technology agnostic approach to support seamless mobility in future wireless networks. Section 3.4 unveils a novel generic DANS algorithm developed to support distributed decision making process between network-terminal entities over the TONA handover architecture. Section 3.5 illustrates the proof of concept, and Section 3.6 concludes this chapter.

3.1 Overview of the Key Aspects in Handover

Traditionally, handover refers to the process of transferring an ongoing user session from one radio channel to another within the same or different cell. In recent times, the notion of seamless mobility has gained much attention due to the idea of ubiquitous, pervasive, and service-oriented future wireless networks envisioned by the ITU. Seamless mobility refers to handover which is both smooth and fast so that end-users are unaware of this process. Specifically, smooth handover minimizes packet loss whilst fast handover minimizes PD during the actual handover process. According to the recent IEEE 802.21 MIH standard [16], seamless mobility also implies the ability to assess and adapt to prevailing network conditions such that the handover decision and execution phases can be effectively optimized. In general, the key aspects of handover [42], [43] can be categorized into two broad perspectives, viz., handover architectures and handover algorithms, as depicted in Figure 3.1. Handover architectures are typically classified by the location of handover control or decision mechanism and their associated handover methodology. Both network-controlled handover (NCHO) and mobile-controlled handover (MCHO), as the names suggest, have their handover controls located in network and terminal entities, respectively. On the other hand, network entity makes handover decision based on information reported by terminal for the case of mobile-assisted handover (MAHO). Handover methodology can be generally categorized as hard and soft handover. In the former, old radio link is removed before new radio link is established, also known as break-before-make handover. This class of handover is susceptible to handover latency since terminal is essentially disconnected during handover and fast handover mechanism is required to ensure seamless handover. The latter allows terminal to maintain more than one radio links with multiple networks simultaneously, also known as make-before-break handover. This class of handover enables smooth handover since terminal engages communication with new radio link well before old link is dropped.

Handover algorithms usually concern the types of handover triggers and the resulting handover performances. HHO refers to handover from any source to target system of the same



**Key aspects for seamless mobility in future wireless networks*

Figure 3.1: Overview of the key aspects in handover.

technology, also known as radio-related handover. Handover triggers for such homogeneous systems have been primarily based on received signal strength indicator (RSSI) from the source and neighboring networks, as well as its derivatives such as signal-to-noise ratio (SNR), signal-to-interference and noise ratio (SINR), bit error rate (BER), and block error rate (BLER). On the other hand, VHO refers to handover from any source to target system of different technologies, also known as service-related handover[1]. Handover triggers for such heterogeneous system introduce new criteria since VHO could be triggered based on QoS-, user-, and terminal-related reasons even though the actual radio link quality is good. The performance of handover algorithms is typically determined by their effect on certain performance measures such as handover latency, handover blocking probability, and call blocking probability, which are related to RT services, and handover frequency which is related to ‘ping-pong’ or unnecessary handovers. Other issues such as maximization of throughput, maintaining QoS guarantee during and after handover,

---

[1] In fact, service-related handover typically arises due to QoS reasons and utilizes VHO as a vehicle to support seamless mobility and QoS transparency advocated by the ITU [1] and the recent IEEE 802.21 standard [16].

load balancing, and minimization of handover frequency are also imperative in achieving seamless mobility in future wireless networks [44].

The existing handover architectures work well for HHO, in a homogeneous network, since it comprises of mainly radio reason handovers where metrics for handover decision can be easily derived. However, a more sophisticated, scalable, and adaptive handover architecture is necessary for future wireless networks with access network heterogeneity as VHO introduces more dimensions such as QoS, load balancing, QoS balancing, user preference, terminal capability, cost, and network policy to the decision space while involving different air interfaces. Moreover, handover latency becomes important for RT services, especially, during VHO where only hard handover is supported. This suggests that maintaining service continuity and QoS transparency are crucial during VHO for the seamless delivery of multimedia service.

3.2 TONA Handover Architecture

In this section, a detailed exposition of a novel distributed TONA handover architecture as shown in Figure 3.2 to support seamless mobility in future wireless networks is presented. Although similar concepts to the TONA handover architecture are briefly introduced in [5] and the architectural aspects have been reported in [45]. These previous works do not present any formal details on the development and implementation aspects of such an architecture, in particular, they do not propose any feasible decision making algorithms. In contrast, the development and implementation of the TONA handover architecture and distributed decision making algorithm between network-terminal entities are first proposed in [46]. Strictly speaking, the TONA handover architecture can been seen as a compromise between the MCHO and NCHO in order to reap the benefits of fully decentralized and centralized handover architectures, respectively. E.g., the MCHO may result in non-optimal resource usage as in the case of WLAN, where STA makes the decision to associate with an AP based on the RSSI, when compared to the NCHO. On the contrary,

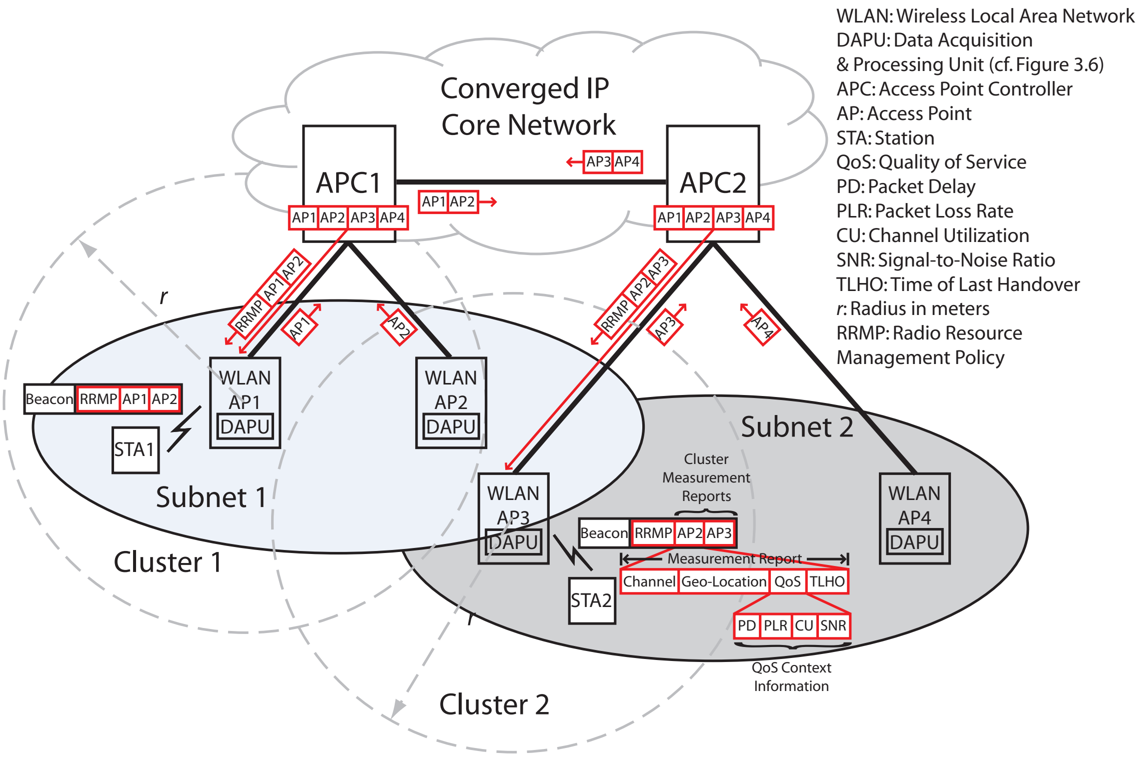


Figure 3.2: Distributed TONA handover architecture.

the NCHO may not offer an optimal RRM decision concerning each end-user as compared to the MCHO in which critical link layer measurements and QoS requirements are readily available. Furthermore, the TONA handover architecture enables inter-network cooperation and facilitates inter-entity cooperation, collectively known as bi-domain cooperation, and is one of the key enablers of the technology agnostic approach.

3.2.1 Inter-Network Cooperation

The key idea is to leverage on the converged IP-based core network to enable inter-network cooperation between access networks by facilitating the cooperative exchange of QoS context information. Note that inter-operator cooperation is not the focus of this thesis. Hence, by inter-network cooperation, it is implicitly assumed that such cooperation exists so that all relevant QoS context information required for network selection is available for exchange between the access networks of different administrative domains. Accordingly, the QoS context information[2] of an AP, together with its channel number, geo-location, and time of last handover event would be encapsulated in a packet, as a measurement report, and periodically transmitted to the access point controller (APC). The APC would collect these measurement reports from every AP in its subnet and facilitate the cooperative exchange of QoS context information between different subnets. The consolidated cluster measurement reports of the source and neighboring APs, as well as the RRM policy would then be disseminated from the APC by using *cluster-based* broadcast. The cluster is defined as a group of 'reachable' APs bounded by the cluster radius r with respect to (w.r.t.) the geo-location of the source AP. Thus, the source AP broadcasts only measurement reports of that cluster so that STA need not monitor the network conditions of distant APs.

Note that the TONA handover architecture is somewhat similar to that of the network-assisted handover (NAHO) architecture which has been briefly discussed in [47]. However, the salient difference is that the network entity is also responsible for the derivation and evaluation of RRM policy in the case of the TONA handover architecture. These are in addition to providing only network information to assist terminal in making handover decision as in the case of the NAHO. Furthermore, although the new 4G architecture proposed in [48] shares similar concept to the TONA handover architecture in terms of IP convergence, it is based on an information 'pulling' approach. Particularly, its access discovery operations are implemented in a three-tier hierarchical fashion which is susceptible to a single point of failure and increased handover latency. On the other hand, the TONA handover architecture is based on a distributed information 'pushing' approach where access discovery is straightforward without the need of additional request-acknowledgment (ACK) signaling overheads, and thus it reduces handover latency. Although the broadcasting of QoS context information will incur additional overheads, it

---

[2] The QoS context information consists only of PD and PLR for the case of bi-domain cooperation (cf. Chapter 4) while it comprises of PD, PLR, channel utilization (CU), and/or SNR for the case of multi-domain cooperation (cf. Chapter 5).

will be shown in Section 4.3.3 that these overheads are indeed insignificant, especially, when outweighed by the remarkable performance improvements. Moreover, the importance of broadcast as a means to disseminate context information is evident from the recent IEEE 802.21 and IEEE 1900.4 standards which provision such support.

3.2.2 Inter-Entity Cooperation

The notions of network-assisted discovery and terminal-oriented decision further facilitate inter-entity cooperation between network-terminal entities in the distributed decision making process. Accordingly, the TONA handover architecture supports: (i) network-assisted discovery such that the source AP broadcasts the QoS context information of neighboring APs together with its own and recommended RRM policy; and (ii) terminal-oriented decision where terminals make network selection decision according to the recommended RRM policy. Note that terminals make the *final* decision in selecting an optimal AP to fulfill their user preferences and QoS requirements while operating within the bounds of the recommended RRM policy. The concept of (i) is compatible with the IEEE 802.21 MIH services [16] where access to information about different networks within a geographical area can help in the handover decision making process. The concepts of both (i) and (ii) are similar to the IEEE 1900.4 standard [26] which aims at enabling terminals to participate in the decision making process autonomously while adhering to some policies and constraints imposed by the network. The intricate similarities of the TONA handover architecture to the IEEE 1900.4 framework will be further discussed in Section 7.2.

As a final remark, network-assisted discovery relies on beacon broadcast to ‘push’ QoS context information to terminals periodically. By listening to the broadcasts, terminals would acquire QoS context information which can then be used as inputs for RRM decisions, e.g., network selection and soft admission control. Such inter-entity cooperation between network-terminal entities provides link layer cognition to coordinate informed VHO and optimize load distribution across a multi-RAT environment in a distributed, *self-adjusting*, and opportunistic manner (cf. Chapter 4). As a result, the advantages of the TONA handover architecture by leveraging on the network-assisted discovery are as follows:

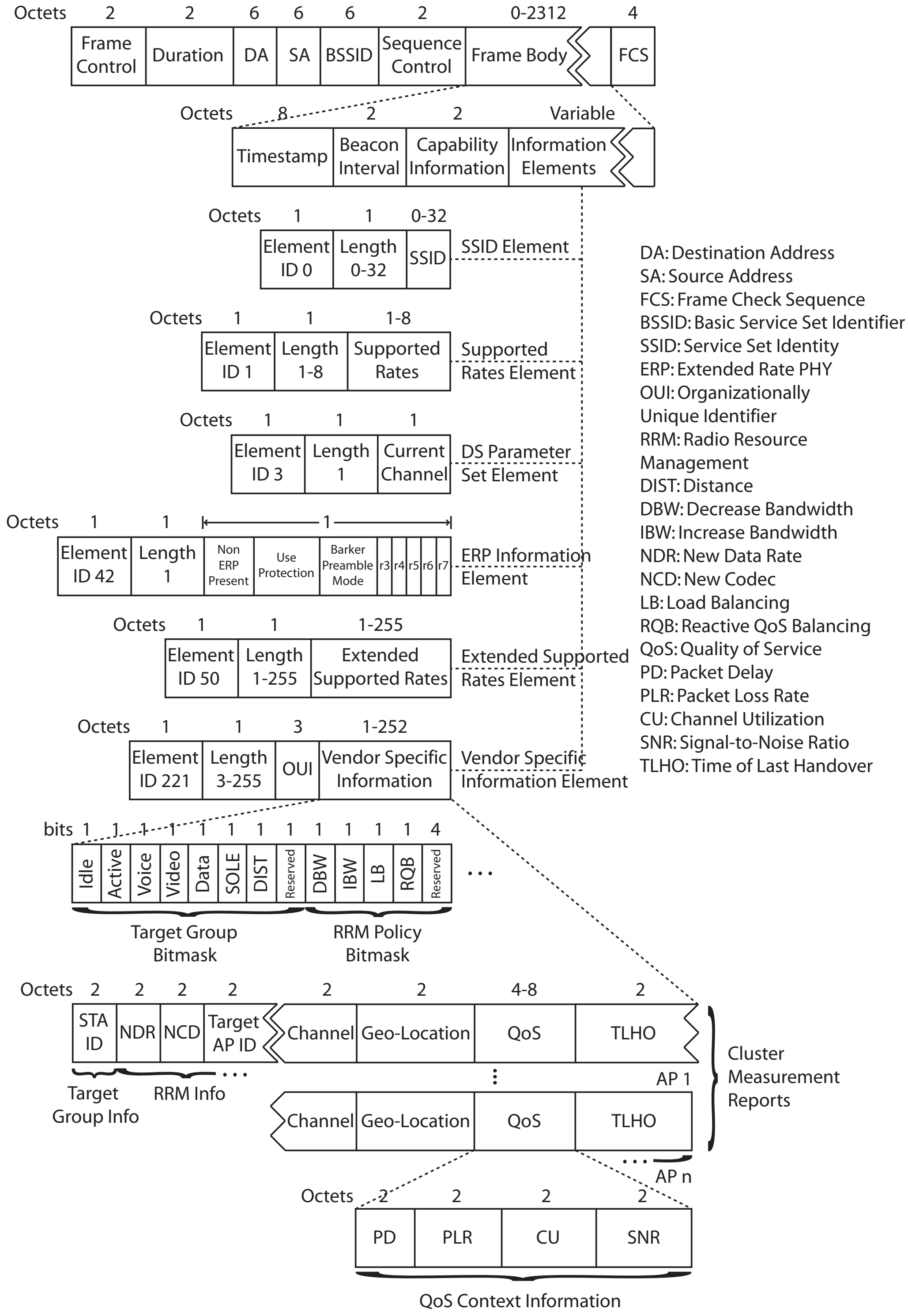
- One of the two main actors for the technology agnostic approach.
- Eliminates the need for terminal to perform any scanning operations or conduct any PHY measurements to discover neighboring access networks.
- Supports fast handover (cf. Section 4.2.1) for RT services as a result of eliminating both detection and scanning delays. This feature has significant importance during inter-technology handover or VHO where soft handover is usually not supported.
- Backward compatibility for conventional non-SDR terminal of a single transceiver.
- 'Green' terminal as power consumption associated with the scanning of available networks will be significantly reduced. Furthermore, in the case of an idle multi-mode terminal, only a single radio interface needs to be active for listening to the broadcast.

In addition, the terminal-oriented decision supports ABC services and is well-suited for decentralized operations in a dynamic multi-RAT environment. However, a caveat for deploying the terminal-oriented decision in a dynamic environment is system stability. E.g., it might happen that a group of terminals frequently switch between networks due to varying network conditions as discussed in Section 3.3.1. Therefore, it is important to consider *system stability* when designing decision making algorithms such as in the proposed technique that will be elaborated in Section 3.4.

3.2.3 QoS Broadcast with Beacon Frame

One of the key features of the TONA handover architecture is inter-entity cooperation between network-terminal entities to support the distributed decision making process. For that purpose, RRM policy and QoS context information are appended to the beacon frame for broadcast by the APs to their associated STAs, referred as QoS broadcast first proposed in [49]. The beacon frame [24] as depicted in Figure 3.3 is part of the management frame subtypes, which allows STA to locate the BSS at any time by broadcasting time and PHY parameters periodically. The frame body of a management frame carries information in both fixed fields and variable length information elements that are dependent on subtypes. The beacon frame consists of the following fixed fields, viz., timestamp, beacon interval, and capability information. The timestamp contains the value of the STA's synchronization timer at the time the frame is transmitted. The beacon interval is the period of beacon transmissions measured in time units of 1024 μs, and the capability information identifies the capability of the STA. The information elements in a beacon frame are the service set identity, supported rates, one or more PHY parameter sets, and some optional information elements such as contention-free parameter set, independent basic service set (IBSS) parameter set, traffic indication map, extended rate PHY (ERP) information, and extended supported rates. The information element is a flexible data structure that occurs in the frame body in order of increasing identifiers. It contains an information element identifier, a length, and the content of the information element.

The fixed fields and variable length information elements data structures allow for flexible extension of the management frame to include new functionality without affecting existing implementations. This is possible as existing implementations will be able to ignore elements with new identifiers. Since the length of the element is also part of the data structure, an existing implementation can disregard new elements without the need to understand its content. Hence, RRM policy and QoS context information can be appended to the beacon frame using the vendor specific information element to ensure interoperability with existing implementations. Note that a new information element could also be adopted in the same way. The vendor specific information set contains an additional mandatory field known as the organizationally unique identifier that distinguishes between different vendors. The RRM policy and QoS context information are stored in



DA: Destination Address
SA: Source Address
FCS: Frame Check Sequence
BSSID: Basic Service Set Identifier
SSID: Service Set Identity
ERP: Extended Rate PHY
OUI: Organizationally Unique Identifier
RRM: Radio Resource Management
DIST: Distance
DBW: Decrease Bandwidth
IBW: Increase Bandwidth
NDR: New Data Rate
NCD: New Codec
LB: Load Balancing
RQB: Reactive QoS Balancing
QoS: Quality of Service
PD: Packet Delay
PLR: Packet Loss Rate
CU: Channel Utilization
SNR: Signal-to-Noise Ratio
TLHO: Time of Last Handover

Figure 3.3: Beacon frame format of the management frame subtypes.

Table 3.1: Encoding of the target group bitmask.

Bit(s)	Target Group	Description	Target Group Info
0	Idle	Idle STAs	-
1	Active	Connected STAs	-
2	Voice	Voice STAs	-
3	Video	Video STAs	-
4	Data	Data STAs	-
5	Sole	Single STA	STA ID
6	Dist	STAs within or outside specific distance	-
7	Reserved	Others	-

the variable vendor specific information field which can accommodate up to 252 octets of information.

The RRM policy has a 16-bit fixed field that contains a 8-bit target group bitmask and a 8-bit RRM policy bitmask as shown in Table 3.1 and Table 3.2, respectively. It also has a corresponding variable information field, which consists of a target group information field and a RRM information field, dependent upon the target group and RRM policy bitmasks. The RRM policy recommends a set of possible actions to the associated STAs. E.g., the APC might trigger congestion control and request voice STAs to decrease their bandwidth by changing from the G.711 to G.723.1 codec in order to cope with the rate anomaly phenomenon [50], [51] known to arise from link adaptation techniques employed to combat diverse wireless channel conditions. Conversely, the APC might allow these voice STAs to increase their bandwidth consumption when spare capacity becomes available. The APC could also trigger preemptive load balancing and request STAs in idle mode to camp on another designated AP with a relatively lower traffic load. In Section 4.2.1, a RRM policy which requests only voice STAs to perform RQB through VHOs by issuing the target group bitmask of (x0000100) and RRM policy bitmask of (xxxx1000) is investigated. It is shown that voice STAs can be opportunistically redistributed to a better quality or less loaded AP, according to prevailing network conditions, to effectuate a QoS-balanced system.

Table 3.2: Encoding of the RRM policy bitmask.

Bit(s)	RRM Policy	Description	RRM Info
0	DBW	Decrease bandwidth	New codec/data rate
1	IBW	Increase bandwidth	New codec/data rate
2	LB	Load balancing	Target AP ID only
3	RQB	Reactive QoS balancing	Target AP ID & cluster measurement reports
4-7	Reserved	Others	-

3.3 Motivations for QoS Parameters Estimation

The fundamental problem of a multi-RAT environment is imputed to the coordination of VHO between different RATs. As discussed in Section 3.1, VHO is imperative to support seamless mobility where its chief advantages are: (i) permitting QoS-based handover or 'better alternative' handover; (ii) increasing the QoS satisfaction level of end-users; (iii) enabling network operators to integrate several RATs into a multi-RAT environment which will improve their coverage and QoS; and (iv) improving trunking efficiency of networks through dynamic load distribution. In general, VHO decision relies on the selection of the 'best' access network that could meet the QoS requirements of end-users, allowing them to enjoy ubiquitous connectivity in the most efficient way, irrespective of time and place. As a matter of fact, an optimal network selection technique is the core component of the generalized CCRRM architecture. To be more specific, it is important for directing informed VHO to exploit heterogeneity within a multi-RAT environment opportunistically and optimize radio resource usage (see, e.g., Chapter 5 of [15] for similar motivation). Particularly, the information required for network selection process could consist of one or more parameters, commonly known as performance metric(s), which can be classified into two broad categories:

- Static QoS parameters in which the values of these parameters vary on a time scale longer than a session lifetime and are usually *independent* on prevailing network conditions. E.g., QoS requirements of end-users, network capacity, security, network policy, user preference, cost of service, battery lifetime, and terminal capability.

- Dynamic QoS parameters in which the values of these parameters vary on shorter time scale, in order of seconds or milliseconds, and are *dependent* on current network conditions. E.g., QoS metrics (PD, PLR, and jitter), radio link measurements (RSSI and SNR), network load, terminal location, and terminal velocity.

3.3.1 Limitations of Existing Vertical Handover Process

Traditionally, the VHO process [43], [52] as illustrated in Figure 3.4 comprises of three main phases:

- Handover information discovery in which the terminal collects information about the reachable neighboring networks. Such information may include geo-location and various context information, e.g., QoS, channel number, available bandwidth, and cost associated to a specific network.
- Handover decision in which the terminal performs network selection based on quantitative network conditions, from the information it gathered during discovery, and qualitative user preferences or QoS requirements. The decision is typically formulated as a MADM problem to rank the available candidate networks.
- Handover execution in which the terminal's connection is transferred from the source network to the target network. Such relocation may involve only Layer 2 handover when the connection is moved within the same IP subnet or Layer 2/3 handover when the connection is moved across different IP subnets.

Early research works relating to VHO have considered network selection as a cost or utility function optimization problem where the cost function[3] is derived from a single

---

[3] The terms cost function and utility function are used synonymously throughout this chapter to refer to an objective function in which one seeks to maximize or minimize depending on the problem statement. Specifically, cost function is derived from classical optimization problem in which it is used to choose the best element among a set of alternatives while utility function is derived from classical utility theory in economics. Given that utility is a measure of user satisfaction level from perceived QoS, the value of a cost function can be seen as a utility value. Consequently, ordinal utility function can also be used to quantify preferences among a set of alternatives.

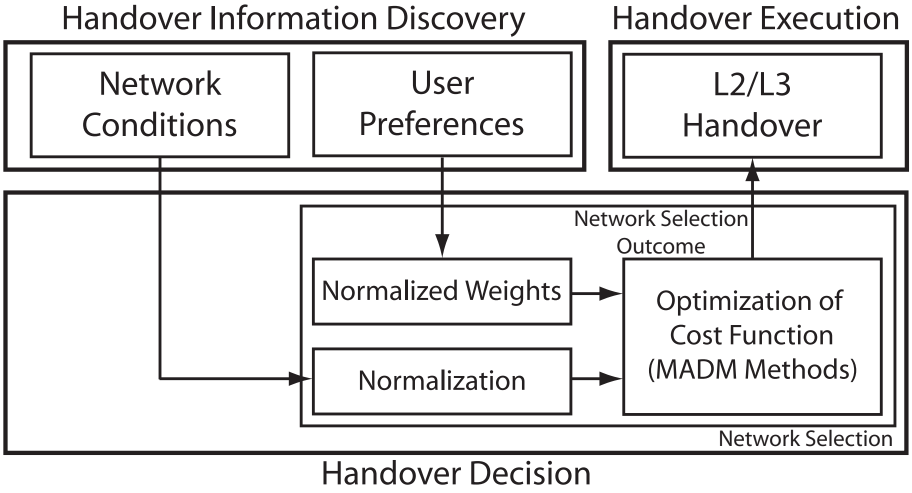


Figure 3.4: Existing VHO process.

performance metric or a weighted combination of various performance metrics. Wang *et al.* [5] propose a policy-based handover system that utilizes a cost function to reflect network parameters such as bandwidth, cost, and power consumption, thereby evaluating the need for handover. McNair and Zhu [43] give an improvement over the cost function proposed in [5] by including network elimination process in their cost function to reduce processing delay while also extending it to cater for different user services. Chen *et al.* [52] notice that network conditions will vary with end-user movements. Subsequently, they introduce a utility function to quantify the perceived QoS of the network w.r.t. the end-user's satisfaction level. This is one of the first adaptive schemes which ensures that VHO is performed only when target network is consistently better than current network by accounting for varying network conditions in the utility ratio which will adjust the stability period accordingly. Lately, Nguyen-Vuong *et al.* [53] investigate different types of commonly used utility functions and conclude that only the sigmoidal form can satisfy all the necessary conditions of a utility function in the context of network selection. However, the tuning of the utility function, i.e., its center and gradient to meet different user preferences, user satisfaction, and network characteristics is a challenging problem.

Recent works relating to ABC services can be viewed as a linear extension of VHO. Specifically, the computation of normalized importance weights in the cost function is refined where the normalized importance weights represent user preferences or application requirements. In addition, dynamic QoS parameters such as PD and PLR are used to characterize the 'goodness' of multimedia traffic under prevailing network conditions so that candidate networks can be properly selected according to the most desirable trade-offs between user preferences and network conditions. Song and Jamalipour [54] propose an access network selection strategy using analytic hierarchy process to evaluate user preferences and grey relational analysis to account for network performances based on an extensive set of QoS parameters. However, such an extensive set of QoS parameters results in duplications of similar information in different forms. Moreover, these QoS parameters are gathered using cross-layer signaling which tends to manifest in the complexity of the protocol stack and non-deterministic processing latency. Furthermore, their access network selection is triggered only if the service class or AP changes. However, this will undermine the varying network conditions if otherwise.

Steven-Navarro and Wong [55] present a comparison of VHO algorithms based on four MADM methods, viz., simple additive weighting (SAW), technique for order preference by similarity to ideal solution (TOPSIS), grey relational analysis, and multiplicative exponent weighting, which exhibit similar results for RT traffic, to provide ranking of candidate networks. Zhang [56] treats VHO decision as a fuzzy MADM problem where fuzzy logic is employed to represent the imprecise information of the network. The proposed algorithm first converts fuzzy data into crisp numbers and subsequently utilizes classical MADM techniques of SAW and TOPSIS to rank candidate networks. Kassar *et al.* [47] also employ such fuzzy MADM framework but rely on analytic hierarchy process to make VHO decisions. Bari and Leung [57] propose an automated network selection scheme by leveraging on both non-compensatory and compensatory MADM algorithms. The basic idea is to use the non-compensatory MADM algorithm for filtering the available candidate networks based on a list of non-QoS-related criteria, followed by ranking the list of available candidate networks based on QoS-related parameters using the compensatory MADM algorithm of TOPSIS. However, the dynamic QoS parameters required for cost function formulation in the above works are assumed to be available during handover information discovery. The issue on how to acquire and process them in an efficient manner is not addressed. As a matter of fact, several authors in the works of [44], [48], [56], [58], and [59] acknowledge that obtaining these dynamic QoS parameters is a non-trivial task. Moreover, some recent works reported in [60] and [61] assume that static and dynamic QoS parameters can be obtained from the media independent information server by leveraging on the recent IEEE 802.21 MIH services framework [16]. However, a recent study in [62], which involves a real IEEE 802.21 MIH implementation, has shown that there is still a lack of robust acquisition and processing techniques to produce reliable predictive event triggers for dynamic QoS parameters.

On the other hand, some works such as [44], [53], and [56] consider only static QoS parameters. Other works in [59], [63], and [64] consider predominantly static QoS parameters whilst including RSSI, SNR, or SINR as the criterion to make network selection decision in heterogeneous networks. However, it is known that RSSI is insufficient to support VHO due to the asymmetric nature of heterogeneous networks. Furthermore, both SNR and SINR are effective only in reflecting the wireless channel quality rather than network conditions which may be a consequence of both wireless channel quality and network congestions. Additionally, most of the pertinent works reported in [47], [48], [52], [54], [55], [57], [59], [60], [61], [63], and [64] which consider dynamic QoS parameters, however, do not address how the non-stationary characteristics of these dynamic QoS parameters will impact on the handover decision resulting from their network selection mechanisms. In particular, it is important to note that although cost function approach suffices when evaluating static QoS parameters (which vary on a time scale longer than a session lifetime), it results in frequent, and often, unnecessary handovers when considering dynamic QoS parameters (which vary on a short time scale) as the handover decision is based on instantaneous cost function value. Consequently, handover is triggered *im-*

mediately when cost function value of the target network is evaluated to be better than the source network, *regardless* of whether source network quality is *sufficient* to meet the QoS requirements of end-users. It is argued in this thesis that triggering a handover whilst the source network quality is still sufficient is unnecessary and non-optimal.

Collectively, the two key limitations of existing vertical handover process can be summarized as follows:

- An efficient way to obtain and process dynamic QoS parameters, which could arise from variability of traffic load, particularly, for multimedia services and variability of wireless channel conditions, required for VHO decision is still not adequately addressed in literature. Consequently, most of the performance metrics used in the prior works are: (i) static such that they vary over a long time scale; (ii) infrequently updated, e.g., only when service class or AP changes, which undermines the dynamic behavior of these QoS parameters; or (iii) assumed to be available during network or handover information discovery phase.
- To this end, it is clear that VHO decisions derived from cost, utility, or MADM methods are based on some form of cost function which will result in frequent, and often, unnecessary handovers when considering dynamic QoS parameters. Such unnecessary handovers will manifest in gratuitous signaling load and handover latency, which inherit negative impacts on system stability and end-user experience.

Therefore, it is imperative to provide a pragmatic and efficient method to obtain and process these dynamic QoS parameters while improving on the frequent unnecessary handovers shortcoming of the existing cost function approach in order to take a step closer to ABC services and realize seamless mobility in future wireless networks.

To the best of the author's knowledge, there is no prior work which: (i) provides a pragmatic approach to acquire dynamic QoS parameters as they are often assumed to be known or available during handover information discovery phase; and (ii) addresses the impact of non-stationary dynamic QoS parameters on the handover decision of the widely adopted cost function approach. The chief contributions of this chapter differ from the related works in seven significant ways: (i) the novel distributed TONA handover architecture provides network-assisted discovery which is compatible with the IEEE 802.21 MIH infrastructure; (ii) the terminal-oriented decision together with the network-assisted discovery mechanisms augment the novel generic DANS algorithm to support distributed decision making process between network-terminal entities which is similar to the IEEE 1900.4 architecture; (iii) the DANS algorithm is a measurement-based network selection technique that provides a pragmatic way in acquiring dynamic QoS parameters such as PD and PLR; (iv) the measurement-based network selection technique also augments the handover decision of existing cost function approach, through handover initiation module, to provide an optimal network selection outcome in the presence of dynamic QoS parameters; (v) the novel technology agnostic approach is coined by amalgamating the main actors of TONA handover architecture and DANS algorithm in which QoS parameters estimation is a cornerstone; (vi) this chapter establishes the key insight which uncovers the fact that the widely adopted cost function approach inherently results in frequent, and often, unnecessary handovers, which will be detrimental to QoS and system capacity, when considering dynamic QoS parameters; and (vii) the implementation of the QoS broadcast mechanism, which ensures interoperability with existing standard, in the TONA handover architecture with detailed beacon frame format and corresponding encodings is shown.

3.3.2 Technology Agnostic Approach

The network state information used to characterize any wireless networks should be independent of the underlying technologies since future wireless networks will be highly heterogeneous. In fact, heterogeneous access networks environment presents a different set of problems pertaining to handover as compared to homogeneous access networks. The traditional method of performing handover based on PHY measurements such as RSSI or SNR works well for homogeneous access networks, but it is no longer sufficient for heterogeneous access networks. One of the main reasons is the non-existence of a common pilot among heterogeneous access networks. This prohibits the use of PHY measurements as handover trigger directly since the reference sensitivity level thresholds of different transmission technologies may not be compared in a meaningful manner without suitable normalization.

In recent years, ABC services have gained much attention as a viable solution for provisioning seamless mobility in a heterogeneous access networks landscape. It advocates the use of user preferences and prevailing network conditions to choose the 'best' available network dynamically, irrespective of place and time, such that users can be connected through the most efficient network. However, the definition of '*best*' could range from user preferences to available network resources. Hence, the key factor to achieve QoS transparency support in heterogeneous access networks lies in *defining* the QoS requirements of end-users and *relating* these to the underlying QoS available within the system. The main QoS parameters describing wireless network conditions have been classified by Chalmers and Sloman [65] into three broad categories, viz., timeliness, bandwidth, and reliability, consisting of eleven QoS parameters. However, not all the information within the listed QoS parameters is necessary as inferences could be made on most of them from a few critical ones. E.g., the effect of BER could be inferred from PLR, and the impact of round-trip time could be inferred from PD. In reality, time and resources could be saved by keeping the critical QoS parameters to a minimum without duplication of information in different forms. Additionally, the RRM system will inevitably become heavyweight and run into scalability issues with increasing number of networks, which is expected in future wireless networks, if a large set of QoS parameters is required. It is expected that future wireless networks would be predominantly based on multimedia traffic. Henceforth, PD and PLR are identified as the critical QoS parameters which could primarily characterize the perceived quality of multimedia applications. The fundamental challenges associated with the acquisition of such dynamic QoS parameters are low latency of data processing for real-time applications, reliability of data, and size of data to be considered pragmatic. Since future wireless networks would be predominantly based on multimedia traffic, it is also important to consider the effects of self-similar traffic [66] apparent in the converged IP-based core network. These reasons provide compelling motivations to estimate the probability distribution of dynamic QoS parameters by statistical inference as it is often unrealistic to observe the entire population.

Normal approximation has been used for network delay estimation by Gibbon [67] to implement a scheduler which manages the retrieval of distributed multimedia data. However, there exist two potential problems. First, Central Limit Theorem does not state how large should the sample size be before it converges to a normal distribution. Particularly, when dealing with real-time applications, the assumption of large samples cannot be established because of time constraints and limiting data. Second, sample median has been widely adopted as an estimator of *average* values for any density functions due to its robustness when considering self-similar traffic which typically manifests as heavy-tailed distributions. Although Central Limit Theorem holds for sample mean, it does not apply to sample median. Therefore, there is no equivalent formula $\sigma\left(F\right) = \left[\mu_2\left(F\right)/n\right]^{1/2}$ that expresses the standard error of sample mean as a simple function of the sampling distribution in the case of sample median. For these reasons, the bootstrap method developed by Efron and Tibshirani [68] is adopted for estimating the probability distributions of critical QoS parameters from the acquired data itself, without the need for unrealistic or unverifiable assumptions. Notably, bootstrap method has been widely used in a multitude of disciplines such as signal processing, biomedical engineering, environmental, and geophysical research to approximate the probability distribution of an estimator or its higher order statistics of some form. A comprehensive treatment of bootstrap applications in signal processing can be found in [69].

From the synthesis of the above requirements, the key principle of the technology agnostic approach first proposed in [49] is conceived within the generalized CCRRM architecture to provide two levels of abstraction from the underlying technologies in future wireless networks: (i) the distributed TONA handover architecture proposed in Section 3.2 and the generic DANS algorithm proposed in Section 3.4 are the key enablers to support this technology agnostic approach through inter-network cooperation and inter-entity cooperation, respectively. The former takes advantage of the IP-based core network to enable the cooperative exchange of QoS context information between access networks. The latter leverages on the notions of network-assisted discovery and terminal-oriented decision to enable distributed decision making process between network-terminal entities. Particularly, the terminal makes RRM decision based on the QoS context information broadcasted by the source network, which eliminates the need to perform any scanning operations or conduct any PHY measurements to discover neighboring access networks; and (ii) the DANS algorithm is a measurement-based network selection process to provide a pragmatic way of estimating dynamic QoS parameters, which are particularly relevant to network selection mechanisms in future wireless networks. Owing to its measurement-based approach, it is applicable to any QoS parameters and for any given wireless networks. Moreover, QoS parameters such as PD and PLR are generally independent of underlying technologies and can be utilized without further normalization, leading to design simplification. Furthermore, it is worth to mention that the influence of PHY channel characteristics can also be implicitly captured in these QoS parameters. Apparently, QoS parameters estimation is a cornerstone of the technology agnostic approach adopted in the generalized CCRRM architecture to support access network heterogeneity and provide link layer cognition for orchestrating informed RRM decisions in future wireless networks.

In summary, the key advantages of QoS parameters estimation from link layer measurements such as PD and PLR are as follows:

- One of the two main actors for the technology agnostic approach.
- Provides a generic approach to characterize the quality of wireless network and its channel, and provide link layer triggers. E.g., the imminent failure of a link is detected by observing the network quality probability of the handover initiation

module, and VHO is triggered when a better quality AP becomes available by monitoring the outcome of the network selection module (cf. Section 3.4).

- Facilitates QoS transparency support by relating the QoS requirements of end-users to underlying QoS of system, which is imperative for user-centric considerations.
- Enables quantification of perceived QoS explicitly and PHY channel characteristics implicitly.
- Bootstrap approximation is a measurement-based technique which means that it is generic to be applied to any QoS parameters and for any given wireless networks as long as they can be measured from the system.
- 'Thin' terminal design as most of the computations associated with QoS parameters estimation, comprising of bootstrap approximation and Bayesian learning, occurs on the network side. Moreover, the closed-form expression of Bayesian learning yields considerable computation efficiency in practice.

3.4 Dynamic Access Network Selection Algorithm

As discussed in Section 3.3, the acquisition and processing of dynamic QoS information[4] required for network selection are typically assumed to be available during handover information discovery and not adequately addressed in literature. Hence, the goals are to: (i) provide a pragmatic approach to estimate such dynamic QoS information; (ii) preclude unnecessary handovers; and (iii) provide a low complexity solution that is suitable for implementation in terminal. These provide the motivation for a novel measurement-based network selection technique first proposed in [46] and [70] that estimates QoS information by bootstrap approximation and filters unnecessary handovers by Bayesian learning in conjunction with CUSUM monitoring. This technique effectively augments the handover decision phase of existing cost function approach, which selects *only* the most suit-

---

[4] The terms QoS information and QoS parameters are used synonymously throughout this thesis to refer to QoS metrics which characterize the perceived quality of either a system or an application.

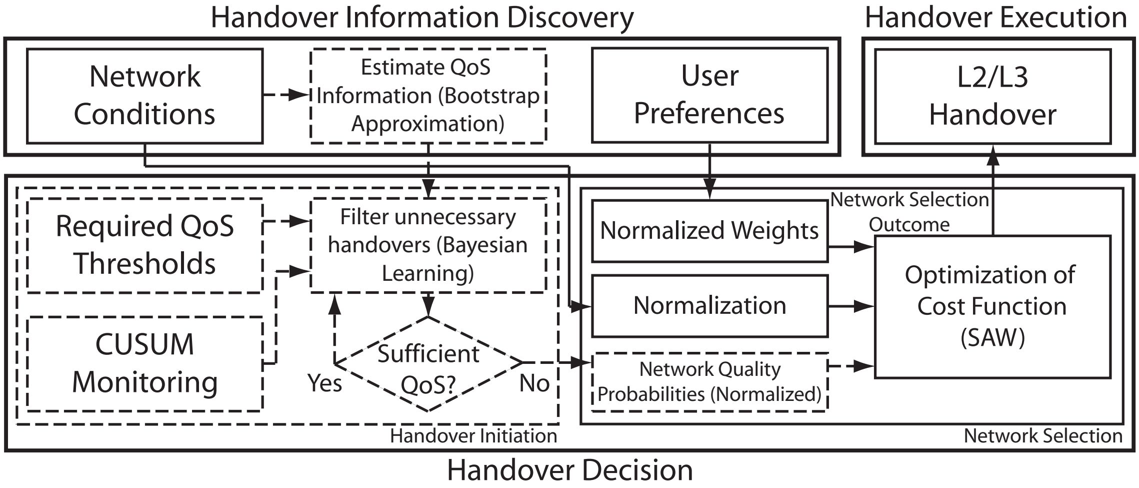

Handover Information Discovery
Handover Execution
Network Conditions
Estimate QoS Information (Bootstrap Approximation)
User Preferences
L2/L3 Handover
Required QoS Thresholds
Filter unnecessary handovers (Bayesian Learning)
Normalized Weights
Network Selection Outcome
Normalization
Optimization of Cost Function (SAW)
CUSUM Monitoring
Sufficient QoS?
Yes
No
Network Quality Probabilities (Normalized)
Handover Initiation
Network Selection
Handover Decision

Figure 3.5: Enhanced VHO process: Solid lines depict the handover decision of existing cost function approaches. Dashed lines depict the additional measurement-based network selection technique proposed to achieve optimal network selection.

able network (with the *network selection* module), by including the ability to determine whether handover is necessary (with the *handover initiation* module) for optimal network selection in the presence of dynamic QoS parameters as illustrated in Figure 3.5.

Note that fuzzy logic-based handover initiation has been proposed by the works reported in [71] and later in [47] to combat the undesirable implications of unnecessary handovers. Fuzzy logic is based on fuzzy set theory to model vagueness or *imprecision of perceptions* by using linguistic while Bayesian learning is based on probability theory to model uncertainty or *randomness of occurrence* by using numeric. Although both are fundamentally different, they appear similar owing to the fact that both membership function and cumulative distribution function (CDF) lie in the same interval $[0, 1]$. In fact, fuzzy logic and Bayesian learning are recently regarded as complementary rather than competitive [72]. However, fuzzy logic relies on a rule-based decision mechanism which may run into scalability issues, from a pragmatic viewpoint, as the number of criteria increases [56].

The concept of the DANS algorithm is a dual-stage estimation process where bootstrap approximation is performed during the first stage in an AP and Bayesian learning in conjunction with CUSUM monitoring are performed during the second stage in the APC. The estimated QoS information is subsequently broadcasted from the source AP known as network-assisted discovery while STA will listen and perform network selection known as terminal-oriented decision. Collectively, these form the inter-entity cooperation between network-terminal entities to support distributed decision making process within the TONA handover architecture. This is of paramount importance in the generalized CCRRM architecture since the outcome of network selection could be used as link layer cognition to coordinate informed VHO and optimize load distribution across a multi-RAT environment in a distributed, self-adjusting, and opportunistic manner. The estimation of the average PD from WLAN is exemplified in this section. Similar approaches can be subsequently used to estimate any QoS parameters for any given wireless networks.

3.4.1 QoS Parameters Estimation: Bootstrap Approximation

The bootstrap method is a computer-based, non-parametric approach where no assumptions are made on the underlying population from which the samples are collected. Here, the measured PD is approximated as independent and identically distributed (i.i.d.) during the data acquisition window. Although network packets traveling between a certain source and destination within the same network could not be truly independent and the statistical distribution of the network QoS information from which sample data are sought would not be truly identical due to varying network conditions, it is an assumption that approximates the actual conditions. The bootstrap notations in what follows are summarized in Table 3.3.

QoS information, in general, is non-stationary. However, it can be considered as stationary when observed over a short time. First, bootstrap approximation is engaged to estimate the *short-term* stationary components of QoS information. Suppose that an inference about the WLAN average PD denoted by unknown parameter θ from a population with

Table 3.3: Bootstrap notations.

F	Unknown distribution
$\hat{F}$	Empirical distribution
$\hat{F}_s$	Smoothed empirical distribution
$X = (x_1, x_2, ..., x_n)$	Original data set
$X^* = \left(x^*_1, x^*_2, ..., x^*_n\right)$	Bootstrap data set
$X^*_b = \left(x^*_{1_b}, x^*_{2_b}, ..., x^*_{n_b}\right)$	*b*th bootstrap data set
$Y^* = \left(y^*_1, y^*_2, ..., y^*_n\right)$	Smoothed bootstrap data set
$Y^*_b = \left(y^*_{1_b}, y^*_{2_b}, ..., y^*_{n_b}\right)$	*b*th smoothed bootstrap data set
$\hat{f}_s(x)$	Kernel density estimate of X
h_{opt}	Optimal smoothing parameter
$\tilde{x}^*$	Median of X^*
θ	Unknown parameter
$\hat{\theta}$	Plug-in estimate of θ
$\hat{\theta}^*$	Bootstrap replicate of $\hat{\theta}$
$\hat{\theta}^*_b$	*b*th bootstrap replicate of $\hat{\theta}$
$\bar{\hat{\theta}}^*$	Sample mean of B $\hat{\theta}^*$
$S\hat{E}_B(\hat{\theta})$	Standard error of $\hat{\theta}$

unknown distribution F, i.e., the WLAN measured PD is of interest, further denoted as $\theta = \theta\left(F\right)$. As discussed earlier, the WLAN measured PD could be approximated as i.i.d. during the data acquisition window, and the bootstrap method for the one-sample situation can be considered where random samples x_i are drawn from a single unknown distribution F to form the original data set $X = (x_1, x_2, ..., x_n)$. Accordingly, $F \rightarrow (x_1, x_2, ..., x_n)$ is used to represent that $X = (x_1, x_2, ..., x_n)$ is a random sample of size n drawn from F. An original data set provides a simple estimate of the entire population based on the assumption that it constitutes the underlying distribution. The *discrete* empirical distribution $\hat{F}$ is then formed by assigning a probability mass of $1/n$ on each x_i of the original data set, such that each x_i has an equal likelihood of being chosen when resampling from $\hat{F}$. The bootstrap data set $X^* = \left(x^*_1, x^*_2, ..., x^*_n\right)$ is subsequently defined to be a random sample of size n drawn *with* replacement from $\hat{F}$. Similarly, $\hat{F} \rightarrow \left(x^*_1, x^*_2, ..., x^*_n\right)$ is used to represent that $X^* = \left(x^*_1, x^*_2, ..., x^*_n\right)$ is a *resampled* random sample of size n from $\hat{F}$.

Bootstrap approximation is a direct application of the plug-in principle which is a simple method of estimating parameters from samples. The plug-in estimate of parameter θ is denoted by $\hat{\theta} = \theta(\hat{F})$ where $\hat{F}$ is used in place of F. Since the interest is to estimate parameter θ by calculating a statistic from a random sample, correspondingly, the same statistic can be calculated from a bootstrap data set X^* to obtain the bootstrap replication of $\hat{\theta}$ as

$$\hat{\theta}^*_{\mathrm{b}} = s\left(X^*_b\right), \qquad b = 1, 2, 3, \cdots, B \tag{3.1}$$

where $X^*_b = b$th bootstrap data set of B independent bootstrap data sets. Given that the statistic of interest $s\left(X\right)$ is the sample median $\tilde{x}$, then $s\left(X^*\right)$ is the median of the bootstrap data set $\tilde{x}^* = x^*_{i+1}$ from the ordered sample values $x^*_1 < x^*_2 < ... < x^*_{2i+1}$. Hence, B bootstrap replicates provide an estimate of the $\hat{\theta}$ distribution, which is the bootstrap estimate of the WLAN average PD distribution, and its standard deviation is the bootstrap estimate of standard error for $\hat{\theta}$ given in [68] by

$$S\hat{E}_B\left(\hat{\theta}\right) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_b - \bar{\hat{\theta}}^*\right)^2}, \qquad \bar{\hat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b. \tag{3.2}$$

In other words, the sample mean of B bootstrap replicates and its standard deviation in (3.2) form the parameter estimates of the WLAN average PD distribution by bootstrap approximation. The number of bootstrap replicates $50 \leq B \leq 200$ has been shown in [68] to be sufficient when estimating the standard error of a statistic.

When sampling with replacement, there is a possibility that some x_i would occur more than once or not at all. Earlier, it is assumed that $\hat{F}$ is a suitable estimate for F. However, the discrete nature of $\hat{F}$ and resampling would manifest in jagged bootstrap replicates distribution of sample median. One possible way of mitigating this problem is to construct bootstrap data sets from a smooth version of $\hat{F}$ instead, which has the effect of smoothing the discreteness of the sample median. This could be achieved with kernel density estimation [73] by taking

$$\hat{f}_s\left(x\right) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{3.3}$$

in which $K(.)$ is a Gaussian kernel with zero mean and unit variance, and h is the window width or smoothing parameter, and subsequently sample with replacement from the smoothed empirical distribution $\hat{F}_s$ rather than $\hat{F}$ such that

$$\hat{F}_s(x) = \int_{-\infty}^{x} \hat{f}_s(y)\, dy, \qquad \hat{\theta}_s = \theta\left(\hat{F}_s\right). \tag{3.4}$$

It is generally agreed that the choice of the smoothing parameter is more crucial as compared to the kernel shape. The optimal value of h as a result of minimizing the approximate mean integrated square error is given in [73] as

$$h_{opt} = \left[\frac{\int K^2(x)dx}{n\left\{\int K(x)\, x^2 dx\right\}^2 \int \left\{f''(x)\right\}^2 dx}\right]^{1/5}. \tag{3.5}$$

If the Gaussian kernel is used, then f is normal and

$$h_{opt} = 1.06\sigma n^{-1/5} \tag{3.6}$$

where σ is estimated from the data using the regular sample standard deviation, resulting in a simple data-based choice for selecting the smoothing parameter. Note that the corresponding simulation procedure could be realized without explicitly solving for $\hat{f}_s$ by taking

$$Y^* = X_{I_i^*} + h_{opt}\varepsilon_i \tag{3.7}$$

where I_i^* is sampled uniformly with replacement from $\{1, ..., n\}$, and ε_i is the random sample generated from the Gaussian kernel $K(.)$ independent of I_i^*. This is referred to as the *smooth bootstrap* which is analogous to adding a small amount of random noise $N(0, 1)$ to each bootstrap data set. Note that $\hat{F}$ can be recovered by setting $h_{opt} = 0$. Without loss of generality, $\hat{F}_s \rightarrow \left(y_1^*, y_2^*, ..., y_n^*\right)$ is used to represent that $Y^* = \left(y_1^*, y_2^*, ..., y_n^*\right)$ is a resampled random sample of size n from *smoothed* empirical distribution $\hat{F}_s$.

Unless otherwise stated, bootstrap approximation procedures implemented in the data acquisition and processing unit (DAPU) are performed with sample size $n = 20$ and number

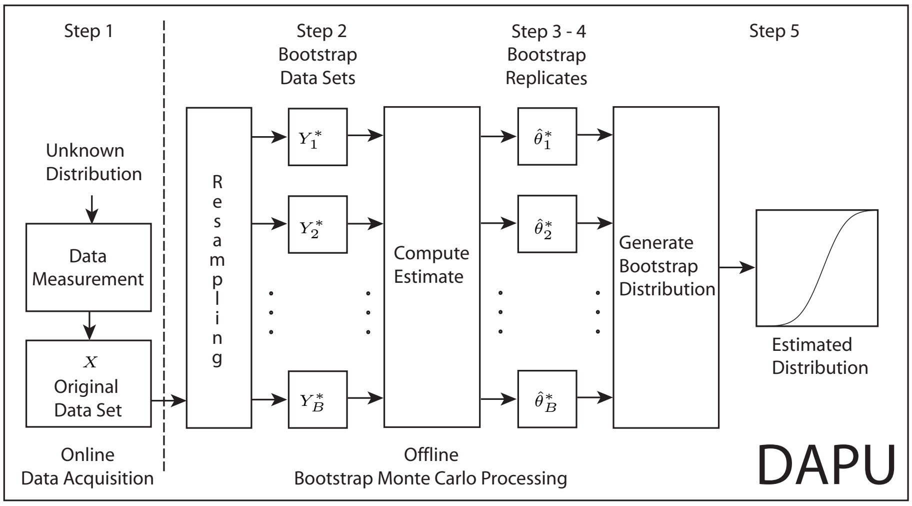


Figure 3.6: Implementation of bootstrap approximation procedures in the DAPU.

of bootstrap replications $B = 50$, requiring a processing time of 180 ms. It will be shown in Section 4.3.3 that there are tradeoffs between QoS performances and QoS broadcast intervals. Consequently, QoS broadcast interval of one second is chosen so that the network is not overwhelmed with storage, communication, and handover signaling overheads. The bootstrap approximation procedures are timed to occur before the QoS broadcast interval by a period of 1.1 times of the processing time. Since the processing time is only a fraction of the QoS broadcast interval, there will be no significant delay introduced by the DAPU. For clarity, the bootstrap approximation procedures are illustrated in Figure 3.6 and described as follows:

Step 1. Obtain the original data set $X = (x_1, x_2, ..., x_n)$ *through online data acquisition, then perform offline bootstrap Monte Carlo processing from step 2 through 5.*

Step 2. Obtain bootstrap data set $Y_b^* = \left(y_{1_b}^*, y_{2_b}^*, ..., y_{n_b}^*\right)$*, each of* n *data values from smoothed empirical distribution* $\hat{F}_s$ *by sampling with replacement.*

Step 3. Calculate the bootstrap replicates $\hat{\theta}_b^*$ *by computing the sample median for each corresponding bootstrap data set obtained in step 2.*

Step 4. Repeat step 2 through 3 B times.

Step 5. Use the distribution of B bootstrap replicates $\hat{\theta}^{*}$ *as parameter estimates to the distribution of* $\hat{\theta}$.

The distribution of bootstrap replicates, where the statistic of interest being the sample median in this thesis, can be shown to exhibit asymptotic normality as a result of employing smooth bootstrap since $\hat{F}_s$ is now continuous.

Asymptotic Normality of Sample Median. Let $F(x)$ and $f(x)$ be the CDF and probability density function of a certain population, respectively whose median is ξ. If $f(\xi) \neq 0$ and $f^{'}(\xi)$ is continuous, then the sample median $\tilde{x}$ has an asymptotically normal distribution with mean ξ and variance

$$\sigma_k^2 = \frac{1}{4\left[f\left(\xi\right)\right]^2\left(2k+1\right)} \tag{3.8}$$

where $2k+1$ is the number of sample values in a random sample.

Proof. See [74]. □

3.4.2 Optimal Network Selection: Bayesian Learning and CUSUM Monitoring

Bayesian learning[5] is a fundamental statistical approach to many difficult data modeling problems. In most real world situations, data and knowledge about the world are incomplete, indirect, and noisy. Hence, uncertainty becomes a fundamental part of the decision making process, and Bayesian learning provides a formal and intuitive way to make de-

---

[5] The term Bayesian learning is used throughout this thesis to refer to sequential Bayesian estimation so that the embedded learning process through the ‘predict-correct’ structure (cf. Section 3.4.3) of sequential Bayesian estimation is accentuated. In some cases, sequential Bayesian estimation is also referred to sequential Bayesian filtering to highlight that the current value is estimated based on past observations.

cision in the presence of these uncertainties. Next, the non-stationary components of QoS information are accounted by employing CUSUM monitoring in conjunction with Bayesian learning. Accordingly, Bayesian learning leverages on the bootstrap estimate of QoS information to build a conditional posterior distribution which is later quantified as network quality probability. The network quality probability is then used to determine: (i) *when* to handover in the handover initiation module by triggering *only* if necessary, i.e., filtering unnecessary handovers; and (ii) *where* to handover in the network selection module by computing the SAW cost function which is based on qualitative user weights and quantitative network quality of each QoS parameter. Jointly, both (i) and (ii) will provide an optimal network selection outcome. It is worth to mention that the central virtue of network quality probability lies in the fact that it does not require further normalization as in other approaches described in Section 3.3.1. This is aligned to the technology agnostic approach in the generalized CCRRM architecture which leads to further computational efficiency.

Suppose that an inference of the network quality based on the observations of the bootstrap estimate of the WLAN average PD is of interest, denoted by y_k^i and parameterized by (μ_k^i,σ_k^{2i}) of network i over time k. Bayes rule can then be applied sequentially as

$$p\left(\mu_k^i, \sigma_k^{2i} | y_k^i\right) \propto p\left(y_k^i | \mu_k^i, \sigma_k^{2i}, y_{k-1}^i\right) p\left(\mu_k^i, \sigma_k^{2i} | y_{k-1}^i\right). \tag{3.9}$$

Recall that the WLAN measured PD is acquired in non-overlapping successive windows during the bootstrap approximation. Hence, successive measurement windows are assumed as independent and the likelihood function simplifies to

$$p\left(y_k^i | \mu_k^i, \sigma_k^{2i}, y_{k-1}^i\right) = p\left(y_k^i | \mu_k^i, \sigma_k^{2i}\right). \tag{3.10}$$

The use of conjugate prior distribution is invoked as the estimation is recursively performed, and the likelihood function data estimated by the DAPU are normally distributed. The sampling variance of observation y_k^i corresponds to the squared of the bootstrap estimate of standard error and is assumed to be constant. The conditional posterior distribution of μ_k^i, given σ^{2i}, which is the Bayes estimate of the WLAN average PD, is shown in [75] as

$$p\left(\mu_k^i|\sigma^{2i}, y_k^i\right) \sim N\left(\hat{\mu}_k^i, \hat{\sigma}_k^{2i}\right) \tag{3.11}$$

where

$$\hat{\mu}_k^i = \frac{\mu_{k-1}^i/\sigma_{k-1}^{2i} + y_k^i/\sigma^{2i}}{1/\sigma_{k-1}^{2i} + 1/\sigma^{2i}} \ , \quad \hat{\sigma}_k^{2i} = \frac{1}{1/\sigma_{k-1}^{2i} + 1/\sigma^{2i}}. \tag{3.12}$$

This closed-form expression is a merit of bootstrap approximation which results in normally distributed data. By acquiring new bootstrap estimate of QoS information, it can update the belief to reflect a better knowledge of the prevailing network conditions. However, it is often desirable to reset the estimation process after the network conditions stabilize in such sequential estimation framework. The challenge in tracking non-stationary QoS information is to devise a change detection mechanism that could optimally reset the estimator such that old information can be forgotten to allow convergence to new estimates.

For the detection of state changes in non-stationary QoS information, two-sided CUSUM monitoring [76], known in the context of quality control, is implemented to detect any change of states and thereby adaptively resetting the Bayes estimator. The input of CUSUM monitoring is the magnitude of the residuals normalized w.r.t. its standard deviation as

$$s_k^i = \frac{y_k^i - \hat{\mu}_{k-1}^i}{\sqrt{E\left[\left(y_k^i - \hat{\mu}_{k-1}^i\right)^2\right]}}. \tag{3.13}$$

This normalization enables the same set of design parameters (δ, h) to be used for different scenarios. For two-sided CUSUM monitoring, a pair of auxiliary test statistics

$$\begin{cases} g_k^{i+} = \max\left(g_{k-1}^{i+} + s_k^i - \delta, 0\right) \\ g_k^{i-} = \max\left(g_{k-1}^{i-} - s_k^i - \delta, 0\right) \end{cases} \tag{3.14}$$

is necessary for the detection of state changes in non-stationary QoS information. The test statistics are initialized with a starting value of zero and will start accumulating their residuals as soon as the Bayes estimate deviates from the bootstrap estimate by more than the drift parameter δ. A positive drift would result in accumulation of residual in g_k^{i+} and a negative drift would result in accumulation of residual in g_k^{i-}. An alarm will then be triggered when either test statistic exceeds the alarm threshold h. After an alarm, the respective test statistic is cleared to zero and the Bayes estimator resets.

Let Ψ_k^i be the normal random variable where its density is the Bayes estimate of the WLAN average PD of network i over time k, i.e., $\Psi_k^i = p\left(\mu_k^i|\sigma^{2i}, y_k^i\right) \sim N(\hat{\mu}_k^i, \hat{\sigma}_k^2 i)$. The Bayes estimate of the WLAN average PD can then be quantified in terms of network quality probability by computing its CDF w.r.t. the PD threshold τ, i.e., $\Pr(\Psi_k^i \leq \tau)$ as

$$F_{\Psi_k^i}\left(\tau\right) = \Phi\left(\frac{\tau - \hat{\mu}_k^i}{\hat{\sigma}_k^i}\right) \tag{3.15}$$

where

$$\Phi\left(\tau\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\tau} \exp\left(-\frac{u^2}{2}\right) du. \tag{3.16}$$

Finally, the optimal network selection for a single QoS parameter at time k can be expressed as an optimization function conditionally by

$$N_k^{opt} = \begin{cases} \max \quad F_{\Psi_k^i}\left(\tau\right) \quad s.t. \quad \Upsilon_k^i \neq 0, \quad if\ \exists i : \Upsilon_k^i = 0 \\ N_{k-1}^{opt}, \quad otherwise \end{cases} \tag{3.17}$$

where

$$\Upsilon_k^i = \mathrm{H}\left[\Pr\left(\Psi_k^i \leq \tau\right) - \frac{1}{2}\right], \quad \mathrm{H}\left[n\right] = \begin{cases} 0,\ n < 0 \\ 1,\ n \geq 0 \end{cases}. \tag{3.18}$$

Υ_k^i is the network anomaly identifier that indicates whether the network quality is sufficient in which $\Upsilon_k^i = 1$ if network quality is sufficient and $\Upsilon_k^i = 0$ otherwise. $\mathrm{H}[n]$ is the unit step function used to implement Υ_k^i. It is important to note that handover is triggered *only* if the source network quality is insufficient, i.e., $\Pr(\Psi_k^i \leq \tau) < \frac{1}{2} \Rightarrow \Upsilon_k^i = 0$.

The network quality probabilities for other QoS parameters, in cases where multiple QoS parameters are utilized, can be evaluated in the same way as discussed in Section 3.4.4.

3.4.3 Bayesian Learning: A Generalized Form of the Kalman Filter

Although not explicitly mentioned, one should note that Bayesian learning, i.e., sequential Bayesian estimation is a generalization of the Kalman filter, which approaches the filtering of unnecessary handovers from a Bayesian viewpoint. It is widely known that the Kalman filter or equivalently Bayesian learning is an optimal linear estimator, particularly, when the bootstrap estimates of the WLAN average PD, i.e., measurement data are essentially normal random variables in this problem. It is optimal in the sense that it combines all available measurement data and prior knowledge about the system to produce an estimate of the desired variables such that the error is statistically minimized as illustrated in the following example.

Generalized Form of the Kalman Filter. Given that the goal is to estimate the average PD of a particular network i, e.g., WLAN from online measurement data at time k, the system state that is of a single dimension can be modeled as

$$x_k^i = x_{k-1}^i + \alpha_{k-1}^i w_{k-1}^i \tag{3.19}$$

where the measurement model is

$$y_k^i = x_k^i + v_k^i. \tag{3.20}$$

The random variables w_k^i and v_k^i represent the system and measurement noise, respectively. The Kalman filter is known to be an optimal linear estimator under the assumption that w_k^i and v_k^i are independent additive white Gaussian noise (AWGN) given by

$$\begin{cases} p\left(w_k^i\right) \sim N\left(0, Q_k^i\right) \\ p\left(v_k^i\right) \sim N\left(0, R_k^i\right) \end{cases} \tag{3.21}$$

where Q_k^i and R_k^i are the state and measurement noise variances, respectively, which are assumed to be constant. Note that the variances Q^i and R^i will influence the Kalman filter behavior. The state noise variance Q^i is typically used to tune the Kalman filter to prevent the 'filter dropping off' problem [77] after the Kalman filter converges to an estimate. Specifically, the Kalman gain K_k^i and error variance P_k^i will have very small stationary values and the filter becomes insensitive to abrupt state changes, which can be observed from the expressions in (3.22). To overcome this problem, an additional discrete variable α_k^i is used to capture state change by using a change detection mechanism known as CUSUM monitoring in which $\alpha_k^i = 1$ if state change is detected and $\alpha_k^i = 0$ otherwise. Accordingly, the state update, Kalman gain, and error variance update expressions can be written as

$$\begin{cases} \hat{x}_k^i = \hat{x}_{k-1}^i + K_k^i \left[y_k^i - \hat{x}_{k-1}^i\right] \\ K_k^i = \frac{P_k^i}{P_{k-1}^i + R^i} \\ P_k^i = \left(1 - K_k^i\right) P_{k-1}^i + \alpha_{k-1}^i Q^i \end{cases} . \tag{3.22}$$

Without loss of generality, it is important to note that expression (3.12) can, in fact, be rewritten as

$$\begin{cases} \hat{\mu}_k^i = \mu_{k-1}^i + K_k^i \left[y_k^i - \mu_{k-1}^i\right] \\ K_k^i = \frac{\sigma_{k-1}^{2i}}{\sigma_{k-1}^{2i} + \sigma^{2i}} \\ \hat{\sigma}_k^{2i} = \left(1 - K_k^i\right) \sigma_{k-1}^{2i} \end{cases} , \tag{3.23}$$

which then reveals its generalized form of the classical Kalman filter.

Proof. See [78]. □

The Kalman filter estimates the system state by using a form of feedback control based on prediction and correction procedures. Based on all previous information, a prediction

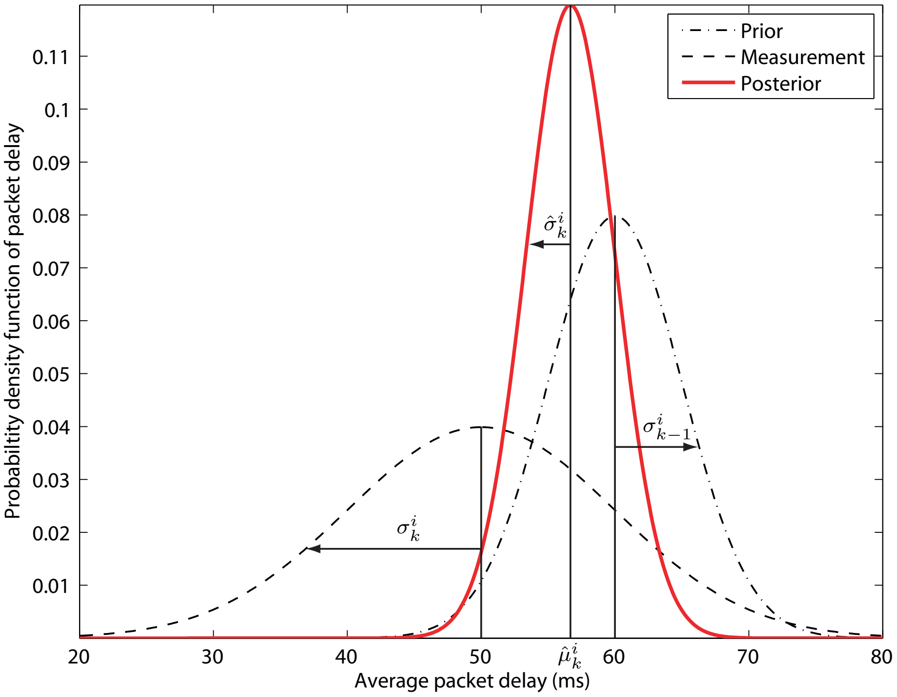


Figure 3.7: Conditional posterior density of the average PD based on the prior and measurement (likelihood) after one observation.

propagates the current state and error variance estimates to obtain an estimate of the prior for the next step. By incorporating the new measurement, i.e., feedback, the difference between the new measurement and the estimate of the prior is then used as correction to obtain an improved estimate of the posterior. Figure 3.7 illustrates that the standard deviation of the posterior $\hat{\sigma}_k^i$ is smaller than both the standard deviation of the measurement σ_k^i and prior σ_{k-1}^i, respectively. This means that the uncertainty in the estimate of the average PD has been reduced by the linear combination of information from two different sources. In other word, the posterior density with mean $\hat{\mu}_k^i$ and variance $\hat{\sigma}_k^{2i}$ provides the 'best' possible or an optimal estimate within any reasonable criterion.

Further examination of expressions in (3.12) appreciates this desirable result. The mean of the posterior is a weighted average of the prior and the measurement in which the weights are proportional to their precisions, i.e., inverse of the variances. On the other hand, the variance of the posterior is the inverse of the sum of their precisions, which again reflects the linear combination of information from two different sources. Suppose that the variance of the measurement σ_k^{2i} is larger than the variance of the prior σ_{k-1}^{2i} as in this example, the mean of the posterior would then be weighted toward the prior as the uncertainty involved in the measurement is larger than the prior. Finally, it is worth to mention that any measurement data, despite its precision, would serve to provide some information which eventually enhances the precision of the posterior as compared to the prior.

3.4.4 Realizing ABC Services with Multiple QoS Parameters

As a final step to truly realize ABC services, it is necessary to consider qualitative user preferences together with quantitative network conditions, i.e., network quality probability derived from (3.15), which would be briefly discussed. As illustrated in Figure 3.8, ABC decision is made based on the most favorable tradeoff between user preferences and prevailing network conditions. The simplest way to gather user preferences is manually going through a user-friendly graphic user interface. However, the major pitfall of this approach is the unwillingness of user to work through such tedious process [79]. One way to mitigate this problem is by mapping a set of user preferences to *stereotypes* or a cluster of characteristics such as traffic classes using a lookup table. The user may then select an appropriate traffic class and inherit the corresponding set of user preferences. Although stereotypes cannot represent all possible scenarios and may not always provide a good fit to the user's preference, it would provide a good baseline for obtaining user preferences. Given the user preferences, it is desirable to prioritize or translate this information into a set of weights describing the QoS requirements from the user perspective. For this purpose, multi-criteria decision making[6] is used, in general, to solve complex and conflicting

---

[6] The term multi-criteria decision making is often used to indicate MADM and sometimes multi-objective decision making. Particularly, MADM is a subset of multi-criteria decision making which usually evaluates a limited number of alternatives while multi-objective decision making involves a larger number of choices.

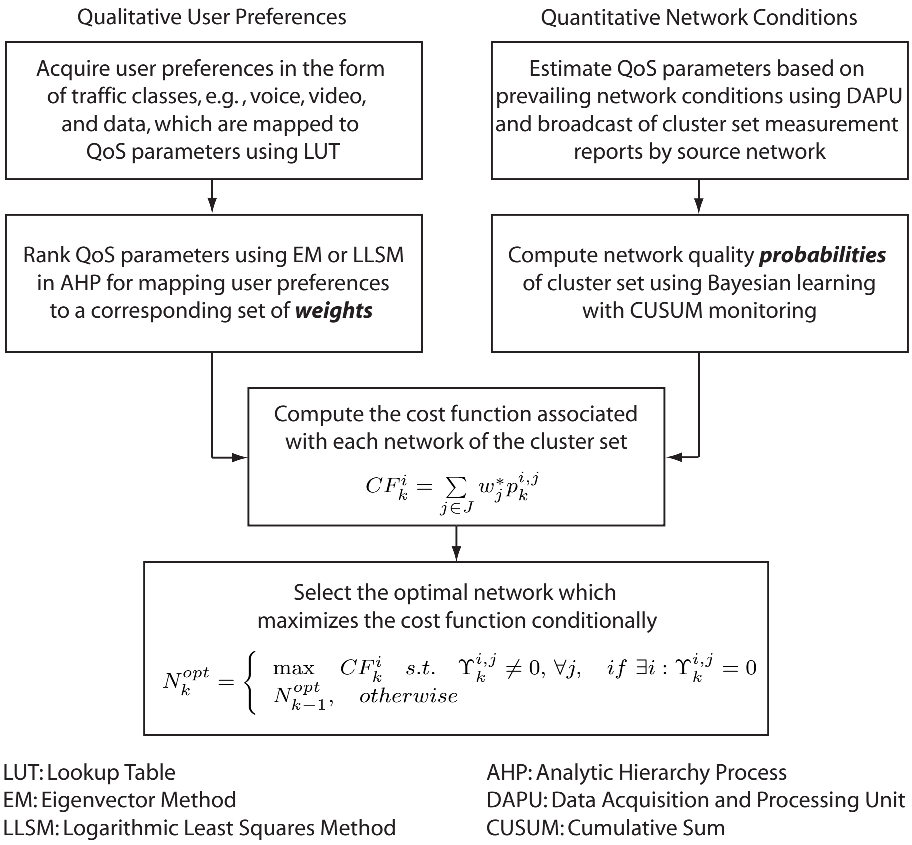

Qualitative User Preferences
Quantitative Network Conditions
Acquire user preferences in the form of traffic classes, e.g. , voice, video, and data, which are mapped to QoS parameters using LUT
Estimate QoS parameters based on prevailing network conditions using DAPU and broadcast of cluster set measurement reports by source network
Rank QoS parameters using EM or LLSM in AHP for mapping user preferences to a corresponding set of ***weights***
Compute network quality ***probabilities*** of cluster set using Bayesian learning with CUSUM monitoring
Compute the cost function associated with each network of the cluster set
$CF_k^i = \sum_{j \in J} w_j^* p_k^{i,j}$
Select the optimal network which maximizes the cost function conditionally
$N_k^{opt} = \begin{cases} \max \quad CF_k^i \quad s.t. \quad \Upsilon_k^{i,j} \neq 0, \forall j, \quad if\ \exists i : \Upsilon_k^{i,j} = 0 \\ N_{k-1}^{opt}, \quad otherwise \end{cases}$
LUT: Lookup Table
EM: Eigenvector Method
LLSM: Logarithmic Least Squares Method
AHP: Analytic Hierarchy Process
DAPU: Data Acquisition and Processing Unit
CUSUM: Cumulative Sum

Figure 3.8: ABC decision concept.

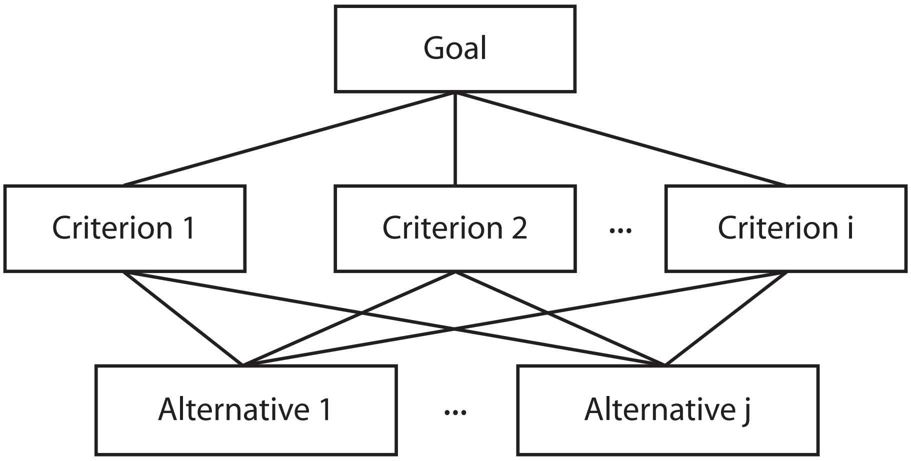


Figure 3.9: Analytic hierarchy process framework.

decision problems. E.g., analytic hierarchy process [80] as illustrated in Figure 3.9 is one of such technique based on the principles of:

- Problem decomposition into a hierarchy of goal, criteria, and alternatives.
- Pairwise comparison of the relative importance of criterion w.r.t. the goal.
- Synthesis of priorities to achieve the weight of each alternative.

Consider the $n \times n$ pairwise comparison matrix A of n criteria at the same hierarchy level. The decision maker's preference of criterion i over criterion j is reflected as a_{ij} and correspondingly $a_{ji} = 1/a_{ij}$ by reciprocal property. If the decision maker has consistent preferences, then all elements $a_{ij} = w_i/w_j$ and $a_{ij} = a_{ik}a_{kj}$ for all i, j, and k. This means there exists a unique set of weights from any column of A, multiplied by a constant. However, the decision maker's preferences are usually inconsistent $a_{ij} \approx w_i/w_j$ and consequently A is also inconsistent. Since inconsistent weights are not unique, they are often derived by using popular prioritization techniques such as eigenvector method and logarithmic least squares method. Readers are referred to [81] for an excellent review on these prioritization techniques which are beyond the scope of this thesis.

With the set of weights representing user preferences and the corresponding network quality probabilities representing prevailing network conditions, the classical method of MADM known as SAW can then be employed to rank the candidate networks. The cost function of a network candidate is determined by the weighted sum of all criteria values

$$CF_k^i = \sum_{j \in J} w_j^* p_k^{i,j}, \quad \sum w_j^* = 1 \tag{3.24}$$

where CF_k^i is the SAW cost function to rank candidate network i at time k as the weighted sum of jth QoS parameter. w_j^* is the normalized user weight of jth QoS parameter. $p_k^{i,j}$ is the network quality probabilities metric for ith candidate network w.r.t. jth QoS parameter at time k, i.e., $\Pr\left(\Psi_k^{i,j} \leq \tau_j\right)$. Similarly, the optimal network selection for j QoS parameters, satisfying both user preferences and prevailing network conditions at time k, is expressed as an optimization function conditionally by

$$N_k^{opt} = \begin{cases} \max \quad CF_k^i \quad s.t. \quad \Upsilon_k^{i,j} \neq 0, \quad \forall j, \quad if\ \exists i : \Upsilon_k^{i,j} = 0 \\ N_{k-1}^{opt}, \quad otherwise \end{cases} \tag{3.25}$$

where

$$\Upsilon_k^{i,j} = \mathrm{H}\left[\Pr\left(\Psi_k^{i,j} \leq \tau_j\right) - \frac{1}{2}\right], \quad \mathrm{H}\left[n\right] = \begin{cases} 0,\, n < 0 \\ 1,\, n \geq 0 \end{cases}. \tag{3.26}$$

For completeness, $\Upsilon_k^{i,j}$ is the network anomaly identifier that indicates whether the network quality is sufficient in which $\Upsilon_k^{i,j} = 1$ if network quality is sufficient and $\Upsilon_k^{i,j} = 0$ otherwise. $\mathrm{H}[n]$ is the unit step function used to implement $\Upsilon_k^{i,j}$. Again, note that handover is triggered *only* if the source network quality is insufficient, i.e., $\Pr(\Psi_k^{i,j} \leq \tau_j) < \frac{1}{2} \Rightarrow \Upsilon_k^{i,j} = 0$.

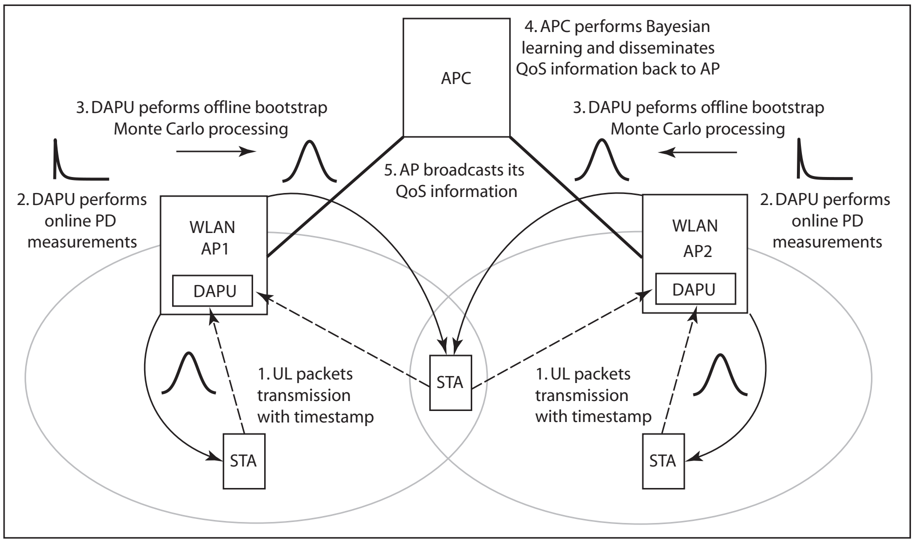

APC
4. APC performs Bayesian learning and disseminates QoS information back to AP
3. DAPU peforms offline bootstrap Monte Carlo processing
3. DAPU peforms offline bootstrap Monte Carlo processing
5. AP broadcasts its QoS information
2. DAPU performs online PD measurements
2. DAPU performs online PD measurements
WLAN AP1
WLAN AP2
DAPU
DAPU
STA
1. UL packets transmission with timestamp
1. UL packets transmission with timestamp
STA
STA

Figure 3.10: Proof of concept for the measurement-based network selection technique (DANS algorithm) based on a homogeneous multi-AP WLAN with two IEEE 802.11b APs.

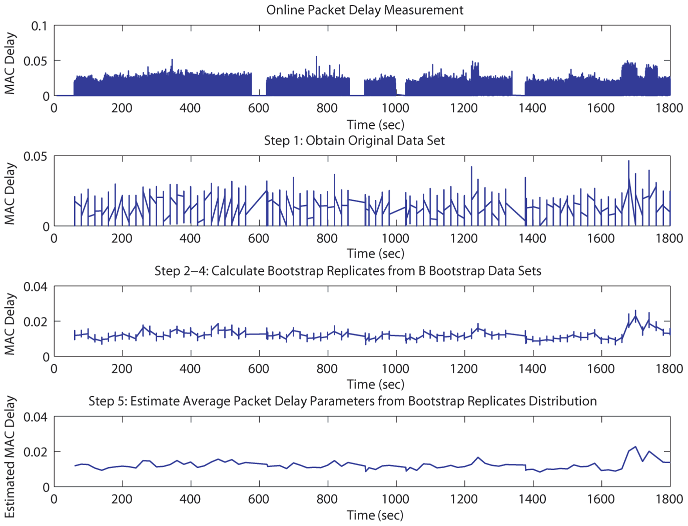
3.5 Proof of Concept

The effectiveness of the proposed measurement-based network selection technique, i.e., DANS algorithm is demonstrated using a homogeneous multi-AP WLAN with two IEEE 802.11b APs as depicted in Figure 3.10. Note that the results presented here would be irrespective of the implementation in a homogeneous or heterogeneous systems, thanks to the technology agnostic approach. The DAPU resides in each AP of the WLAN and performs *online* PD measurements. Timestamps are encapsulated in packets transmitted by STAs to enable uplink (UL) measurements. However, this is unnecessary for downlink (DL) packets since measurements are taken from the AP itself. When the DAPU acquires a sufficient sample size of 200 samples, it performs *offline* bootstrap Monte Carlo processing to estimate the probability distribution of average PD. This estimated QoS information would then be encapsulated in the beacon for broadcast in the DL where STAs could listen and perform network selection.

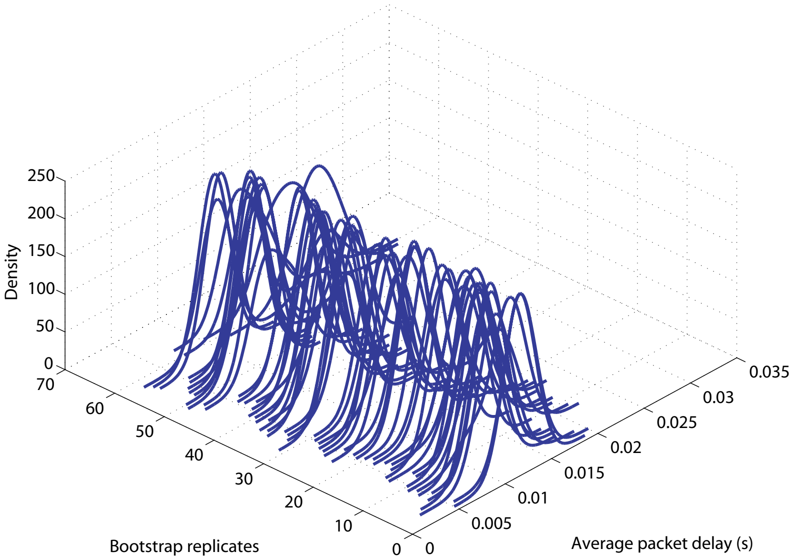
The simulation model is developed by using OPNET™ Modeler® 12.1 with Wireless Module. Minor modifications to the existing DCF models are performed without departure from the IEEE 802.11b standard for integration with the custom DAPU model. The robustness of the proposed technique is evaluated by utilizing three standard application models, viz., voice over internet protocol (VoIP), video conferencing, and video streaming to generate RT traffic. The first scenario simulates four STAs per AP with mixed traffic load. One pair of STAs transmits VoIP traffic and two pairs of STAs transmit video conferencing traffic where each VoIP and video conferencing session contains both UL and DL streams. The fourth STA of each AP does not transmit any traffic. This is to investigate the performance of the proposed technique under extreme condition in which both networks have almost similar load and PD. In the second scenario, the fourth STA of each AP engages in video streaming at two distinct instances in time to increase the load of the respective APs. For an infrastructure BSS, the DL becomes the capacity bottleneck in the presence of many two-way communications such as VoIP and video conferencing (see, e.g., Figure 6.11 of Section 6.4.1, [82], and [83]). Hence, the PD of interest is taken as the MAC delay experienced by the AP. Further details on the general simulation models can be found in Appendix A-2.

3.5.1 Verification of Bootstrap Approximation

The simulation results at different stages of the DAPU residing in AP 2 are presented in Figure 3.11(a). The DAPU first performs online PD measurements of the MAC delay in AP 2. The offline bootstrap Monte Carlo processing is triggered every 20 seconds which results in an original data set of 200 samples and consequently 200 bootstrap data sets after sampling with replacement from the original data set. Bootstrap replicate for each corresponding bootstrap data set is then calculated by computing the sample median. Finally, the average PD is estimated from the sampling distribution of 200 bootstrap replicates, which evidently shows that the DAPU is able to capture the dynamic behavior of the prevailing MAC delay. Next, Figure 3.11(b) illustrates that the sampling distribution of bootstrap replicates follows a normal distribution without recourse to the Central Limit Theorem. This is a result of employing smooth bootstrap, and it follows that the



(a) Bootstrap Monte Carlo procedures.



(b) Sampling distribution of bootstrap replicates.

Figure 3.11: Bootstrap approximation in DAPU of AP 2 for simulation scenario 1.

sample median of any continuous functions exhibits asymptotic normality [74]. This important property provides analytical tractability when incorporating the bootstrap estimates of QoS information from the DAPU to build a conditional posterior distribution during Bayesian learning. In fact, the closed-form expression for the Bayes estimate of QoS information is obtained in (3.12).

3.5.2 Verification of Bayesian Learning and CUSUM Monitoring

Both Figure 3.12 and Figure 3.13 demonstrate the main features of the synergy between Bayesian learning and CUSUM monitoring, which provide smoothing of bootstrap estimates in stationary conditions and fast adaptation when drastic state changes are encountered. In order to appreciate this synergy, consider the case of Bayesian learning without CUSUM monitoring. It follows that a small value of state noise variance Q^i provides accurate estimates, i.e., lower error variance P^i_k in stationary conditions but slow transient response during abrupt state change. On the contrary, a large value of Q^i ensures rapid response to abrupt state change but at the expense of inaccuracies, i.e., higher P^i_k. Hence, the introduction of CUSUM monitoring augments Bayesian learning by making it adaptive to abrupt state changes. Upon the detection of state change, Q^i which takes a sufficiently large value (of 1 in this case) will act to reset the Bayesian learning process by allowing it to forget the old estimate and converge rapidly to a new estimate. The sensitivity of CUSUM monitoring is governed by the drift parameter δ and alarm threshold h. A small value of δ provides faster detection to the fluctuations of the normalized residual s^i_k in (3.13) whilst a larger value of h will prevent false alarm but at the expense of a longer response time for detecting state changes. The rule of thumb for tuning CUSUM monitoring is to choose a large value of h and set δ to half of the magnitude of the expected state change [76]. However, the magnitude of expected change is usually not known a priori therefore the drift parameter $\delta = 0$ is chosen for faster detection. The alarm threshold $h = 5$, which is selected empirically, has been found to achieve satisfying results over an exhaustive range of simulation scenarios.

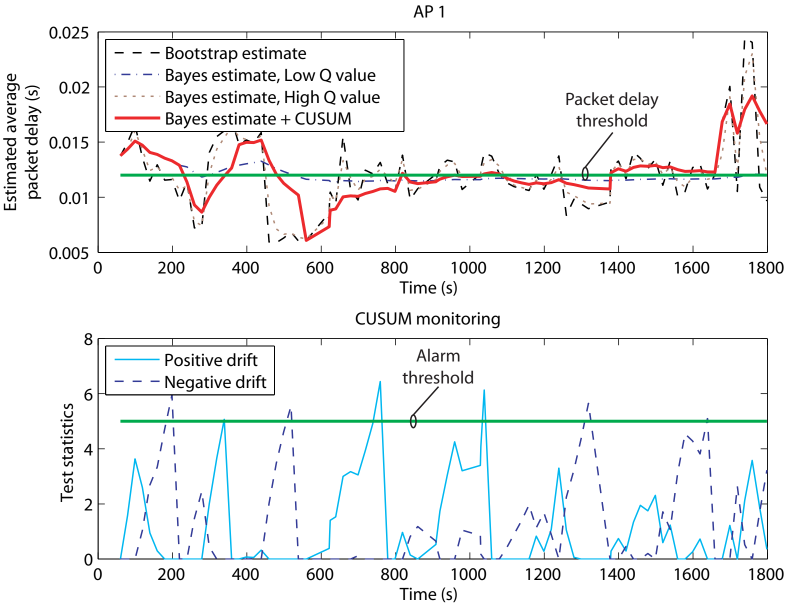


Figure 3.12: Bootstrap estimate, Bayes estimate, and CUSUM monitoring of AP 1 for simulation scenario 1.

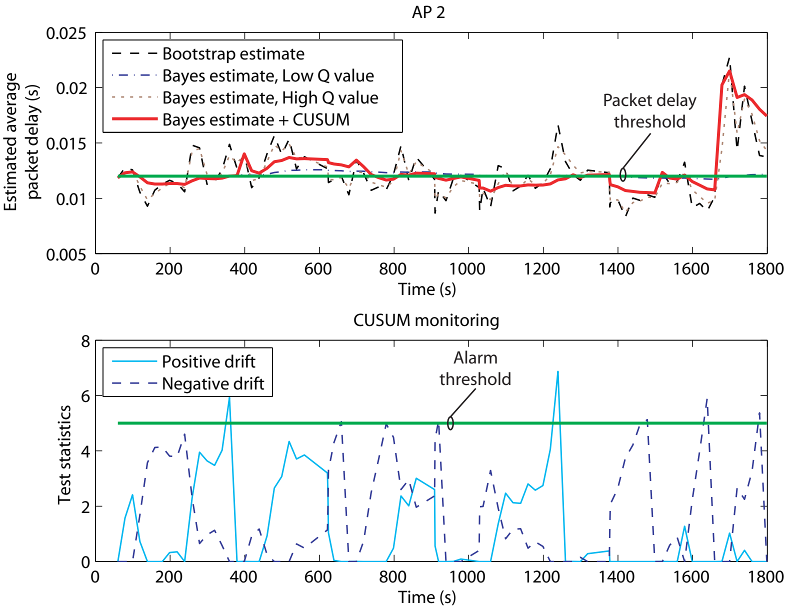
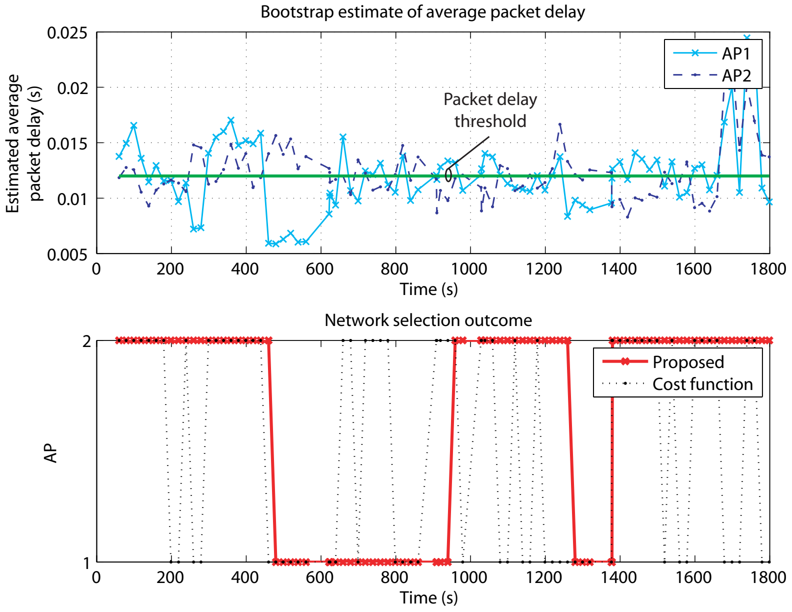


Figure 3.13: Bootstrap estimate, Bayes estimate, and CUSUM monitoring of AP 2 for simulation scenario 1.

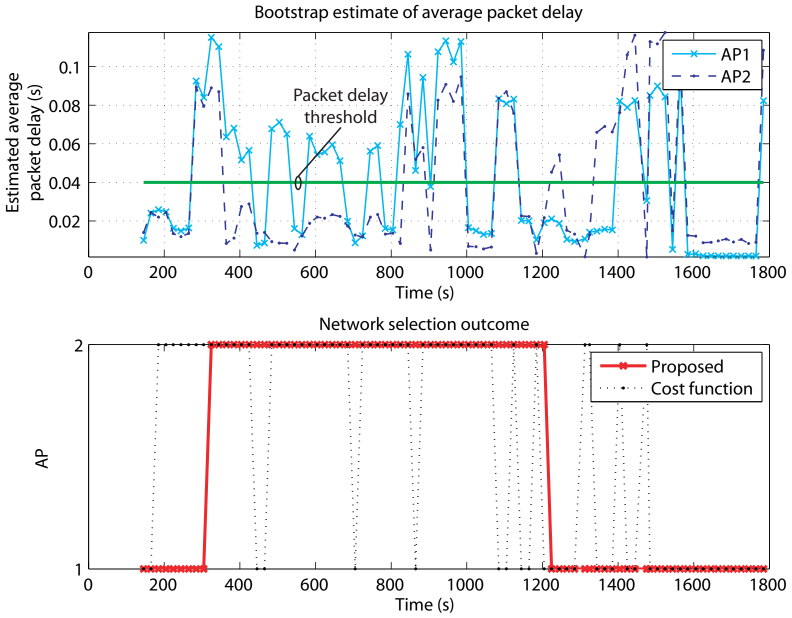
3.5.3 Evaluation of Network Selection Schemes

The outcome of the proposed technique for simulation scenario 1 is depicted in Figure 3.14(a). The proposed technique selects an optimal AP according to (3.17). A PD threshold of 12 ms, which coincides with the mean level of average PD in both APs, referred as the gray decision zone is chosen. The outcome of the proposed technique is then compared against the cost function approach which selects the ‘best’ AP as the one having the lowest MAC delay. The virtue of the proposed technique can be seen from its effectiveness in preventing unnecessary network selections which could well result in ‘ping-pong’ handovers if triggered. Clearly, the proposed technique can make a more informed network selection by considering both prior knowledge and measurement data, whereas the cost function approach acts on the instantaneous value of its evaluation. For simulation scenario 2, the PD threshold is raised to 40 ms, which again corresponds to the gray decision zone in both APs, as the fourth STA of each AP now participates in video streaming. It is apparent from Figure 3.14(b) that the proposed technique still yields better performance than the cost function approach in terms of a more reliable network selection outcome. To this end, a reasonable conjecture is that the filtering of unnecessary handovers is crucial for achieving system stability, minimizing the unjustified usage of scarce resources, and improving QoS performance. This conjecture will be further investigated in Section 3.5.4.

The robustness of the proposed technique has been validated using two scenarios indicative of the worst case in scenario 1 and the more general case in scenario 2. Next, the performance of the proposed technique is compared to four other network selection schemes. The cost function approach selects the ‘best’ AP as the one having the lowest MAC delay. The static selection scheme selects the ‘best’ AP as the one whose average MAC delay maximizes the CDF in (3.15). In other words, the static selection scheme is equivalent to the proposed technique without Bayesian learning. The two-state Markov chain selection scheme is implemented as in [84]. Lastly, the moving average selection scheme selects the ‘best’ AP in a similar fashion as the static selection scheme but with the exception



(a) Simulation scenario 1.



(b) Simulation scenario 2.

Figure 3.14: Network selection outcome.

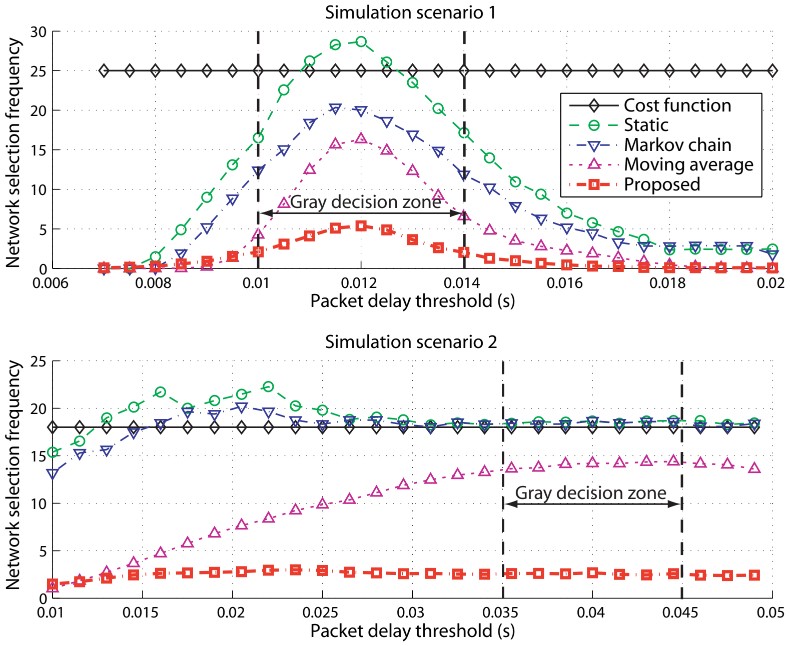


Figure 3.15: Performance evaluation of various network selection schemes.

that the mean of average MAC delay is first computed over a short window. In this study, the estimated MAC delays from the previous simulation scenarios are used to generate the uniformly distributed, pseudo-random scenarios of different MAC delay realizations. The comparison between network selection frequency of different schemes is based on an average of 1000 iterations per PD threshold. An additional 20% hysteresis is included in the static, Markov chain, and moving average selection schemes for supplementing stability in these network selection outcomes. Figure 3.15 illustrates that the proposed technique is superior in precluding unnecessary handovers or ‘ping-pong’ effect with the lowest network selection frequency without recourse to additional network selection reduction mechanism such as hysteresis. In particular, the proposed technique yields a significant improvement of at least twofold when the PD threshold is near the gray decision zone which is vulnerable to other network selection schemes.

3.5.4 Evaluation of Handover and QoS Performances

An efficient and reliable network selection technique is pivotal in a dynamic multi-RAT environment where heterogeneity can be exploited to improve the QoS and system capacity, particularly, when time-varying traffic and wireless channel conditions exist in practice. Although handover decision derived from the cost function approach can exploit heterogeneity promptly, it suffers from frequent network selection as presented in Section 3.5.3. Earlier, it is postulated that frequent network selections will cause unnecessary handovers, which have serious implications on the QoS and system capacity, in the presence of dynamic QoS parameters. To prove this conjecture, the handover and QoS performances of the proposed technique are compared to four cost function approaches, formed by incorporating three well-known network selection reduction mechanisms, viz., hysteresis, exponential weighted moving average (EWMA), and dwell time [43] to the baseline cost function approach, i.e., without any network selection reduction mechanisms.

A typical hotspot scenario of mixed IEEE 802.11b/g APs and STAs with the simulation parameters as summarized in Table 3.4 is simulated. Voice STAs are modeled as variable bit rate (VBR) sources to generate VoIP stream by using G.711 codec with silence suppression and packetization interval of 10 ms. Video STAs generate traffic according to the motion picture experts group (MPEG)-4 trace (Jurassic Park) [85] at 25 frames/sec, and data STAs generate best effort file transfer protocol (FTP) traffic with a mean data rate of 600 Kbps in the DL and 60 Kbps in the UL. The MAC service data unit (MSDU) lifetime limit mechanism is incorporated to discard MSDUs from the transmitter queue if they exceed the MSDU lifetime before successful transmission. A more detailed description of general simulation models is available in Appendix A-2.

From Figure 3.16, the proposed technique again demonstrates its effectiveness in providing reliable network selection, which is robust even under highly dynamic multimedia traffic, whilst the baseline cost function approach gives erratic network selections. As

Table 3.4: Simulation parameters.

WLAN Model (DCF)	Standard	Data Rate (Mbps)	PHY	Quantity
AP (STA)	IEEE 802.11b	11	HR/DSSS	1 (12)
AP (STA)	IEEE 802.11g	54	ERP-OFDM	1 (11)
Traffic Model (VBR)	**Codec /Protocol**	**Avg. Data Rate (Kbps)**	**MSDU Lifetime (s)**	**Quantity**
Voice	G.711	64	0.05	19
Video	MPEG-4	770	0.1	2
Data-UL (DL)	FTP	60 (600)	1	2

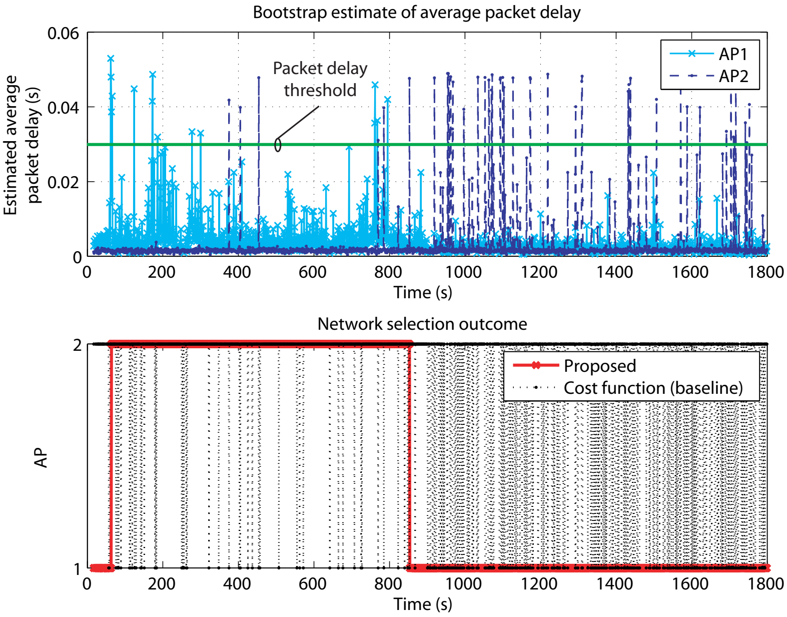
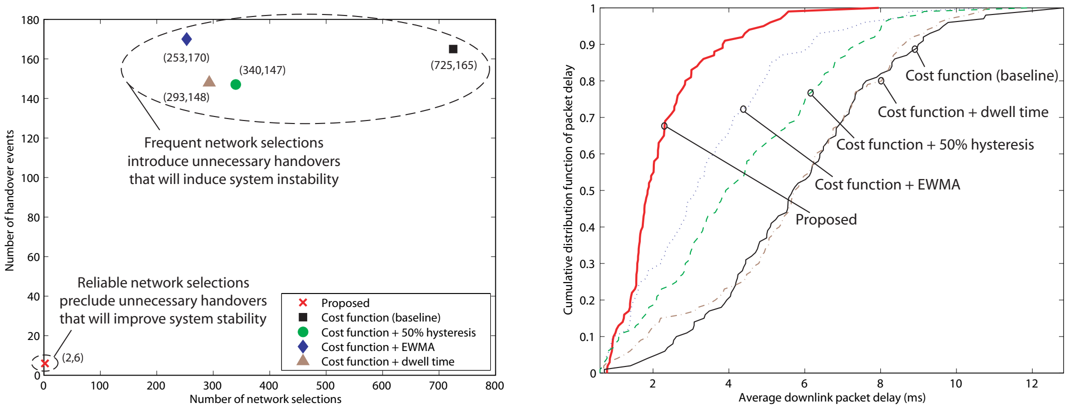


Figure 3.16: Estimated average PD and network selection outcome.

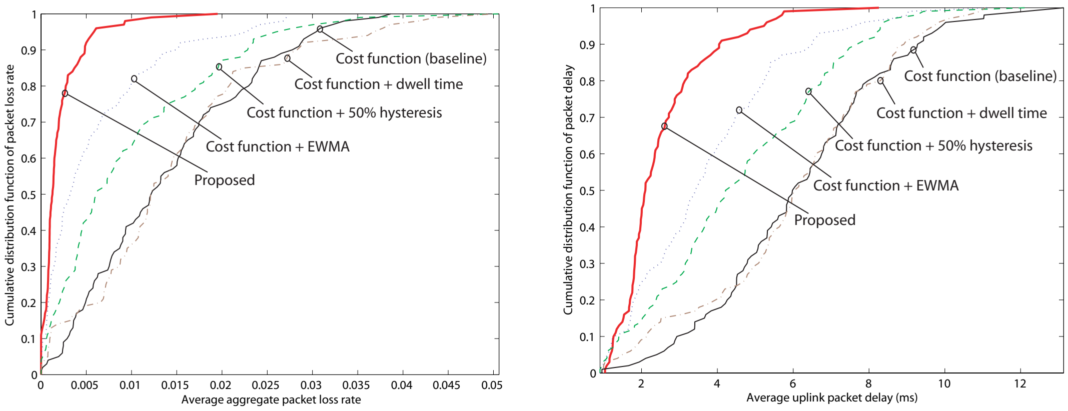
summarized in Figure 3.17(a), it is apparent that frequent network selections with all cost function approaches, despite the introduction of network selection reduction mechanisms, cause unnecessary handovers. On the other hand, the number of network selections and handovers are both dramatically reduced by more than 95% with the proposed technique. Figure 3.17(b) and Figure 3.17(c) illustrate that unnecessary handovers induce system instability which has a strong negative impact on the QoS performance of all the evaluated cost function approaches. In contrast, the proposed technique improves system stability which yields a significant improvement of 32% (61%) and 63% (84%) over the cost function+EWMA (baseline cost function) approach in both average DL PD and aggregate PLR, respectively. The average UL PD in Figure 3.17(d) shows similar trend as the average DL PD. Clearly, the salient benefits of these handover and QoS improvements translate to the reduction of signaling overheads and retransmissions, respectively which can be used to transmit useful traffic and boost system capacity. Therefore, these results confirm the conjecture in Section 3.5.3 where filtering of unnecessary handovers is crucial for achieving system stability, better utilization of radio resources, and improvement in QoS performance. It is also important to realize that the real benefit of terminal-oriented decision as discussed in Section 3.2.2 can be harnessed only by improving the system stability appropriately [9]. Although the effectiveness of the proposed technique is exemplified with a single QoS parameter in the context of WLAN, the results for multiple QoS parameters still hold according to expressions (3.25) and (3.26) for any given wireless networks.

3.6 Chapter Summary

This chapter has introduced a novel distributed TONA handover architecture to support the convergence of heterogeneous access networks through the IP-based core network. In addition, a novel generic DANS algorithm has been developed to provide a pragmatic way to estimate dynamic QoS information, which is particularly relevant to network selection mechanisms in future wireless networks. The key principle of a technology agnostic ap-



(a) Handover events vs. network selections. (b) Average DL PD.



(c) Average aggregate PLR. (d) Average UL PD.

Figure 3.17: Handover and QoS performances evaluated at an average system utilization of 90%.

proach has been conceived from the unification of both TONA handover architecture and DANS algorithm. This fusion forms a bi-domain cooperation in which QoS parameters estimation is a cornerstone of the technology agnostic approach adopted in the generalized CCRRM architecture to support access network heterogeneity and provide link layer cognition for coordinating informed RRM decisions in future wireless networks.

The key insight of this chapter has revealed that although the widely adopted cost function approach suffices when evaluating static QoS parameters, it inherently results in frequent, and often, unnecessary handovers which will be detrimental to the QoS and system capacity when considering dynamic QoS parameters. Extensive simulations have demonstrated that the DANS algorithm can effectively augment the existing cost function approach with an additional handover initiation module (cf. Figure 3.5) to offer link layer cognition for filtering unnecessary handovers. This provides an optimal network selection outcome and consequently improves the QoS and system capacity. The link layer cognition is attributed to the conditional posterior distribution derived from the Bayesian learning process, which is essentially a generalization of the Kalman filter, within the handover initiation module. Moreover, the proposed dual-stage estimation process is computationally efficient to implement in practice as the Bayes estimates, which is used to quantify the network quality, can be derived in closed-form expression. To be more specific, the proposed technique has a linear time complexity of $\mathrm{O}\left(\mathrm{n}\right)$ where n is the number of AP which will be bounded in practice due to limiting SNR.

The next chapter investigates the benefits of bi-domain cooperation in the context of QoS provisioning for delivering both VoIP and multimedia traffic over WLAN. In particular, the concept of RQB, which has favorable intrinsic properties for optimizing radio resource usage and effectuating a QoS-balanced system, is advocated. It is worth mentioning that the concept of RQB which quantifies the state of balance between access networks is also based on QoS parameters. Hence, this further reinforces the technology agnostic approach of the generalized CCRRM architecture.

CHAPTER 4

BI-DOMAIN COOPERATION

The technology abstraction and link cognition module of the generalized CCRRM architecture presented in Chapter 3 is based on the key principle of technology agnostic approach, which is obtained from the fusion of inter-network cooperation and inter-entity cooperation to cast the concept of bi-domain cooperation. The distributed TONA handover architecture enables inter-network cooperation to facilitate the cooperative exchange of QoS context information between heterogeneous access networks while the generic DANS algorithm enables inter-entity cooperation between network-terminal entities to make an informed initial access or handover decision. Through the notion of bi-domain cooperation, an efficient iLB scheme can be devised to offer RQB by incorporating fast handover in conjunction with soft admission control. The novel concept of RQB will harmonize VHO in future wireless networks so that the end-to-end goal of maintaining a QoS-balanced system can be accomplished.

This chapter examines the benefits of RQB based on bi-domain cooperation which forms the baseline design of the generalized CCRRM architecture. The objective of the generalized CCRRM architecture is to exploit the heterogeneity within complementary future wireless networks to coordinate better utilization of radio resources in an opportunistic yet altruistic manner. The importance of *cooperation* is demonstrated through its ability to exploit heterogeneity as an enabler for improving the overall system capacity and QoS of end-users. Furthermore, the generalized CCRRM architecture benefits from the unified actions of joint optimization and results in a QoS-balanced system by enabling different functional entities to form synergies and multiple access networks to interact. Consequently, the notion of QoS balance as the criterion to quantify the state of balance in future wireless networks is strongly advocated. In recent times, there is also an emerging need for cooperation in wireless networks [9], [14], [34], particularly, between different RATs to support: (i) VHO for service continuity and QoS transparency in order to realize seamless mobility; and (ii) efficient management of pooled resources [86].

This chapter is outlined as follows. Section 4.1 presents the challenges associated with multimedia service delivery over WLAN. Section 4.2 describes an efficient iLB scheme to provision QoS for VoIP over WLAN, also known as VoWLAN, and multimedia service delivery over WLAN based on the concept of bi-domain cooperation. Through simulations, Section 4.3 shows that statistical QoS guarantee can be provisioned for both VoIP and multimedia traffic. Moreover, both throughput and QoS fairness can be achieved, and overall system capacity can be maximized by using the iLB scheme, if cooperation is adopted to maintain a QoS-balanced system. Finally, the main conclusions from this chapter are listed in Section 4.4.

4.1 Challenges of Delivering Multimedia Services over WLAN

With the emerging IEEE 802.11n standard [87], WLAN is poised as a promising ubiquitous networking technology to support multimedia applications where providing QoS becomes imperative. The increasing popularity of multimedia applications such as VoIP, video streaming, and data have made their unification over WLAN compelling since they can now leverage on the pervasive WLAN of high bandwidth for user mobility. However, the support of RT VoIP and video services over WLAN poses numerous challenges such as QoS provisioning, admission control, and load balancing since WLAN is not designed to support delay sensitive traffic.

One of the main challenges in QoS provisioning for WLAN is to support RT connections with seamless handover as dynamic network conditions may result in unacceptably high PD and consequently packet loss. E.g., VoIP requires one-way end-to-end delay of less than 150 ms [88] but can tolerate some PLR of up to 2% [89]. This implies that the total handover latency and packet loss should not exceed these bounds in order to sustain an undisrupted VoIP call of acceptable QoS. Moreover, this problem is magnified during a handover which typically results in excessive handover latency and packet loss. Here, the focus of this thesis is on minimizing Layer 2 handover latency, which composes of detection delay, scanning delay, authentication delay, and reassociation delay where both detection and scanning delays are found to be the dominating costs in [90] and [91]. It will be shown in Section 4.2.1 that fast handover can be achieved for RT connections by eliminating: (i) detection delay when link layer detection is exploited to trigger VHO; and (ii) scanning delay when DANS algorithm is employed to provide the information of an optimal target AP without the need to invoke the scanning phase.

In addition, the WLAN handover process is predominantly based on PHY detection without QoS considerations. This often causes the overloading of APs and consequently all the associated connections would suffer from high delay. Garg and Kappes [92] show that it is crucial to determine the network capacity that can be supported by the DCF, in terms of the maximum number of simultaneous VoIP connections, since its effective bandwidth is significantly reduced by inherent overheads which limit the maximum number of VoIP calls to a small number. Their study also suggest that admission control is vital for an infrastructure BSS WLAN to protect existing VoIP connections. Similarly, Zhai *et al.* [93] find that WLAN attains maximum throughput and low delay when operating in the non-saturation mode due to low collision probability, suggesting that admission control is a suitable strategy for RT traffic due to its low bandwidth but strict delay requirements. Interestingly, Chen *et al.* [94] show that although the IEEE 802.11e standard supports prioritized QoS, it cannot guarantee strict QoS required by RT services under heavy load without an appropriate network control mechanism. Here, the sporadic overloading of

APs is mitigated by introducing a unified fast handover and soft admission control iLB scheme to perform RQB.

There are numerous research works on enhancing QoS support for WLAN either through admission control or load balancing. However, a unified approach to provision QoS through a comprehensive and generalized CCRRM architecture has not been adequately studied in literature. In general, the choice of an appropriate load metric is pivotal in any admission control and load balancing schemes as it serves to estimate the available network capacity. For circuit-switched cellular networks such as GSM, load balancing is traditionally based on the number of active calls per cell as its load metric since the load contributed by each user is the same. However, Bianchi and Tinnirello [95] demonstrate that load balancing in packet-switched wireless networks such as WLAN can be improved by using additional 'packet level' load metrics such as gross load, which considers retransmissions, and packet loss. Subsequently, Garg and Kappes [96] offer a CU estimation technique which gives the best representation of the effective network load.

Accordingly, the CU that measures the fraction of channel occupation time per observation interval has been widely used as the load metric for both load balancing and admission control algorithms due to its simplicity and high accuracy in estimating effective network load. Zhai *et al.* [97] employ channel busyness ratio as the load metric for their admission control and rate control scheme to provide statistical QoS guarantee for VoIP traffic and maintain high throughput for best effort flows in the IEEE 802.11b/e WLANs. Bazzi *et al.* [98] develop a measurement-based call admission control, which uses either the channel occupancy or queue size of AP as the load metric, to protect the QoS of existing connections by denying incoming calls when resources are low. However, the parameters of their call admission control require tuning for different traffic mixes, and hence they are not adaptive to dynamic network conditions. Moreover, the works reported in [97] and [98] do not consider load balancing feature which makes it unlikely to optimize overall system capacity.

On the other hand, Daher *et al.* [99] consider load balancing and incorporate admission control within a centralized architecture using medium busy time as the load metric. Balachandran *et al.* [100] present an adaptive load balancing solution where a centralized admission control server contains the load information of all APs and is solely responsible for making RRM decisions, but it utilizes the throughput of AP as their load metric. However, this approach requires STA to perform service level negotiation with the admission control server prior to both initial access and handover. Although the authors propose using retransmissions to trigger handover, the associated handover latency is not investigated and may be detrimental to RT connections. Furthermore, the works reported in [99] and [100] require a fully centralized RRM which is prohibitive in handling time critical information necessary to make detailed RRM decisions concerning the end-users or APs. In contrast, Velayos *et al.* [101] propose a decentralized load balancing scheme which also uses the throughput of AP as their load metric. However, throughput according to [96] is not a suitable load metric as it is highly influenced by the data rate of STAs running different applications and variable transmission data rate due to dynamic wireless channel conditions which affect link quality. Moreover, the major pitfall of this scheme is that STA will experience service outages during a handover since it must first disassociate from the source AP and can reassociate only with an underloaded target AP after some searching time has elapsed.

To the best of the author's knowledge, there is no prior research work on QoS balancing scheme with intrinsic properties of providing statistical QoS guarantee while optimizing system utilization by considering fast handover in conjunction with soft admission control to maintain a QoS-balanced system. The principal contributions of this chapter differ from the related works in five significant ways: (i) QoS provisioning for VoWLAN with heterogeneous voice codecs and multimedia service delivery over WLAN are treated from a single unifying generalized CCRRM architecture; (ii) bi-domain cooperation is identified within the harmonizing generalized CCRRM architecture to promote a QoS-balanced system by exploiting the heterogeneity of a multi-AP WLAN; (iii) the notion of RQB is introduced in the proposed iLB scheme, which has intrinsic properties of providing statistical QoS guarantee for both VoIP and multimedia traffic while maximizing overall system capacity, as the criterion to quantify the state of balance in a multi-AP WLAN; (iv) the iLB scheme is lightweight and adaptive to dynamic network conditions by leveraging only on the estimated critical QoS information of PD as the criterion to select an optimal target network for handover and as the load metric for soft admission control; and (v) an evaluation of the system cost involved in the generalized CCRRM architecture and an analysis of tradeoffs between QoS performance including the number of handover events and QoS broadcast intervals are given.

4.2 Integrated Load Balancing Scheme

The design philosophy of the iLB scheme first proposed in [102] and [103] is based on the novel concept of RQB to exploit heterogeneity within a multi-AP WLAN where APs are physically co-located in an opportunistic yet altruistic manner. Under the notion of RQB which promotes a QoS-balanced system, a handover will be triggered *only* if: (i) the QoS requirements of STAs cannot be sustained; (ii) a better quality AP exists; and (iii) the requested handover will not disadvantage the existing connections of the target AP. The context of disadvantage here refers to the situation when existing connections *fail* to meet their QoS requirements as a result of that handover. The opportunistic yet altruistic exploitation is achieved when all the above conditions are met. It is important to note that the first two conditions will preclude unnecessary handovers due to the reactive and opportunistic handover triggering approach. In addition, the first two conditions will invoke the self-adjusting nature of the generalized CCRRM architecture, which will adapt to both traffic and wireless channel variations. The use of QoS parameters to capture both traffic and wireless channel variations is the chief advantage of the generalized CCRRM architecture in realizing the technology agnostic approach to support access network heterogeneity.

Accordingly, the proposed iLB scheme incorporates: (i) fast handover to support seamless handover by exploiting link layer detection to eliminate detection delay and employing DANS algorithm to eliminate scanning delay from the WLAN handover process; and (ii) soft admission control to protect the QoS of existing connections when resources are low. The basic idea of this synergy is to protect the QoS of RT services from network overloading by performing RQB to trigger VHO in an opportunistic yet altruistic manner. More importantly, the QoS requirements of STAs are statistically guaranteed *during* handover by enabling seamless handover with fast handover and *after* handover by operating WLAN in the non-saturation mode with soft admission control.

The advantages of RQB to achieve the end-to-end goal of maintaining a QoS-balanced system are manifold which will be sequentially unraveled in this and the subsequent chapters. First, it inherently provides statistical QoS guarantee for both VoIP and multimedia traffic. Second, it maximizes the overall system capacity through better utilization of radio resources by maintaining QoS and throughput fairness. Third, it precludes unnecessary handovers. Fourth, it mitigates rate anomaly in multirate environment through synergetic interactions between link adaptation and load adaptation (cf. Section 5.5.1). Fifth, it preserves baseline QoS (cf. Section 7.5.3).

4.2.1 iLB Algorithm

The algorithm of the proposed iLB scheme is depicted in Figure 4.1. The shaded blocks refer to network entities while the unshaded blocks refer to terminal entities. ABC services which consider both network conditions and user preferences during network selection in dashed lines can also be supported as discussed in Section 3.4.4. However, these are outside the scope of this chapter. The proposed iLB scheme can be triggered by two events, viz., *initial access* to network where STA would choose the optimal AP according to its PD (QoS requirement) and *handover* when the PLR of AP (network QoS) exceeds 2% for the case of RT services such as VoIP and video. Soft admission control located

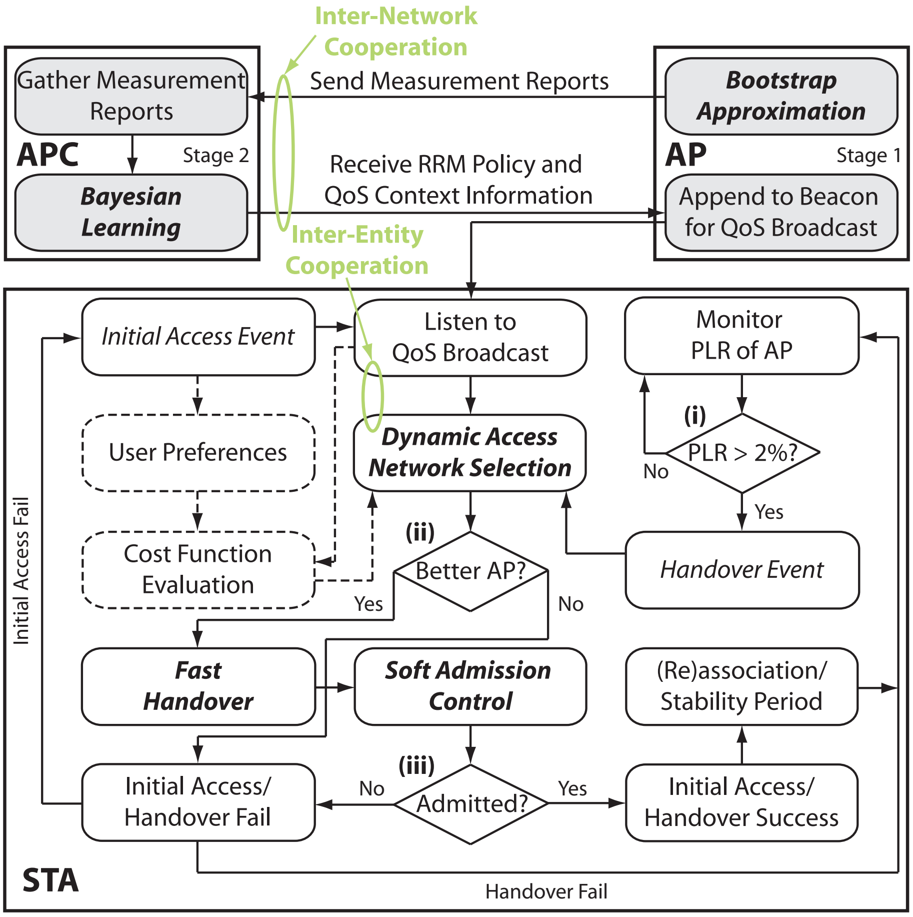


Figure 4.1: Algorithm of the iLB scheme based on bi-domain cooperation.

in each STA would arbitrate the prevailing QoS, in terms of PD, between the source and target APs. Upon admission, STA will perform (re)association with the optimal target AP during (handover) initial access. Otherwise, STA will continue to monitor the PLR of its associated AP when handover fails or listen to the QoS broadcasts for RRM policy and QoS context information when initial access fails. A stability period of ten QoS broadcast intervals is enforced, before other STAs can make the next handover attempt, to prevent the handover synchronization problem [5] or better known as ‘ping-pong’ effect.

The iLB scheme is built upon the concept of bi-domain cooperation within the generalized CCRRM architecture, viz., inter-network cooperation with the TONA handover architecture and inter-entity cooperation with the DANS algorithm. Specifically, it performs RQB by using PLR for link layer detection and *only* the estimated PD as the criterion for network selection (cf. Section 3.4) and as the load metric for soft admission control, making it lightweight and adaptive to dynamic network conditions. It is worth noting that the RRM policy adopted in this chapter is one which requires *only* voice STAs to perform RQB through VHOs. The motivation is to examine whether the legacy DCF WLAN without service prioritization could provision QoS guarantee for both VoIP and multimedia traffic by leveraging on the novel concept of RQB.

Fast Handover

As exposited in Section 3.2, STA will listen to the QoS broadcasts, which contain RRM policy and QoS context information, and select the optimal AP according to its PD estimates. This enables the obviation of both detection and scanning phases in the IEEE 802.11 handover process as STA listening to the broadcast would be able to get the information of prospective neighboring APs. Consequently, this leads to significant Layer 2 handover latency reduction and optimizes the STA's power consumption, which will be illustrated in the following by comparing the handover processes between the existing IEEE 802.11 WLAN and iLB scheme.

Figure 4.2 depicts the existing IEEE 802.11 WLAN handover process with active scan mode. Note that a total handover latency of more than 1000 ms [91] is expected when link layer detection is used only to trigger handover. The main reason is because link layer detection is based on the positive ACK mechanism of the IEEE 802.11 DCF. Hence, the transmitting STA cannot differentiate between collision, congestion, or being outside the coverage of an AP during occasions when the ACK is not received. As a result, most of the proprietary algorithms in commercial devices first perform link adaptation procedures, followed by using the request to send (RTS)/clear to send (CTS) mechanism to probe the link after repeated transmission failures in order to eliminate collision as the cause of transmission failure. Subsequently, the scanning phase will be invoked only after several unsuccessful RTS/CTS handshakes. Although PHY detection is widely deployed to exclude the delay of link layer detection, it results in non-uniform load distribution due to the sporadic congestion of APs and the total handover latency, which is dominated by the scanning delay, can still be as high as 420 ms [90].

To be more specific, there are two types of scan modes supported by the IEEE 802.11 WLAN, viz., passive or active mode. The average probe delay of passive scan can be expressed simply as a function of the beacon interval BI, number of available channels n, and channel switching time CST, which are dependent on implementation, as $T_{passive} = n\,(BI + CST)$. In contrast, the average probe delay of active scan is determined by the $MinChannelTime$ and $MaxChannelTime$ values, which are again dependent on implementation, bounded by $n\,(MinChannelTime + CST) \leq T_{active} \leq n\,(MaxChannelTime + CST)$. E.g., $T_{passive} =$ 1120 ms and 130 ms $\leq T_{active} \leq$ 230 ms for the IEEE 802.11b with typical values of $BI =$ 100 ms, $n =$ 10 (excluding current channel), $MinChannelTime =$ 1 ms, $MaxChannelTime =$ 11 ms, and $CST =$ 12 ms [104].

Hence, the salient advantage of the proposed iLB scheme is the ability to support fast handover for RT services by eliminating both detection and scanning delays, thanks to the generalized CCRRM architecture. This is possible since the fact that RT VoIP and video services can tolerate some PLR of 2% is exploited and utilized as link layer detection to trigger handover. Given that the optimal target AP information is available from the DANS algorithm at the same instance, detection delay will not be incurred. Consequently, the total Layer 2 handover latency illustrated in Figure 4.3 is significantly reduced to approximately 16 ms to 30 ms. Moreover, this fast handover feature has significant importance during inter-technology handover or VHO as soft handover is usually not supported.

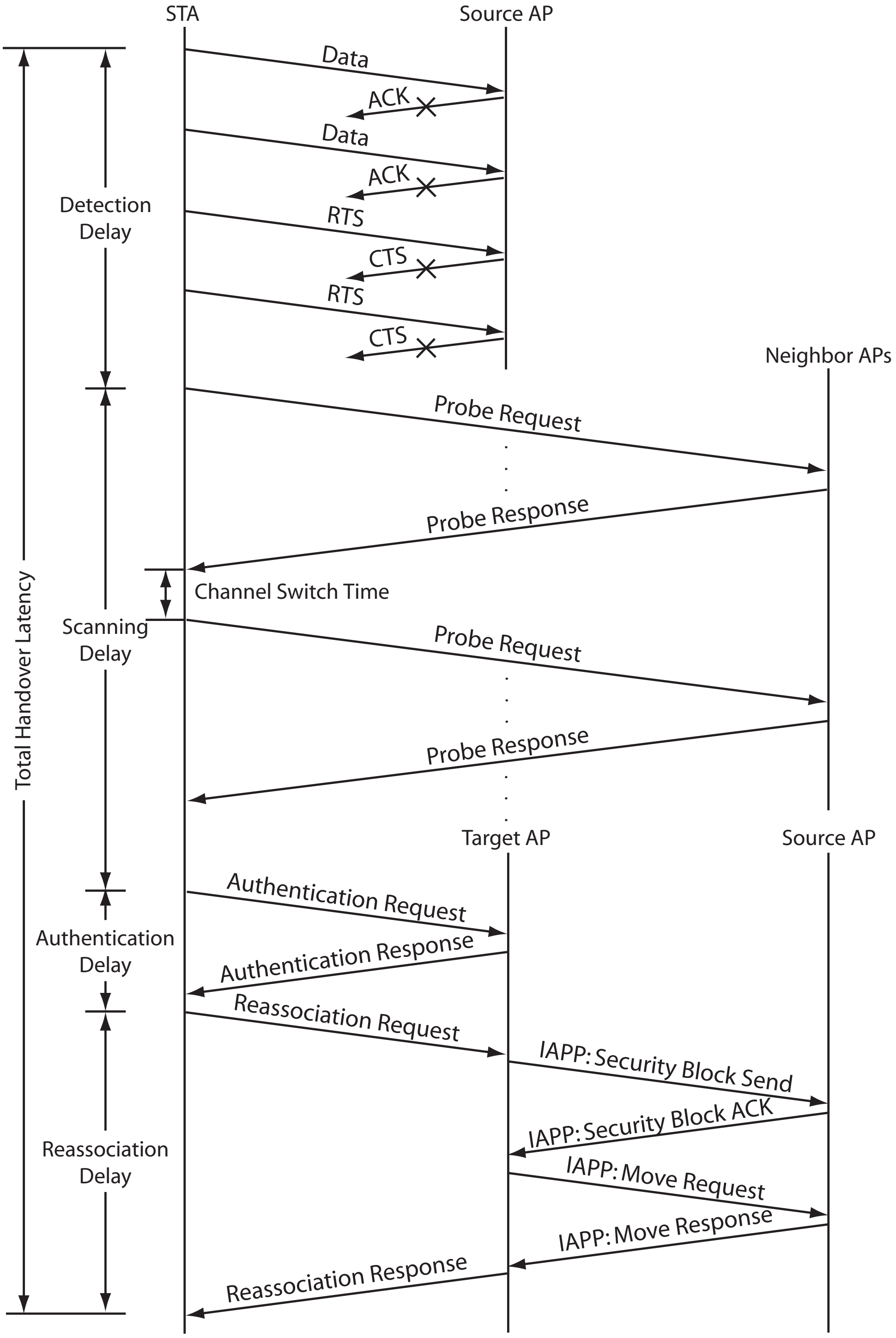


Figure 4.2: Handover process of the existing IEEE 802.11 WLAN.

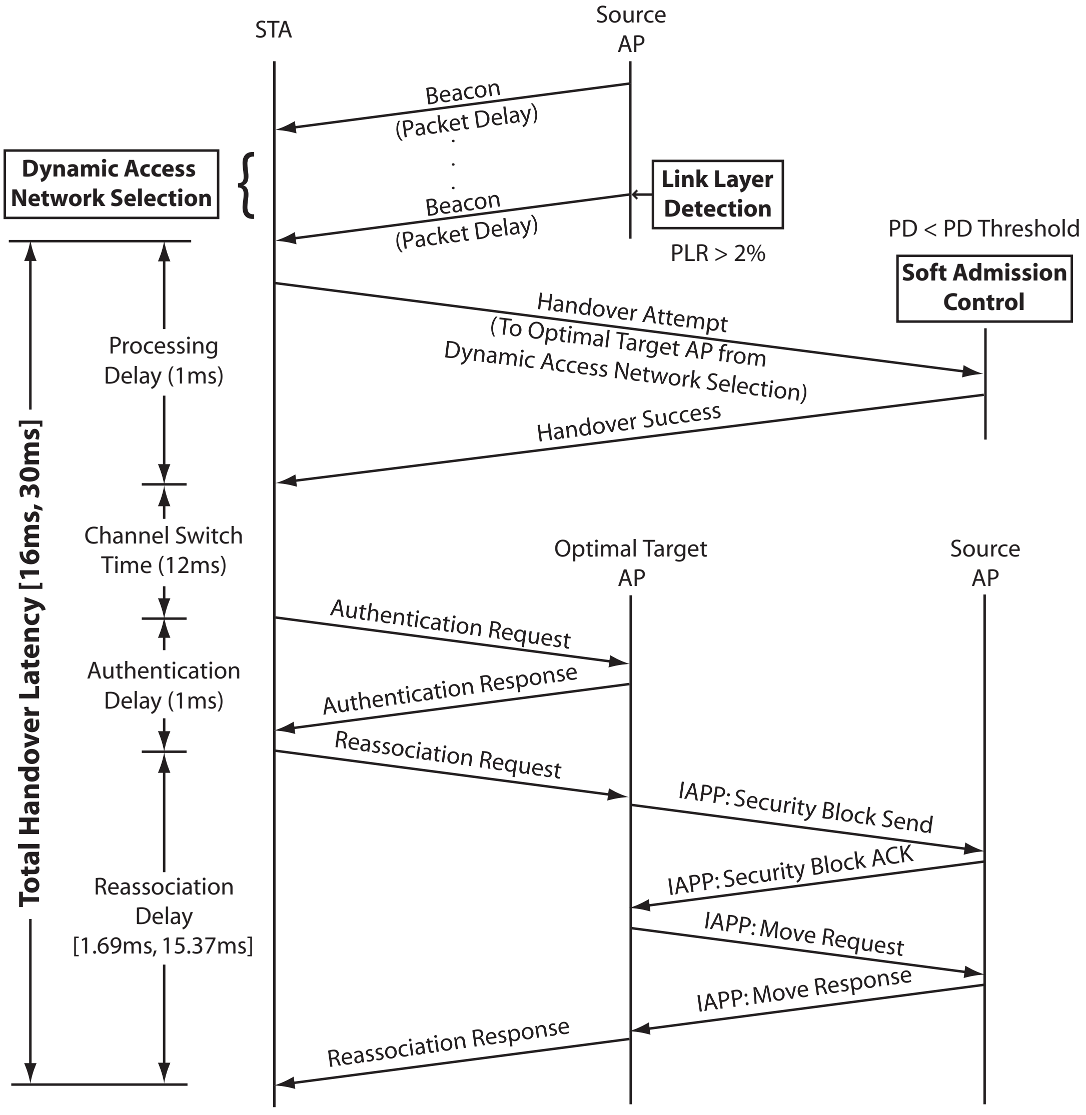


Figure 4.3: Seamless handover process of the iLB scheme.

Soft Admission Control

As previously mentioned, the estimated PD is also augmented as the load metric to devise a soft admission control which is simple yet effective as it considers dynamic network conditions prevalent in broadband WLANs. Here, a distinction is drawn between the hard admission control and soft admission control. The former is traditionally used for homogeneous voice traffic where network capacity can be easily pre-determined. This enables the number of connections, which is typically fixed, to be used directly as admission threshold. The latter mitigates the difficulty of estimating the actual bandwidth occupancy of services when considering heterogeneous traffic of different arrival rates, packet lengths, and burstiness in the presence of prevailing wireless channel conditions. In particular, the variability of data transmission rates and frame retransmissions will add to the complexity of estimating any pre-determined network capacity.

Accordingly, *soft* refers to the number of admissible connections which is not fixed but variable depending on the class of services, e.g., RT and NRT, the type of traffic sources, e.g., constant bit rate (CBR) and VBR, the proportion of traffic mixes, and prevailing wireless channel conditions. The key idea is to ensure that the PD threshold of an AP is not violated when accepting new connections, which effectively protects the QoS of existing connections, by maintaining WLAN in the non-saturation mode. Thus, soft admission control is important when considering VoIP traffic with heterogeneous voice codecs and multimedia traffic since traditional hard admission control, which applies pre-determined network capacity directly as admission threshold, is ineffective against such dynamic network conditions.

4.3 Performance Evaluation of the iLB Scheme

To evaluate the performance and effectiveness of the iLB scheme, built on the basis of bi-domain cooperation, two separate studies are conducted under ideal channel condi-

Table 4.1: VoIP traffic generation parameters.

Traffic Type	Packet Size (Bytes)	Inter-arrival (ms)	Avg. Data Rate (Kbps)
G.711 (VBR)	80	10	64
G.729 (VBR)	20	20	8
G.723.1 (VBR)	24	30	6.4

tions. The first study concentrates on the VoWLAN while the second study focuses on the multimedia service delivery over WLAN. The simulation models are developed by using OPNET™ Modeler® 14.0 with Wireless Module. Modifications are performed to the existing DCF models for integration with the custom DAPU model. Additionally, QoS support with network selection based on the DANS algorithm (cf. Section 3.4), fast handover, and soft admission control is provided.

In the first study, a typical hotspot scenario, which consists of a homogeneous multi-AP WLAN with two IEEE 802.11b APs operating at the maximum data rate of 11 Mbps, is simulated according to Figure 4.4. The VoIP packets are encoded by using three popular voice codecs, viz., G.711, G.729, and G.723.1 with the packetization interval of 10 ms, 20 ms, and 30 ms, respectively as shown in Table 4.1. In this simulation, an unbalanced load of thirteen G.711, two G.729, and one G.723.1 STAs in BSS 1 while five G.711, one G.729, and two G.723.1 STAs in BSS 2 is initially introduced. At time 300 s, two G.711 connections are started in BSS 2. Subsequently, at time 600 s, one G.711 connection is started in BSS 1 while three G.711 connections are stopped in BSS 2.

The second study simulates a typical hotspot scenario which consists of a heterogeneous multi-AP WLAN with one IEEE 802.11b AP and one IEEE 802.11g AP operating at the maximum data rates of 11 Mbps and 54 Mbps, respectively as depicted in Figure 4.5. The simulation is subjected to multimedia traffic sources as summarized in Table 4.2. Voice STAs are modeled as VBR sources to generate VoIP stream by using the G.711 codec with silence suppression and packetization interval of 10 ms. Video STAs generate traffic according to the MPEG-4 trace (Jurassic Park) [85] at 25 frames/sec, and data STAs generate best effort FTP traffic. In this simulation, an unbalanced load of seven voice,

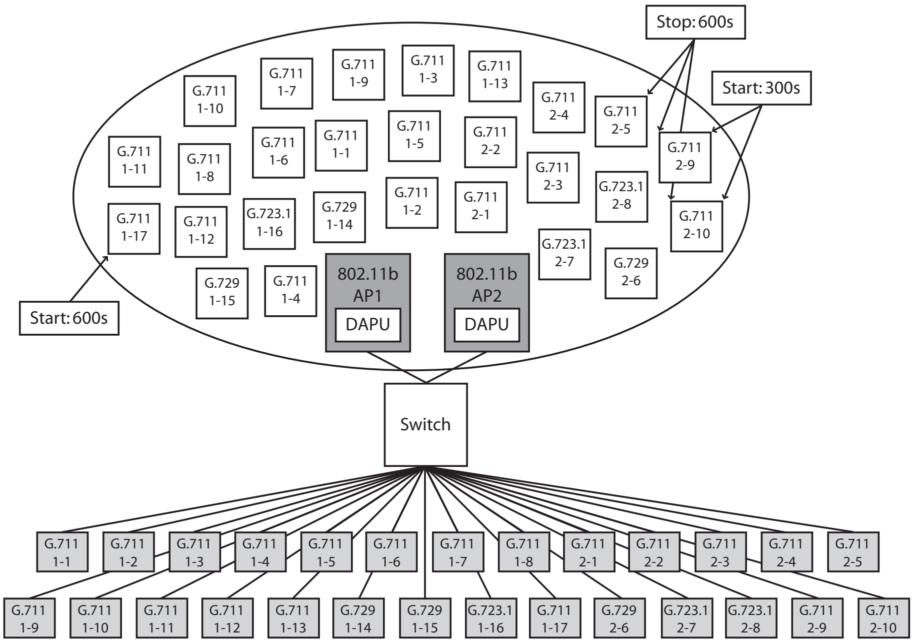


Figure 4.4: Simulation model of a homogeneous multi-AP WLAN with the IEEE 802.11b APs.

Table 4.2: Multimedia traffic generation parameters.

Traffic Type	Packet Size (Bytes)	Inter-arrival (ms)	Avg. Data Rate (Kbps)
Voice-G.711 (VBR)	80	10	64
Video-High Quality	MPEG-4 trace	40	770
Data-FTP (UL)	750	100	60
Data-FTP (DL)	3750	50	600

two video, and two data STAs in BSS 1 whilst seven voice STAs in BSS 2 is initially introduced. At time 900 s, five voice, one video, and one data connections from BSS 1 are stopped while five voice connections from BSS 2 are started. Note that these discrete events induce imbalance traffic load during the simulation for evaluating the responsiveness of the iLB scheme under such dynamic network conditions. Further, notice that no perturbations are injected after 600 s and 900 s for the first and second studies, respectively in order to observe the steady state performance. The MSDU lifetime limit mechanism is incorporated to discard MSDUs from the transmitter queue if they exceed the MSDU lifetime before successful transmission. Additional details on the general simulation models can be found in Appendix A-2.

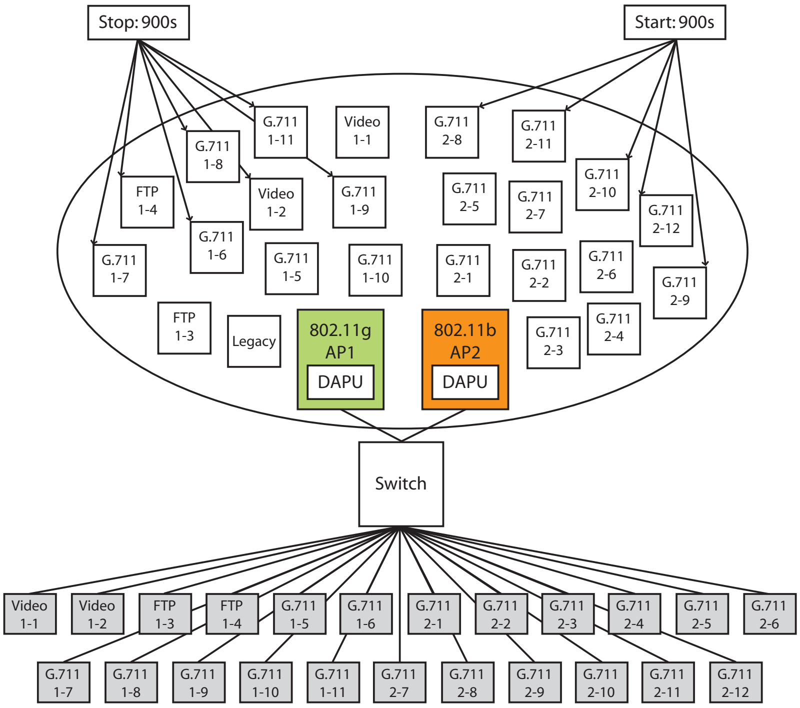


Figure 4.5: Simulation model of a heterogeneous multi-AP WLAN with the IEEE 802.11b/g APs.

The performance of the proposed iLB scheme is investigated from two critical aspects. First, the QoS performance in terms of the PD and PLR of APs, which reflects the capability of WLAN to support RT services, is examined as the APs are the bottleneck links. In the study of VoWLAN, the throughput performance is also presented. Second, the effect of RQB on the overall system utilization is quantified by adopting the QoS balance index (QBI) in (A-9) of Appendix A-3.1 to reflect the fairness of throughput and QoS conditions among the APs.

4.3.1 Voice over WLAN with Heterogeneous Voice Codecs

In this study, the QoS performance of the iLB scheme is evaluated in terms of PD, throughput, and packet loss. A comparative analysis between the performance of the iLB scheme and DCF, which represents the case without RQB, is also investigated. The comparison between iLB and DCF is of interest due to the fact that majority of the existing

WLANs are DCF-based. Moreover, they support only PHY detection and lack admission or load control mechanism to prevent the overloading of APs.

First, the average UL and DL PDs associated with each AP are investigated. Each VoIP connection has duplex traffic which eventually results in higher DL load, leading to classical capacity bottleneck at the AP of an infrastructure BSS VoWLAN (see, e.g., Figure 6.11 of Section 6.4.1, [82], and [83]). As such, the average DL PD is worse than its average UL PD as shown in Figure 4.6. The results also demonstrate that bounded average UL and DL PDs of less than 20 ms are achievable throughout the simulation. This corresponds very well to the induction that the total Layer 2 handover latency is less than 30 ms and confirms the ability of iLB to support fast handover. Furthermore, it is noticed that AP 1 is overloaded while AP 2 is under-utilized for the case without iLB. This is predominantly due to PHY detection of the existing IEEE 802.11 WLAN handover process which lacks QoS considerations. As a result, no handover is triggered since all STAs are within the good coverage region of their APs. Therefore, AP 1 has significantly higher average UL and DL PDs as compared to AP 2. It is worth noting that an average DL PD of up to 550 ms is experienced in AP 1 for the case without iLB.

Second, both the UL throughput and DL throughput are investigated. The former refers to the aggregate throughput of APs and the latter refers to the aggregate throughput of STAs. The simulation results reveal that UL throughput performance is similar for both iLB and DCF as shown in Figure 4.7(a). This is not surprising as the DL instead of the UL is the capacity bottleneck for an infrastructure BSS VoWLAN. This is a direct consequence of fairness in the IEEE 802.11 basic access scheme of DCF. To be more specific, although APs carry much higher traffic load than STAs in an infrastructure BSS VoWLAN, both contend with the *same* priority. For DL throughput, it is noticed that iLB brings about an average enhancement of 6% as shown in Figure 4.7(b). This enhancement over the case without iLB is due to the buffer overflow phenomenon in AP 1 which is operating beyond its maximum capacity, and hence it experiences excessive packet loss. Although the DL throughput enhancement is moderate, it is worth noting that both average UL and DL PDs

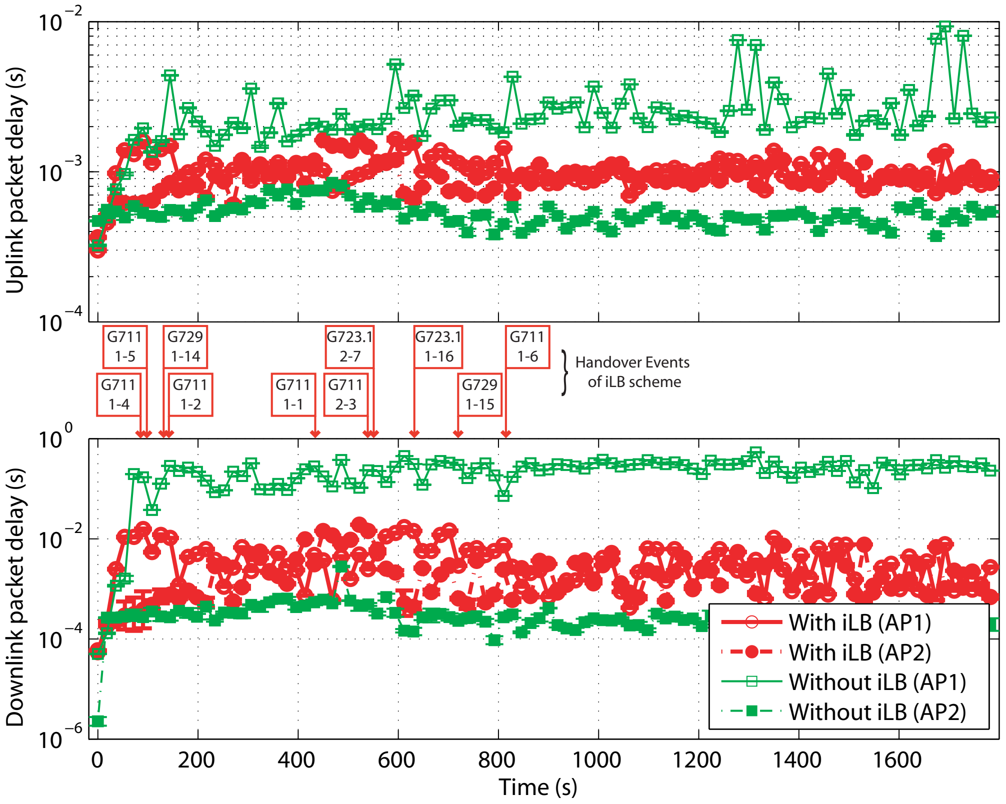
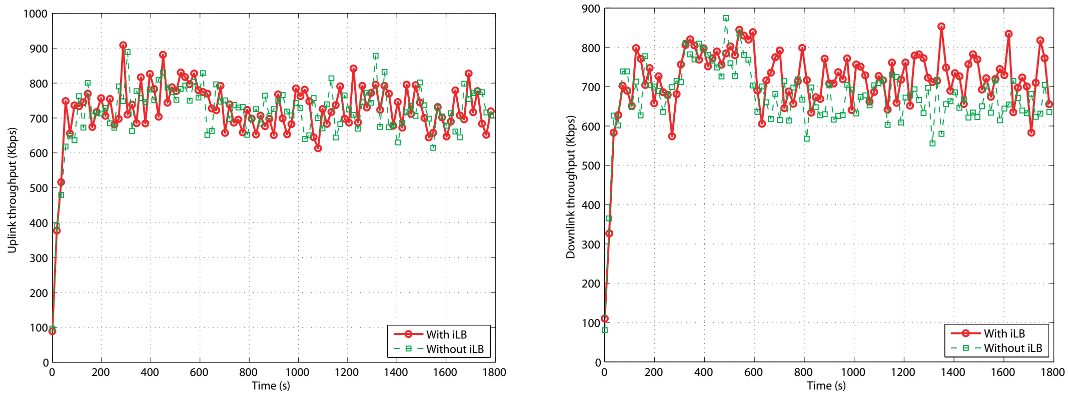
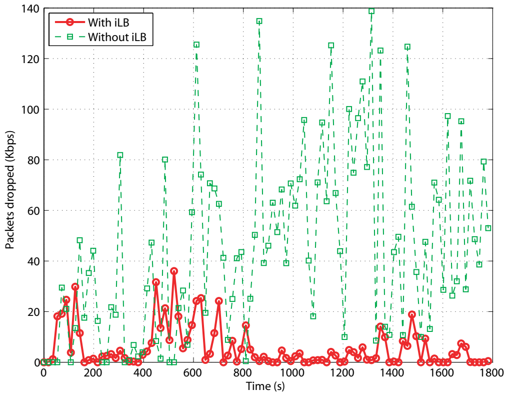


Figure 4.6: Average UL and DL PDs.



(a) Average UL throughput.

(b) Average DL throughput.



(c) Average packet loss.

Figure 4.7: Average throughput and packet loss.

of AP 1 for the case without iLB as shown in Figure 4.6 are not acceptable in terms of call quality for any VoIP connections. Moreover, a comparison of packet loss as shown in Figure 4.7(c) reveals that the total packets dropped from both APs for the case without iLB are eight times higher than for the case with iLB.

Finally, Figure 4.8 illustrates that iLB effectuates the optimal balance of network throughput in contrast to the case without iLB where network throughput is significantly imbalanced. Note that iLB delivers good steady state performance as there is no unnecessary handovers when the throughput between APs is balanced. By virtue of the iLB scheme, an optimal load distribution is attained as a result of soft admission control which is adaptive to dynamic network conditions through the QoS parameters estimation process.

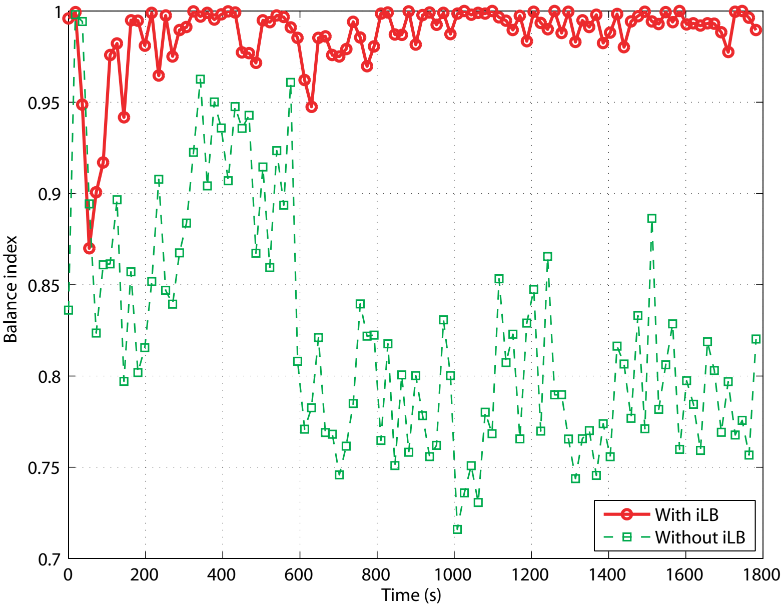


Figure 4.8: Balance index of network throughput.

Apparently, the self-adjusting feature of the soft admission control enables the support of heterogeneous voice codecs with different packetization intervals and packet lengths, which is not possible with the traditional hard admission control technique. As a final remark, the iLB scheme offers balance from a QoS perspective. Particularly, it achieves QoS balance in both delay and throughput, which jointly act to improve overall system utilization.

4.3.2 Multimedia Service Delivery over WLAN

In this study, the QoS performance of the iLB scheme is first evaluated in terms of both PD and PLR. The RQB performance of the iLB scheme is then evaluated in terms of both throughput and QoS fairness. Finally, a comparative analysis on the performance of the iLB scheme with the IEEE 802.11a/b/g DCF and the IEEE 802.11e EDCA, both of which represent the cases without RQB, is conducted. The comparison between iLB and DCF is of interest due to the fact that majority of the existing WLANs are DCF-based which lack service prioritization necessary to support multimedia traffic. The comparison between iLB and EDCA serves to show that although EDCA can support service differentiation, it cannot guarantee strict QoS required by RT services under heavy load without an appropriate QoS balancing scheme. Moreover, the adoption of EDCA by the industry remains uncertain due to the significant cost that will be incurred in replacing the existing IEEE 802.11a/b/g hardwares for additional QoS support.

Evaluation of QoS Performance

To verify the capability of the iLB scheme in providing statistical QoS guarantee for multimedia service delivery over a multi-AP WLAN, the average UL and DL PDs associated with each AP as shown in Figure 4.9 and Figure 4.10, respectively are first examined. It is found that the average DL PD is worse than the average UL PD for iLB, DCF, and EDCA. Particularly, it is noted that the average UL PD of iLB, DCF, and EDCA are well within the acceptable WLAN PD limit of 60 ms in order to meet the one-way end-to-end delay requirement of VoIP packets. Again, this is a direct consequence of the asymmetric traffic load on both links for an infrastructure BSS since the DL becomes the classical capacity bottleneck in the presence of many two-way communications such as VoIP (see, e.g., Figure 6.11 of Section 6.4.1, [82], and [83]). Hence, from this point onwards, the focus is on the average DL PD and PLR which is shown in Figure 4.11 since they are the limiting factors.

In this simulation, AP 1 with multimedia traffic is overloaded during the first 900 s while AP 2 with voice only traffic is overloaded during the last 900 s for both DCF and EDCA as a result of the bursty nature of the offered load. Similarly, the overloading is predominantly due to PHY detection of the existing IEEE 802.11 WLAN handover process which lacks QoS considerations. As a consequence, no handover is triggered since all STAs are within the good coverage region of their APs. On the contrary, VHOs are observed with iLB since it supports link layer detection which allows STA to trigger handover when

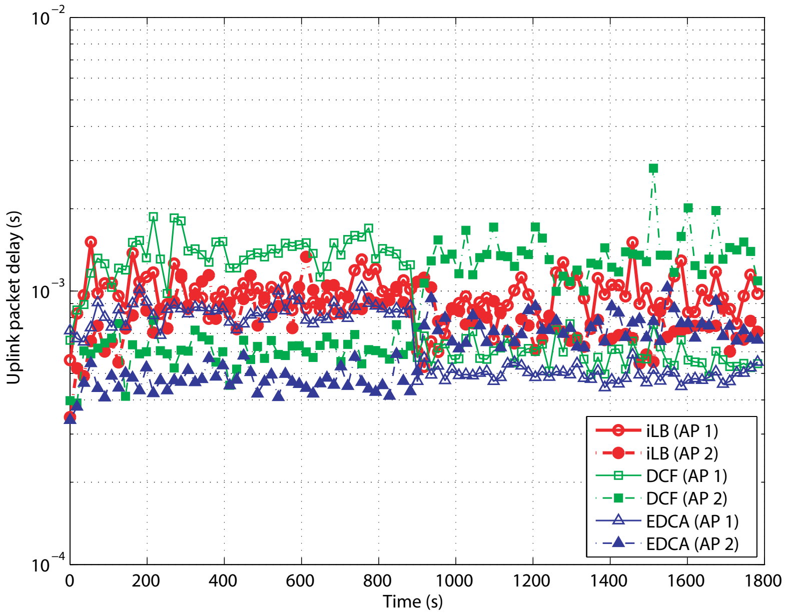


Figure 4.9: Average UL PD.

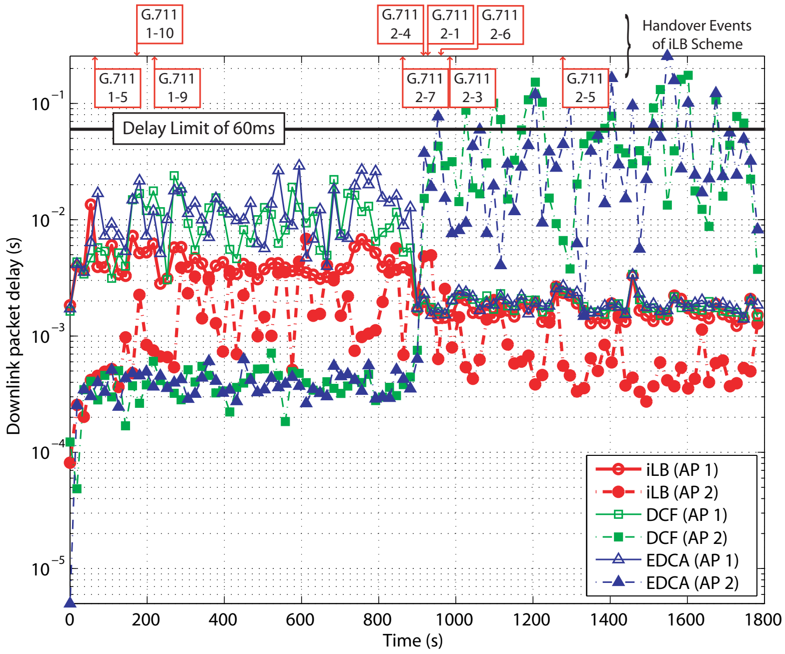


Figure 4.10: Average DL PD.

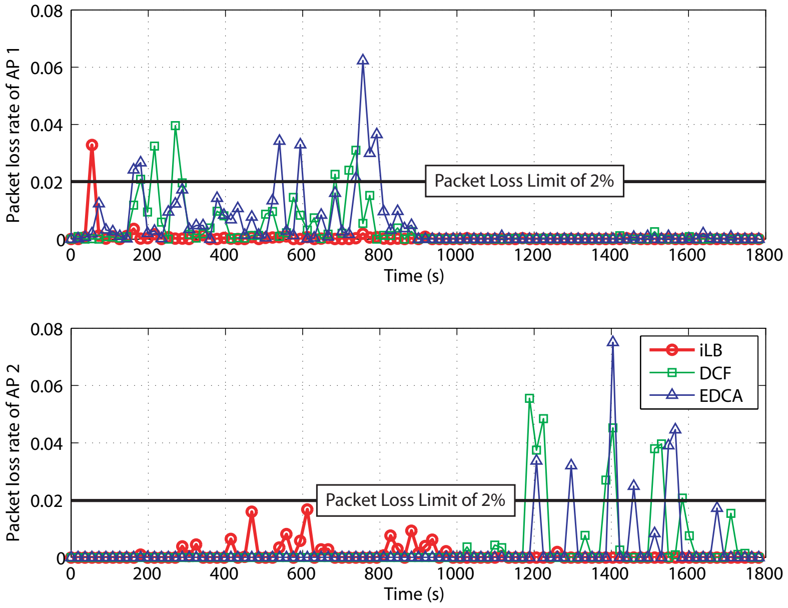


Figure 4.11: Average DL PLR.

the PLR of its associated AP exceeds 2%. This together with the soft admission control effectively mitigate the overloading of both APs. Note that based on the notion of RQB by opportunistic yet altruistic exploitation, a handover will be triggered only on the conditions that: (i) the PLR of source AP is more than 2%; (ii) there exists a target AP which can better meet the delay requirement of VoIP services; and (iii) the handover attempt can be completed only if the target AP can still accept connections when subjected to soft admission control. As such, there will be no additional loss associated with a particular handover when successfully triggered, and the QoS shall be statistically guaranteed after handover since WLAN will operate in the non-saturation mode to protect the QoS of existing connections.

It is evident from Figure 4.10 and Figure 4.11 that both DCF and EDCA are unable to support the strict QoS requirements of RT VoIP services where the PD incurred by WLAN should be less than 60 ms and the PLR should be less than 2%. Accordingly, DCF and EDCA have an average DL PD of up to 170 ms and 250 ms in AP 2, respectively as shown in Figure 4.12. In addition, DCF and EDCA have an average DL PLR of more than 4%

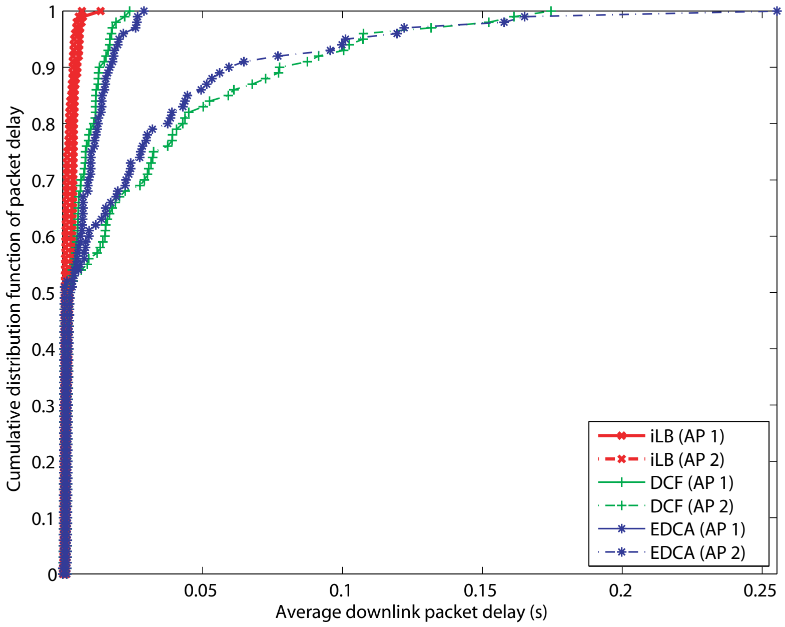


Figure 4.12: CDF of average DL PD.

and 6%, respectively in both APs as shown in Figure 4.13. These observations are due to the buffer overflow phenomenon in both APs which are operating beyond their maximum capacity, and hence they are experiencing excessive PD and consequently packet loss. Although EDCA with QoS prioritization achieves the best UL performance in both APs as shown in Figure 4.14, it has the worst DL performance in terms of average PD and PLR when subjected to heavy load as shown in Figure 4.12 and Figure 4.13, respectively. As a matter of fact, DCF performs *better* than EDCA in AP 2 with voice only traffic, suggesting that the smaller contention window (CW) sizes in EDCA cause increased collisions which have a strong negative impact on the DL performances. It is expected that iLB could effectively mitigate this problem, especially, when EDCA is utilized only for voice traffic of the same priority which reduces to the classical DCF scenario.

With introduction of the iLB scheme, an average DL PD of less than 14 ms together with an average DL PLR of less than 2% are achieved in both APs throughout the simulation as shown in Figure 4.10 and Figure 4.11, respectively. This corroborates the ability of iLB

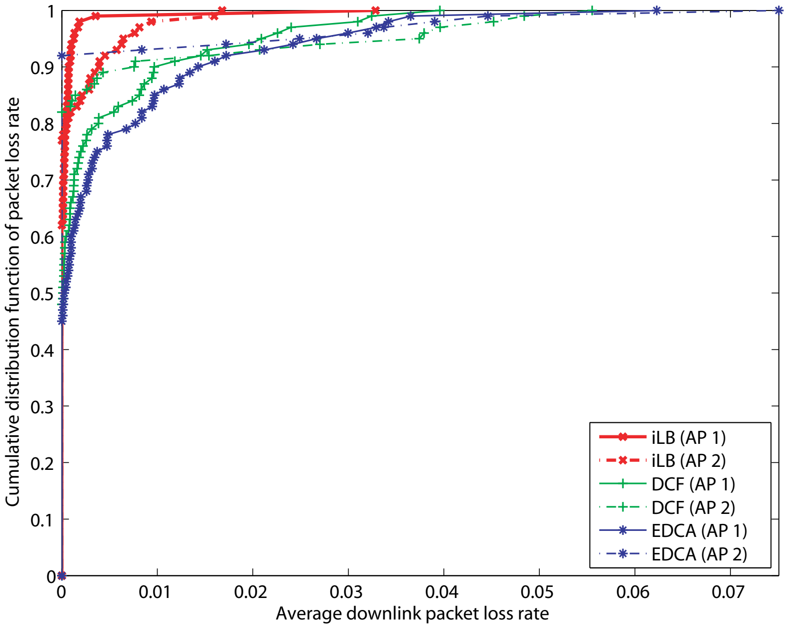


Figure 4.13: CDF of average DL PLR.

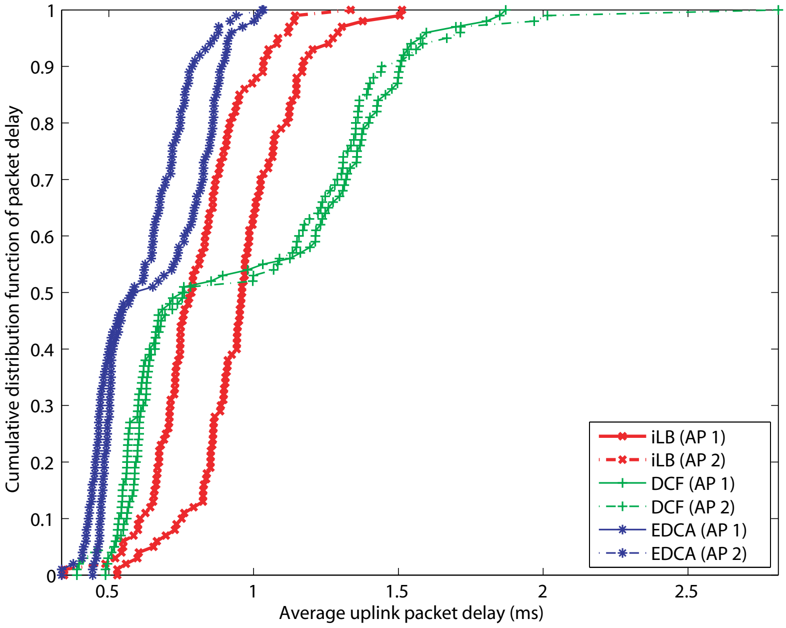


Figure 4.14: CDF of average UL PD.

to support fast handover and the induction that the total Layer 2 handover latency is less than 30 ms. Essentially, this means that QoS shall also be statistically guaranteed during handover. The initial PLR of 3.2% is a result of the link layer detection that triggers handovers in a reactive and opportunistic manner. Clearly, the APs which used to be the capacity bottlenecks are now able to support RT VoIP connections in the presence of multimedia traffic with bounded average PD and PLR.

Evaluation of RQB Performance

To quantify the effect of RQB on the overall system utilization, the definition of (A-9) is adopted to reflect the fairness of throughput and QoS conditions among APs. It is observed that iLB exhibits both throughput fairness as shown in Figure 4.15 and QoS fairness as shown in Figure 4.16, which jointly improve overall system utilization in contrast to DCF and EDCA. Again, note that iLB delivers good steady state performance as there are no unnecessary handovers when QoS between APs are balanced. The balance index of network throughput for DCF and EDCA without RQB is 0.86 which improves to 0.96 with iLB. Similarly, the balance indexes of network delay for DCF and EDCA without RQB are 0.56 and 0.58, respectively which improve to 0.81 with iLB. Similar to the VoWLAN study, an optimal load distribution is attained because the estimated PD metric directly optimizes the expected PD, making it adaptive to dynamic network conditions. This also augments the soft admission control to support multimedia traffic of high variability in a self-adjusting manner, which is not possible with the traditional hard admission control technique. The simulation results also indicate that RQB leads to uniform traffic distribution which in turn maximizes trunking gain by reducing call blocking probability and maintaining lower average delay in the network. In addition, it precludes unnecessary handovers by the reactive and opportunistic handover triggering approach. These advantages could be harnessed by adopting the notion of bi-domain cooperation where the QoS context information of each AP is shared between network entities, i.e.,

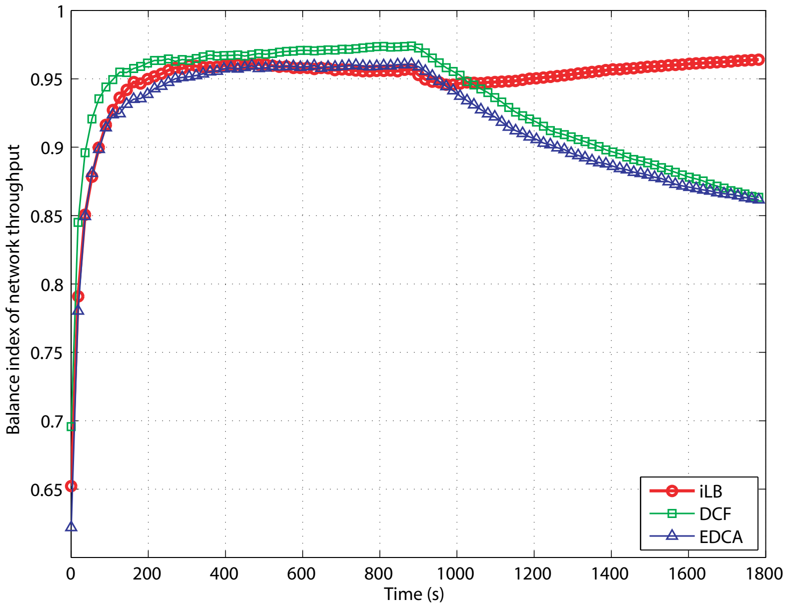


Figure 4.15: Balance index of network throughput.

inter-network cooperation, and between network-terminal entities, i.e., inter-entity cooperation to facilitate joint optimization within the generalized CCRRM architecture.

The effect of RQB on the overall system utilization can also be inferred from the number of retransmission attempts. Accordingly, lower retransmission attempts signify better utilization of radio resources which in turn leave more potential to maximize overall system capacity. From Figure 4.17, it is apparent that iLB has the lowest number of aggregate retransmission attempts. In fact, it attains a 33% and 24% reduction in retransmission attempts as compared to DCF and EDCA, respectively. These reductions in retransmission attempts can be used to transmit useful traffic which essentially boost the effective system capacity. Clearly, iLB can exploit the heterogeneity of a multi-AP WLAN by redistributing voice STAs to a better quality or less loaded AP in an opportunistic yet altruistic manner. This is possible as the generalized CCRRM architecture benefits from the unified actions of joint optimization to promote a QoS-balanced system by enabling multiple APs to interact and network-terminal entities to form synergies. To this end, the simulation results have shown that RQB has intrinsic properties of providing statistical

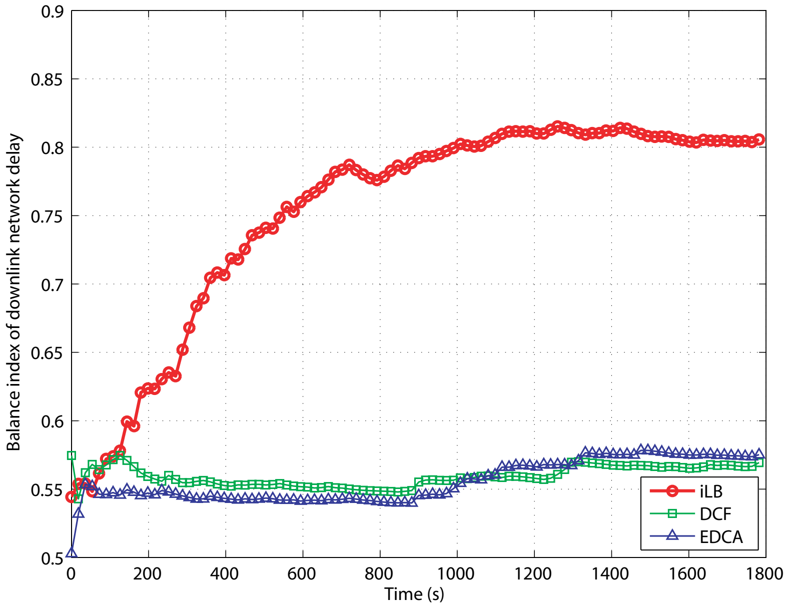


Figure 4.16: Balance index of DL network delay.

QoS guarantee for multimedia traffic, as well as both throughput and QoS fairness which jointly maximize overall system capacity. This reiterates the importance of maintaining a QoS-balanced system in future wireless networks. As a final note, the iLB scheme provides a generic approach to effectuate a QoS-balanced system, irrespective of access network heterogeneity, as shown in the simulation comprising of a mixture of the IEEE 802.11b and IEEE 802.11g APs. This generalization is a direct consequence of the technology agnostic approach as discussed in Section 3.3.2. Therefore, it is clear that the iLB scheme can be fully extended to support VHO in future wireless networks envisaged as an IP-based multi-RAT environment.

4.3.3 Evaluation of System Cost and QoS Broadcast Interval

The importance of iLB scheme in a multi-AP WLAN which is indicative of future wireless networks envisaged as an IP-based multi-RAT environment has been demonstrated. However, any derived benefits come at a cost to the system in terms of both network and

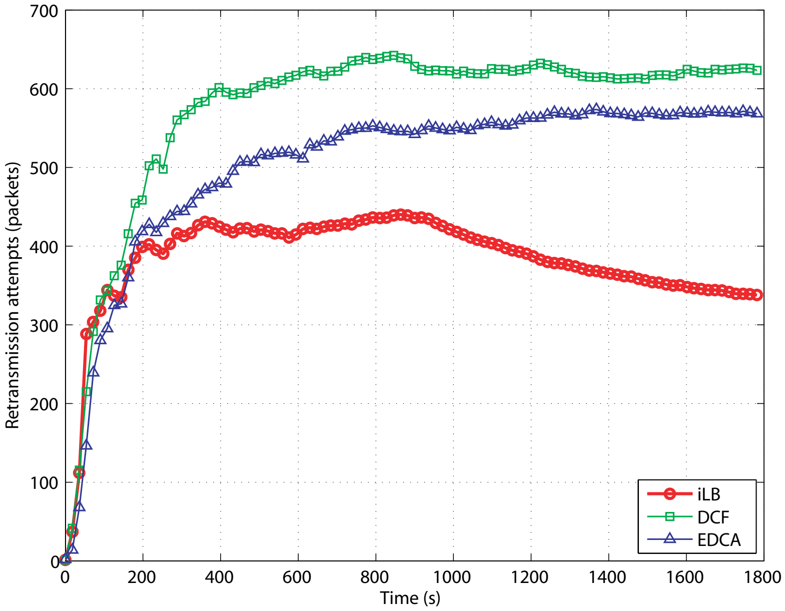


Figure 4.17: Aggregate retransmission attempts.

terminal, which would now be discussed. For network, there will be no additional signaling overheads associated with the broadcast of RRM policy and QoS context information as they are appended to beacons which are periodically broadcasted by an AP to announce its existence. However, there would be some storage and communication overheads for updating these RRM policy and QoS context information. As discussed in Section 3.2.3 (cf. Figure 3.3), the vendor specific information consists of a fixed field and a variable field, which are dependent upon the target group and RRM policy bitmasks. In the event when RQB is required, the cluster measurement reports containing QoS context information of n APs will be restricted only to those within a geo-localized area with cluster-based broadcast. This amounts to a total of 60 octets of network state information[1] per AP, if the cluster comprises of five APs. There would also be some signaling overheads associated with handover events arising from RQB. However, these would be infrequent since the

---

[1] Network state information of 60 octets applies for the case of bi-domain cooperation where QoS context information consists only of PD and PLR. However, network state information will increase to 70 octets for the case of tri-domain cooperation and 80 octets for the case of quad-domain cooperation (cf. Chapter 5) where each additional QoS information requires 2 octets (cf. Figure 3.3).

notion of RQB promotes a QoS-balanced system which will preclude unnecessary handovers due to the reactive and opportunistic handover triggering approach as mentioned in Section 4.2.

Finally, the question of identifying an optimal signaling frequency or QoS broadcast interval of RRM policy and QoS context information also needs to be addressed. To answer this question, the impact of different QoS broadcast intervals on the QoS performance and number of handover events is investigated. From Figure 4.18 and Figure 4.19, it is observed that the QoS performance in terms of both average DL PD and aggregate PLR degrades with increasing QoS broadcast interval. This is not surprising as short-term fluctuations cannot be effectively exploited when the QoS broadcast interval increases. From Figure 4.20, it is noticed that the number of handover events increases with decreasing QoS broadcast interval. It is now obvious that tradeoffs exist between the QoS performance including the number of handover events and the QoS broadcast intervals. Specifically, the QoS performance improves with decreasing QoS broadcast interval at the expense of the increasing number of handover events. Although it may be possible to achieve better QoS performance by reducing the QoS broadcast interval, the storage and communication overheads for updating RRM policy and QoS context information, as well as signaling overheads associated with handover events will bound to increase. Hence, a favorable tradeoff here would be selecting a QoS broadcast interval that gives good QoS performance with a reasonable amount of storage, communication, and handover signaling overheads. Accordingly, the QoS broadcast interval of 1 second is chosen in this thesis as it yields a significantly better QoS performance, without generating more handover events, as compared to QoS broadcast interval of two seconds. In addition, most of the commercially deployed WLANs operate with a default beacon interval of 100 ms. With the generalized CCRRM architecture, the RRM policy and QoS context information are required to be broadcasted only once every ten beacon intervals so that the network is not overwhelmed with storage, communication, and possibly handover signaling overheads. It is worth mentioning that the broadcast of QoS context information within the

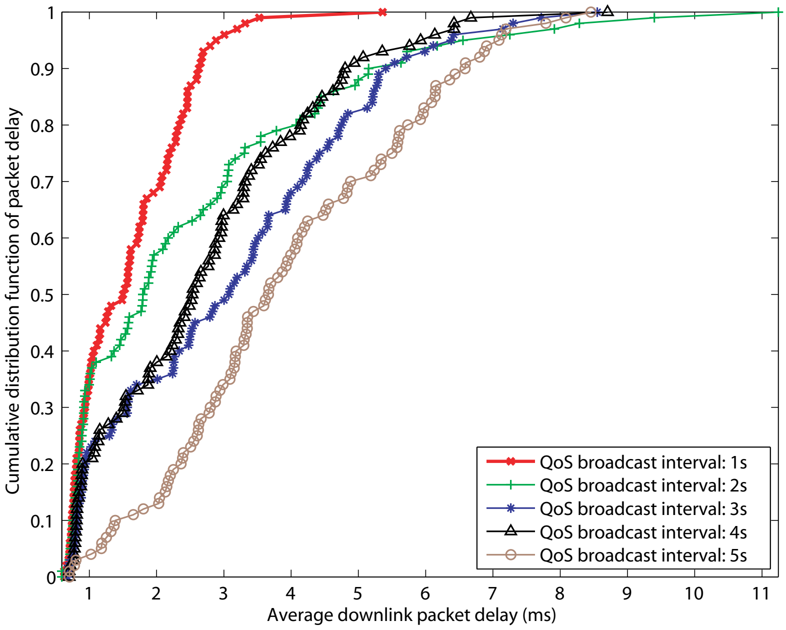


Figure 4.18: CDF of average DL PD.

cluster measurement reports is optional depending on the RRM policy. E.g., cluster measurement reports containing the QoS context information will be broadcasted only during RQB (cf. Table 3.2). When RQB is not required, the network state information reduces only to a total of 10 octets. Hence, the additional network state information required in the generalized CCRRM architecture does not impose heavy loads on the network.

For terminal, the computational complexity which would manifest as power consumption is considered. Although the proposed iLB scheme requires additional computations to perform network selection, it is expected to be minimal since the network selection procedure has a linear time complexity of $O(n)$. Moreover, n will be bounded since cluster-based broadcast is restricted only to APs within a geo-localized area. Furthermore, the exclusion of scanning phase in the fast handover design more than offsets this incremental computational cost.

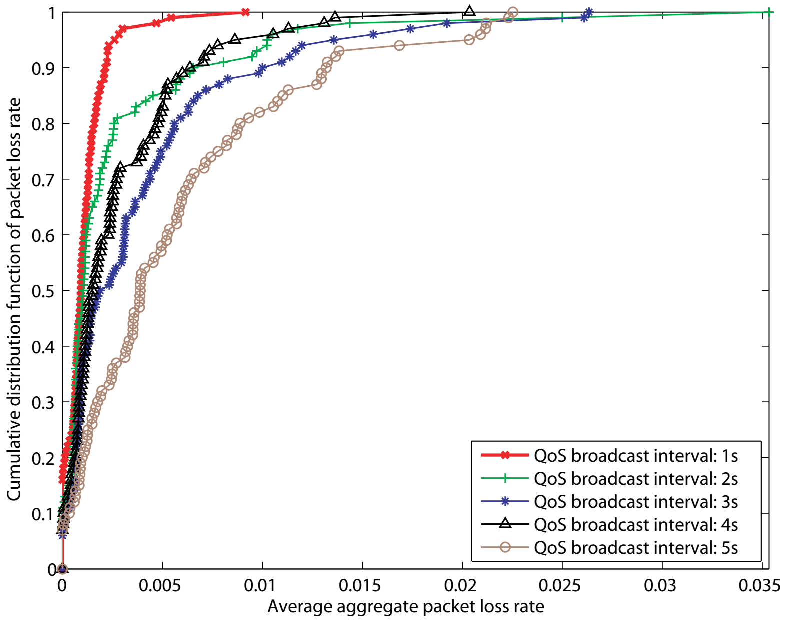


Figure 4.19: CDF of average aggregate PLR.

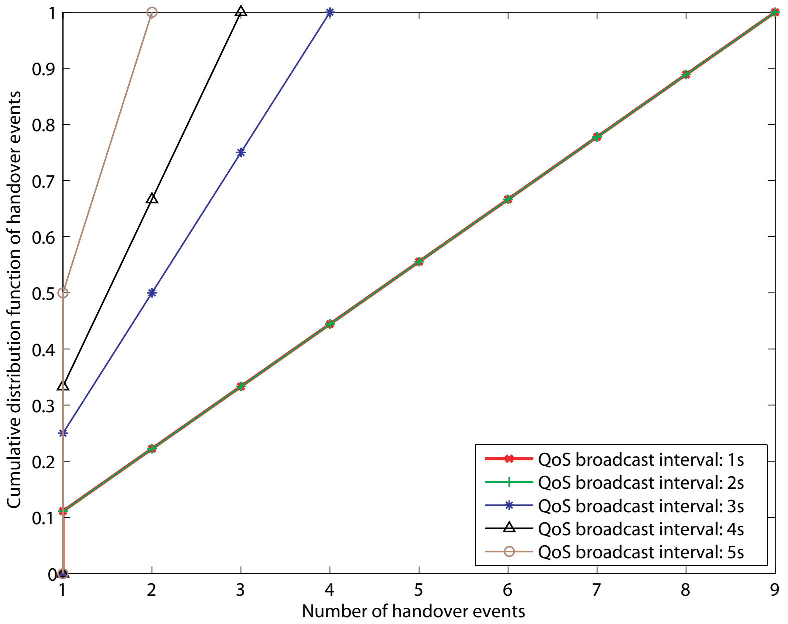


Figure 4.20: CDF of handover events.

4.4 Chapter Summary

This chapter has presented the baseline design of the generalized CCRRM architecture and motivated the importance of cooperation which can exploit heterogeneity as an enabler to improve the overall system capacity and QoS of end-users in a multi-AP WLAN (or future wireless networks). Leveraging on the TONA handover architecture and DANS algorithm proposed in Chapter 3, which provide bi-domain cooperation, an efficient iLB scheme has been devised to offer RQB by incorporating fast handover in conjunction with soft admission control to trigger VHO in an opportunistic yet altruistic manner. The iLB scheme is lightweight and adaptive to dynamic network conditions by using only PD estimates as: (i) the criterion during network selection which eliminates both detection and scanning phases from the WLAN handover process; and (ii) the load metric to devise a robust soft admission control which supports VoIP traffic with heterogeneous voice codecs and multimedia traffic, otherwise not conceivable with hard limiting approaches. It has been shown by induction that the iLB scheme is able to support seamless handover with a total Layer 2 handover latency of 16 ms to 30 ms. Simulation results have further demonstrated that a bounded average DL PD of less than 14 ms and a bounded average DL PLR of less than 2% are achievable. These satisfy the stringent QoS requirements of RT VoIP connections in both VoWLAN and multimedia service delivery over WLAN scenarios (cf. Appendix A-2.1 for details).

The simulation results have also indicated that RQB has intrinsic properties of providing statistical QoS guarantee to enable seamless delivery of both VoIP and multimedia traffic while maximizing overall system capacity. Hence, the notion of employing QoS balance as the criterion to quantify the state of balance in a multi-AP WLAN (or future wireless networks) in which network conditions vary significantly for both mobile and stationary terminals has been advocated. In summary, the iLB scheme offers four main benefits, viz., (i) statistical QoS guarantee during handover with fast handover; (ii) statistical QoS guarantee after handover with soft admission control; (iii) exhibits both throughput and QoS fairness to jointly improve overall system utilization; and (iv) the novel concept of

RQB provides a generic solution to achieve the end-to-end goal of a QoS-balanced system, thanks to the technology agnostic approach of the generalized CCRRM architecture.

Building on the concept of inter-network and inter-entity cooperations, the following chapter extends from bi-domain cooperation to multi-domain cooperation where both QLO and LAS frameworks illustrated in Figure 2.5 will be presented. The evolved, generalized CCRRM architecture will include intra-layer cooperation and inter-layer cooperation to induce synergetic interactions between different functional blocks and layers of protocol stack, respectively. How to leverage on multi-domain cooperation in the evolved generalized CCRRM architecture to exploit all possible heterogeneity in a multi-AP WLAN (future wireless networks) is the focus of the next chapter.

CHAPTER 5

MULTI-DOMAIN COOPERATION

The extensive deployment of the IEEE 802.11 WLAN has positioned itself as the de-facto wireless access network for the ‘last mile’ connections. The potential for WLAN to deliver multimedia contents such as VoIP, video conferencing, video streaming, and data services will become a reality with the advent of the IEEE 802.11n standard [87], promising data rates of up to 600 Mbps. In fact, QoS provisioning at the MAC layer is critical to provide guarantee for QoS requirements of multimedia services [105]. However, QoS provisioning for multimedia traffic delivery in the IEEE 802.11 networks is a non-trivial and challenging task due to the stochastic nature of the random backoff process. The iLB scheme which leverages on bi-domain cooperation to offer RQB has been presented in Chapter 4. In particular, RQB exhibits the salient traits of providing statistical QoS guarantee for both VoIP and multimedia service delivery over WLAN while maximizing overall system utilization as a result of the improvement in throughput and QoS fairness. However, the studies in Chapter 4 are conducted using the legacy IEEE 802.11 with basic access scheme of DCF, which does not support any form of service prioritization, and under ideal channel conditions. Hence, the aim of this chapter is to incorporate service prioritization and link adaptation within the generalized CCRRM architecture to offer a more comprehensive QoS guarantee for delay-sensitive RT services and explore the effects of channel impairments on RQB, respectively.

By building on the concept of bi-domain cooperation, this chapter explores the different domains of cooperation and incrementally extends the generalized CCRRM architecture to support multi-domain cooperation. First, service prioritization is introduced and intra-layer cooperation between different RRM functional blocks within the MAC layer is integrated with bi-domain cooperation to form the QLO framework based on *tri-domain cooperation*. Next, link adaptation is introduced and inter-layer cooperation between the PHY and MAC layer is incorporated with tri-domain cooperation to devise the LAS framework based on *quad-domain cooperation*. Such modular design of the generalized CCRRM architecture enables flexibility in adaptation to different deployment scenarios while treating the QoS provisioning of multimedia service delivery over WLAN from a unified perspective. In other words, the modular design of the generalized CCRRM architecture enables different levels of customization and allows opportunities for dynamic composition of network configurations or policies, which could be delivered on the fly according to dynamic network conditions, in a highly adaptive manner.

This chapter is outlined as follows. Section 5.1 discusses some implementation aspects of service prioritization, in particular, the caveats of EDCA and emphasizes on the need to support service prioritization in the legacy DCF. Section 5.2 inspires tri-domain cooperation and proposes the QLO framework which introduces intra-layer cooperation for improving multimedia service delivery in a single rate WLAN under dynamic network conditions. Section 5.3 presents the performance evaluation of the QLO framework under the effect of network traffic and wireless channel variations. Section 5.4 highlights some of the key challenges of multirate WLAN-based cognitive networks. Particularly, those which will arise from the rate anomaly phenomenon. Section 5.5 motivates quad-domain cooperation and presents the LAS framework which incorporates inter-layer cooperation to exploit the benefits of both link adaptation and load adaptation on-demand for mitigating rate anomaly and enhancing multimedia service delivery in a multirate WLAN-based cognitive network under dynamic network conditions. Section 5.6 discusses the performance evaluation of the LAS framework, and Section 5.7 concludes this chapter with the key findings.

5.1 Service Prioritization: EDCA vs. DCF

A new MAC layer function known as the hybrid coordination function proposed in the IEEE 802.11e standard is an enhancement over the DCF to support differentiated QoS. The EDCA is a contention-based access mechanism provided by the hybrid coordination function to offer QoS station (QSTA) prioritized QoS access to the wireless medium while still supporting best-effort traffic to non-QSTA. The QoS support in the EDCA is realized with the introduction of four different first in first out (FIFO) transmit queues and access categories, corresponding to four different priorities, viz., AC_BK, AC_BE, AC_VI, and AC_VO in the order of increasing priority. These four access categories are mapped to eight user priorities, which may take integer values from 0 to 7, according to the IEEE 802.1D standard. Service prioritization is achieved by directing higher layer data traffic into one of the four transmit queues w.r.t. the UP – AC mappings as illustrated in Figure 5.1. Each of the transmit queues is then processed by an enhanced variant of the DCF known as the enhanced distributed channel access function (EDCAF) with AC-specific parameters known as the EDCA parameter set. Essentially, each EDCAF can be thought of as a virtual DCF contenting with differentiated medium access. On the contrary, the legacy DCF with a single FIFO transmit queue can provide only medium access with the same priority resulting in best-effort delivery, regardless of traffic types, rendering it ineffective to support different QoS requirements.

The EDCA parameter set defines the priority differentiation in channel access by varying three key parameters: (i) CW backoff parameters (CW_{min} and CW_{max}) which influence the average time required to successfully deliver a packet; (ii) arbitration interframe space (AIFS) which determines the duration of time a QSTA needs to perform carrier sensing, i.e., defer access following a busy medium before initiating backoff; and (iii) transmission opportunity (TXOP) limit which specifies the duration of time a QSTA may transmit after it acquires a TXOP. Note that the QSTA may transmit multiple frames within its TXOP allocation. Although this feature could provide temporal fairness between QSTAs to alleviate the rate anomaly problem of DCF [106], there are limitations

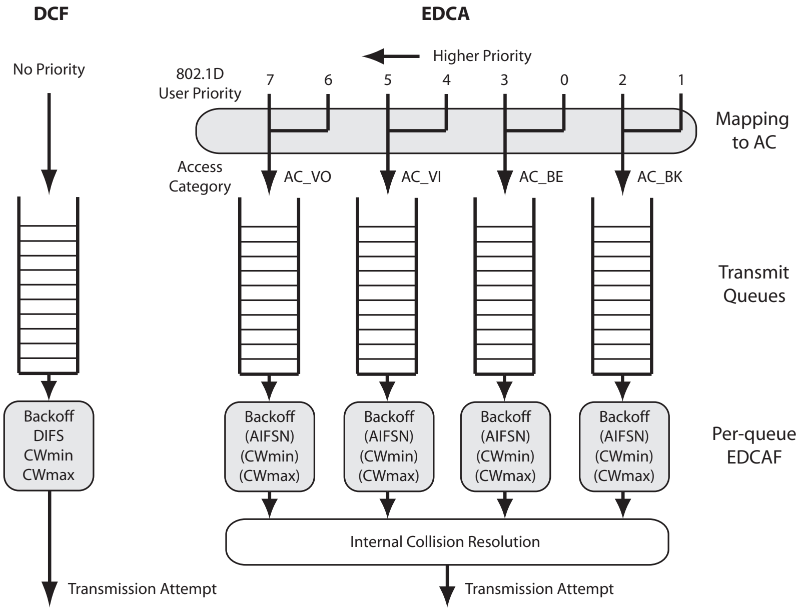


Figure 5.1: Comparison of the IEEE 802.11 DCF and IEEE 802.11e EDCA.

in practice (cf. Section 5.5). In fact, the default values of EDCA parameter set recommended by the standard and many related works, e.g., [107] and [108] are heuristically derived and do not guarantee optimized performance. Moreover, these static values are not adaptive to dynamic network conditions. Hence, the question of how to achieve an optimal EDCA parameter set for a given network configuration and a set of QoS constraints under varying network conditions still remains as an open research problem [105], [107]. The complexity in configuring the open EDCA parameter set optimally can be best appreciated by understanding the underlying principle of each tunable parameter and its influences on service prioritization. Accordingly, service prioritization can be achieved in one or a combination of the three ways, viz., CW_{min} differentiation, AIFS differentiation, or TXOP differentiation. The basic idea is that the AC with a smaller CW_{min} and arbitration interframe space number (AIFSN) corresponds to higher priority and has a better chance to access the wireless medium earlier while a longer TXOP limit enables the AC to seize the wireless medium for a longer period of time. However, there are caveats associated with each differentiation mechanism.

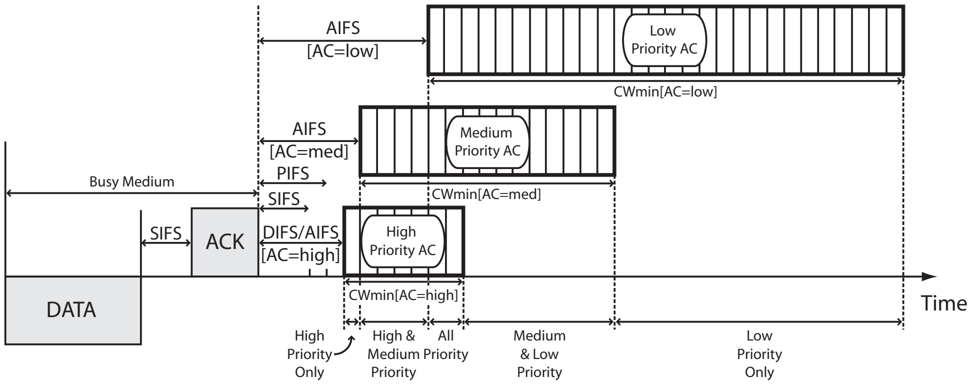
Bianchi *et al.* [109] show that CW_{min} differentiation would lead to aggregate throughput degradation in both low and high priority STAs as the reduction of initial CW size will increase the probability of collision in the medium, and thus it reduces the effectiveness of the random access mechanism. In fact, they demonstrate that AIFS differentiation is superior to CW_{min} differentiation as the former reserves channel slots for high priority STAs exclusively without modifying the random backoff process. However, the authors also point out that the number of reserved channel slots will bound to increase, which implies that the number of idle slots between two consecutive transmissions will decrease, under heavy load. In fact, Engelstad *et al.* [110] highlight that AIFS differentiation would lead to the starvation of lower priority STAs under heavy load. This is because lower priority STAs, with a higher AIFSN, will be denied from accessing the medium since the probability that their AIFSN is larger than the number of idle slots between two consecutive transmissions will increase significantly under heavy load. Hence, they do not have a chance to decrement their backoff counter and eventually drop their packets. Nafaa [105] cautions that although TXOP limit differentiates throughput among the access categories such that higher priority STAs occupy the medium for a longer duration to achieve higher throughput, it should not be purely based on throughput considerations as this can lead to the QoS degradation of low priority STAs with increased delay.

Above all, the key problem of EDCA is that it cannot guarantee strict QoS prioritization between the access categories due to the overlapping of their initial CWs, as depicted in Figure 5.2(a), based on the recommended values of AIFS and CW_{min} in the standard. In other words, the EDCA provides only relative (statistical) prioritization to high priority traffic where prioritized access is guaranteed in the long term but not for every contention. Another possible cause for loose QoS prioritization is attributed to the fact that each STA will freeze its backoff counter only when the medium is busy, but it will continue to count down its backoff timer whenever the channel becomes idle. Thus, it might happen that a lower priority packet that arrives earlier has the least number of remaining backoff slots as compared to a newly arrived but higher priority packet. Consequently, the lower priority packet will seize the medium over the higher priority packet. This problem will be exacerbated during high traffic load or increased contention between the EDCAFs of the same AC. Particularly, a negative impact exists for the higher priority access categories as they will experience more collisions, which lead to the doubling of their CWs upon transmission failures, owing to their smaller initial CW sizes. Moreover, the higher priority AC may operate using a larger CW while the lower priority AC operates with the minimum CW size CW_{min} after a successful transmission as the number of consecutive collisions in each STA may be different, i.e., not all the EDCAFs increase their CWs at the same time. This priority reversal phenomenon [111] will create ambiguity in fairness between packets of the same AC among STAs [112], and it will affect any schemes that employ CW_{min} differentiation to support service prioritization as soon as the CWs of different access categories overlap either initially or upon collisions.

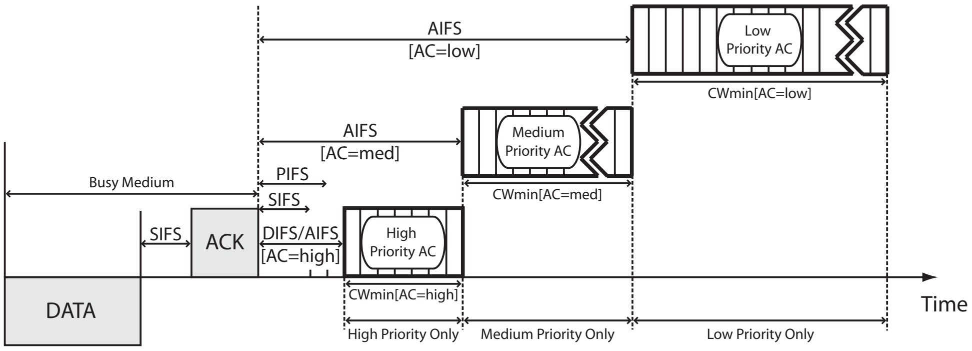
The impact of the priority reversal phenomenon is detrimental to high priority RT services such as VoIP. To overcome this problem, Lee *et al.* [108] propose to enforce strict prioritization by imposing the relation

$$AIFS\left[AC_i\right] \geq AIFS\left[AC_j\right] + CW_{\max}\left[AC_j\right], \quad j > i \tag{5.1}$$

where i and j denote the priorities of the access categories. This strict prioritization as illustrated in Figure 5.2(b) will be effective against priority reversal phenomenon as the CWs between the different access categories are guaranteed to be non-overlapping. However, this scheme leads to poor bandwidth utilization, particularly, when a large value of CW_{max} for higher priority STAs is chosen. Hence, lower priority STAs have to defer their access to the medium unnecessarily on occasions when higher priority STAs have no packets to transmit. On the other hand, a small value of CW_{max} will result in higher contention between high priority STAs [111]. To mitigate these undesirable effects of such strict prioritization, Wang *et al.* [113] enforce strict prioritization by employing the



(a) Relative (statistical) prioritization: The priority reversal phenomenon.



(b) Strict prioritization.

Figure 5.2: Service prioritization of the IEEE 802.11e EDCA.

concept of black-burst contention. However, this scheme requires modifications and is not compatible with the original EDCA mechanism as specified by the standard.

Although the upcoming IEEE 802.11e standard has been proposed to deal with some shortcomings of the legacy DCF, the studies in Chapter 4 and numerous previous works, e.g., [94] and [114] have revealed that the EDCA cannot support strict QoS guarantee under heavy load without an appropriate network control mechanism such as QoS balancing scheme or admission control. Furthermore, the adoption of EDCA by the industry remains uncertain due to significant capital expenditure that will be incurred in replacing the existing IEEE 802.11a/b/g WLAN hardwares for additional QoS support. As a result, there are various works which are devoted to provide service prioritization in the legacy DCF.

Deng and Yen [115] provide service prioritization by differentiating CW sizes. Although the initial CW sizes are designed to be non-overlapping, such scheme is still vulnerable to the priority reversal phenomenon upon collisions as discussed earlier. Nyandoro *et al.* [116] propose to support service prioritization based on the capture effect by allowing STAs to transmit using one of the two different power levels according to their service classes. However, this scheme requires high priority STAs to transmit with a higher power. Hence, the impact of the hidden terminal problem, inter-cell interferences, and association procedures need to be addressed. Yu and Choi [117] implement a modified dual queue strategy at the AP and employ a simple scheduling policy known as strict priority queueing to serve RT queue, as long as it is non-empty, in favor of NRT queue. By far, this is the simplest and most intuitive way to provide service prioritization, which offers performance comparable to the EDCA and requires only a software upgrade in the device driver level, making it an attractive alternative to the EDCA. Moreover, the study in [112] clearly shows that priority-based mechanisms in the EDCA cannot guarantee fairness between different traffic classes, and fair queueing scheme such as the one proposed in [117] is required.

Given that the focus of this chapter is to investigate how multi-domain cooperation can be exploited to provision QoS for delivering multimedia service over WLAN, the legacy DCF with strict priority queueing is preferred over the EDCA as the legacy DCF, other than its prevalence, serves as a better benchmark to study any performance gains.

5.2 Tri-Domain Cooperation

The IEEE 802.11 WLANs are already pervasive in many diverse environments such as enterprises, universities, hotels, and public hotspots due to their low cost deployment. The forthcoming IEEE 802.11n standard [87] offering data rates of up to 600 Mbps will accentuate their benefits for high-speed ubiquitous broadband wireless access. However, as discussed in Section 4.1, the delivery of QoS-demanding multimedia applications such as VoIP and video over IP in the presence of data services over WLAN is very challenging. In addition, WLAN requires service differentiation which is lacking in the basic access scheme of DCF that accounts for a majority of the currently deployed WLANs. A possible alternative is implementing a modified dual queue strategy [117] to provide service prioritization in the existing DCF-based APs as explained in Section 5.1. Moreover, WLAN requires knowledge about the radio environment in order to improve performance and provide guarantee to multiple QoS requirements of multimedia applications. E.g., measurements such as the load information of neighboring APs are important for load balancing function, and the average MAC delay of APs could be used for selecting an optimal AP that meets the QoS requirements of end-users while preventing unnecessary handovers as shown in Section 3.5.3. More importantly, channel impairments, which are apparent in hotspot deployments and indoor environments, due to frequent non-line-of-sight (NLOS) transmissions caused by structures and obstacles should be considered together with the network load, achievable network QoS performance, and desired QoS requirements of end-users when performing load distribution.

5.2.1 Intra-Layer Cooperation

As a matter of fact, many previous works, e.g., [99] and [101] have considered load balancing and admission control in the context of ideal, similar wireless channel conditions where a single load metric suffices. However, it is argued in this thesis that using a single load metric such as CU under dissimilar wireless channel conditions has catastrophic effect. In this section, a holistic approach to provision QoS in the presence of the multifaceted challenges, as discussed in Sections 4.1 and 5.2, is proposed to redistribute load across a multi-AP WLAN opportunistically. The design of the generalized CCRRM architecture is based on the two key requirements of seamless mobility and QoS transparency support in order to meet the ‘anywhere and anytime’ concept. These requirements place very strict constraints over the efficient use of radio resources, particularly, in the presence of increasing demands for multimedia service delivery over wireless networks, and thus more stringent QoS requirements. To address these concerns, the distributed QLO framework that supports tri-domain cooperation is proposed as illustrated in Figure 5.3 by extending from the iLB scheme. It leverages on an additional intra-layer cooperation within the MAC layer to optimize load distribution according to: (i) different services using the network-QoS entity; and (ii) dynamic network conditions through the connection-QoS entity. E.g., load control performs RQB by triggering VHO to the optimal target AP based on network selection outcome. Finally, a handover attempt can be completed only if the target AP can still accept connections when subjected to admission control.

The basic idea is to exploit heterogeneity[1] within such multi-AP WLAN, through the cooperative exchange of QoS context information and opportunistic network selection, i.e., bi-domain cooperation, to promote a QoS-balanced system by optimizing load distribution in a self-adjusting manner. To the best of the author’s knowledge, there is no prior work on load distribution scheme that offers statistical QoS guarantee and optimizes system capacity from a unified perspective, under network heterogeneity and dissimilar

---

[1] The context of heterogeneity here refers to radically different data rates of mixed IEEE 802.11b/g WLANs, traffic variations from multimedia flows, and diverse channel conditions prevalent in hotspots and indoor propagation environments.

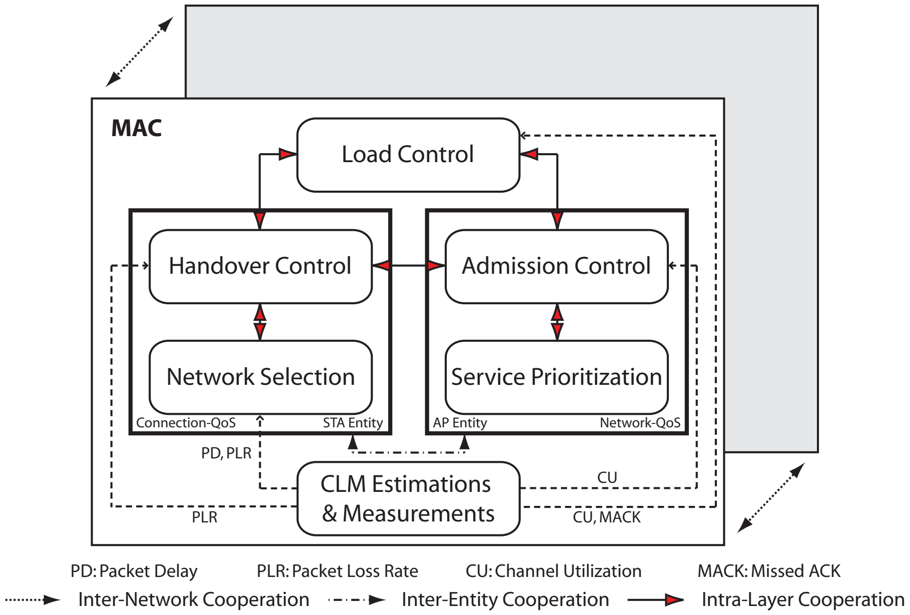


Figure 5.3: Tri-domain cooperation: The distributed QLO framework.

wireless channel conditions. The main contributions differ from the related works in three significant ways: (i) the QLO framework gives a unifying treatment in QoS provisioning for multimedia traffic delivery over the widely deployed DCF-based single rate WLANs; (ii) the cooperative load metric (CLM) consisting of PD, PLR, CU, and missed acknowledgment is exploited for adaptation to dynamic network conditions which include traffic and wireless channel variations; and (iii) the notion of RQB, which has intrinsic properties of providing statistical QoS guarantee while maximizing overall system capacity, is validated as a suitable criterion to quantify the state of balance in a heterogeneous multi-AP WLAN.

5.2.2 QoS-Inspired Load Optimization Framework

The fundamental of the QLO framework first proposed in [118] is based on network-assisted discovery over the TONA handover architecture as presented in Section 3.2. By listening to the QoS broadcast (cf. Section 3.2.3), STAs will acquire the global QoS context information of PD, PLR, and CU. An additional missed acknowledgment (MACK) metric, a local information to an AP, is necessary for the load control function of the AP to distinguish between traffic and wireless channel variations. The idea of MACK is similar to automatic rate fallback employed in Lucent WaveLAN-II [119]. The MACK together with the QoS context information form the CLM to serve as inputs for network selection, handover control, admission control, and load control to optimize load distribution across a heterogeneous multi-AP WLAN opportunistically in a distributed and self-adjusting manner. Note that in Section 4.2.1, soft admission control is based only on a single PD metric. However, CU is employed here instead of PD as a consequence of implementing service prioritization over the DCF. It follows that the PD of RT packets will be relatively lower than NRT packets after service prioritization. This implies that the average PD of an AP will also be low when RT traffic dominates. Under such conditions, the PD metric could lose its effectiveness by the over admission of RT flows, which eventually leads to the starvation of NRT flows.

The algorithm of the proposed QLO framework depicted in Figure 5.4 aims to redistribute load by nominating candidate STAs for handover to a better quality or less loaded AP. The load distribution is performed by leveraging on intra-layer cooperation between the network-QoS and connection-QoS entities, in addition to bi-domain cooperation. The network-QoS entity consists of both service prioritization and admission control to deal with the different service profiles of end-users. Multimedia traffic can be classified into RT or NRT according to its delay requirements. RT traffic such as VoIP and video conferencing is delay-sensitive whereas NRT traffic, also known as elastic traffic, can tolerate a relatively longer delay. Hence, it is important to introduce service prioritization in the DCF-based WLAN so that RT traffic can be handled with higher priority than NRT traffic in order to support a more comprehensive, differentiated QoS guarantee of multimedia traffic. For this purpose, the modified dual queue strategy in [117] is adopted with a slight modification to introduce an additional granularity of service prioritization to support voice, video, and data flows. The admission control regulates the network load by operating under the admission threshold, typically set below the saturation point of

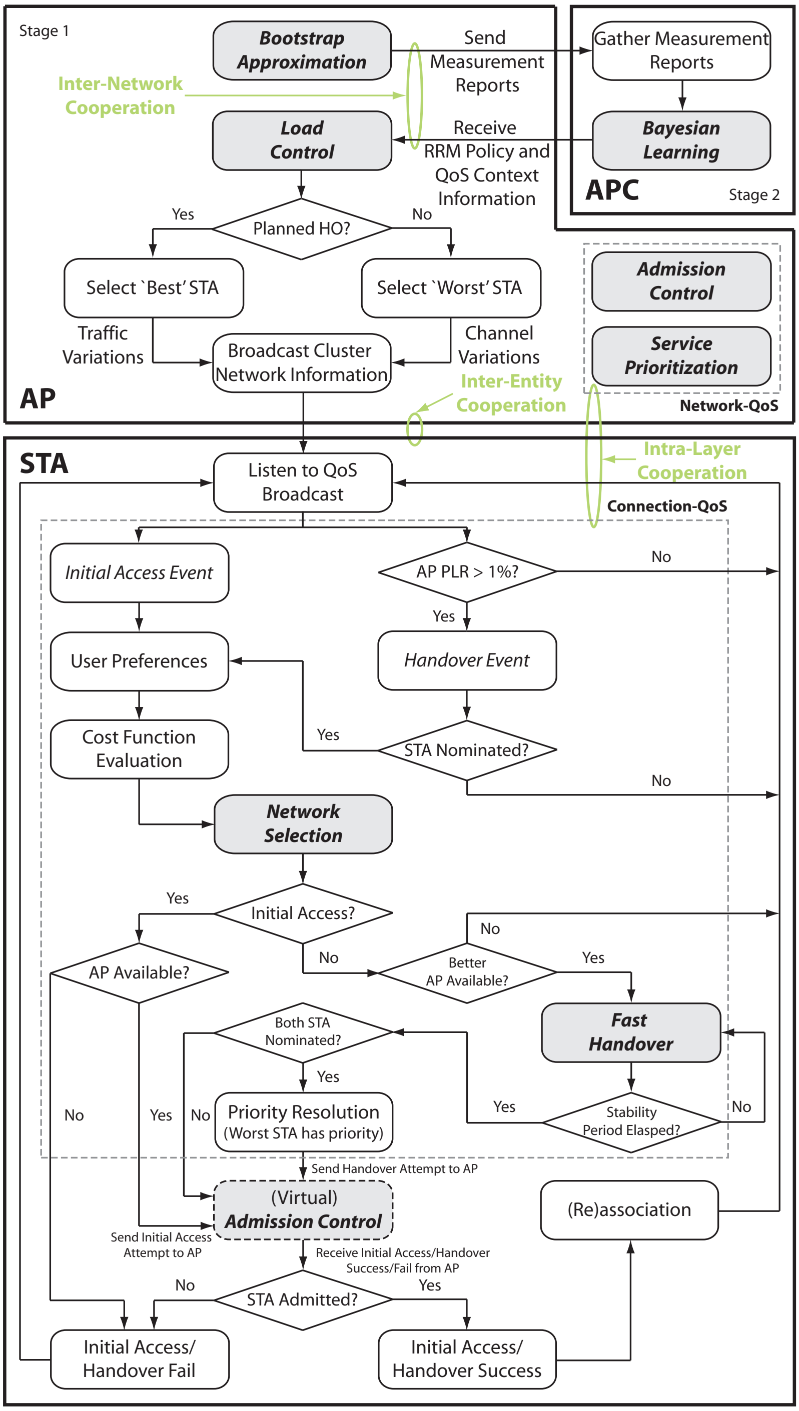


Figure 5.4: Algorithm of the QLO framework based on tri-domain cooperation.

WLAN, in order to protect the QoS of existing flows. In fact, admission control and load control are often not dissociable. The main reason is that both rely on the knowledge of the load metric in order to make their decision. As mentioned in Section 4.1, the CU estimation technique in [96] is chosen as the load metric for both admission control and load control due to its simplicity and high accuracy in estimating effective network load. Accordingly, the CU of each flow and the corresponding network capacity are estimated at every beacon interval as

$$CU_{total}^{n} = \sum_{k \in Flows} CU_{k}^{n}, \quad n \in APs, \tag{5.2a}$$

$$CU_{j}^{n} + CU_{total}^{n} < CU_{max} \tag{5.2b}$$

where $0 \leq CU_{total}^{n} \leq 1$ is the total CU of nth AP, CU_{j}^{n} is the CU of jth flow, and CU_{max} is the admission threshold. A new flow can be accepted without affecting the QoS of existing flows if (5.2b) holds.

The CU is defined as the fraction of channel occupation time required for successful transmissions per observation interval. Therefore, the CU corresponding to the bandwidth requirement of each flow is simply the product of its packet arrival rate and the average time required to transmit a packet successfully given as

$$CU_{j}^{n} = \frac{R_j}{L_j} \times T_{s_j}^{n}, \tag{5.3a}$$

$$T_{s_j}^{n} = T_{DIFS} + T_{BO}^{n} + 2T_{PHY} + T_{DATA_j}^{n} + T_{SIFS} + T_{ACK} + 2\delta, \tag{5.3b}$$

$$T_{BO}^{n} = \frac{CW_{min}}{2} \times T_{SLOT} \tag{5.3c}$$

where R_j is the average data rate, L_j is the packet length of each flow, $T_{s_j}^{n}$ is the average time required for a successful packet transmission as depicted in Figure 5.5, δ is the propagation delay, and T_{BO}^{n} is the average backoff time as in [120]. T_{DIFS}, T_{SIFS}, T_{SLOT}, and CW_{min} are defined in the IEEE 802.11 standard. Note that the CU accounts

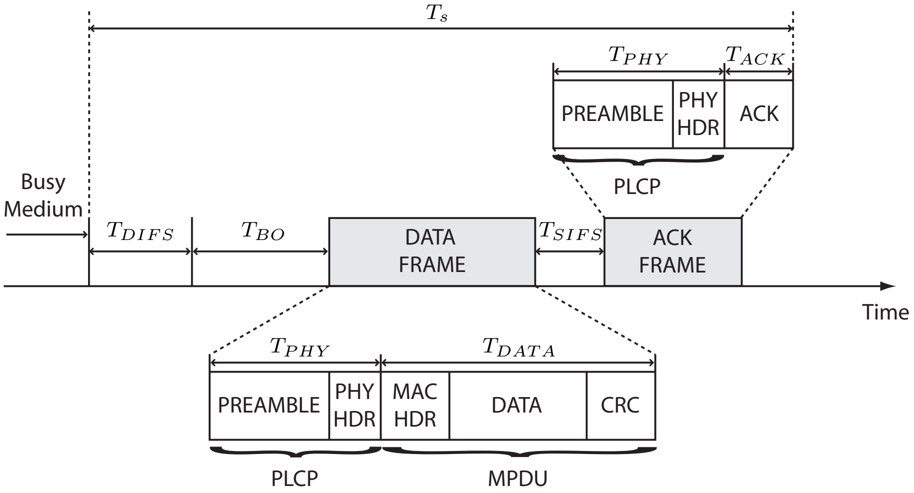


Figure 5.5: Successful transmission time for the IEEE 802.11 basic access scheme of DCF.

only for channel occupation time during successful transmissions while channel busyness ratio proposed in [93] accounts for channel occupation time during both successful transmissions and collisions. However, the work of [93] has shown that the CU is almost the same as the channel busyness ratio when WLAN is operating in the non-saturation mode as the probability of collision is very small. Hence, the channel occupation time during collisions can be disregarded when admission control is in place.

On the other hand, the connection-QoS entity consists of both network selection and handover control to deal with dynamic network conditions associated with channel impairments and network congestions. The PD and PLR, which are the critical QoS information for multimedia traffic, are utilized as metrics for network selection developed in Section 3.4 to reliably select an optimal AP according to the QoS requirements of different traffic classes. Fast handover as described in Section 4.2.1 is incorporated to support RT services by eliminating both detection and scanning delay as the information of an optimal target AP is known.

Similar to the iLB scheme, the design philosophy of the QLO framework is based on the key principle of RQB (cf. Section 4.2) to exploit heterogeneity within a multi-AP

WLAN in an opportunistic yet altruistic manner. To be more specific, imminent handover is detected when the PLR of the source AP, which is the capacity bottleneck of an infrastructure BSS in the presence of many two-way VoIP connections (see, e.g., Figure 6.11 of Section 6.4.1, [82], and [83]), exceeds 1%. As previously mentioned, the AP leverages on an additional MACK metric to distinguish between traffic and wireless channel variations. Specifically, consecutive MACKs signify bad wireless channel conditions while high PD denotes serious QoS degradation due to abrupt traffic variations. Although MACK could also occur due to collisions, the probability of collisions will be small when WLAN is non-saturated [97] as admission control can accurately regulate input traffic with the CU. Hence, when bad wireless channel condition is detected, the source AP would nominate a 'worst' STA as

$$\begin{aligned} \max \quad & MACK_i^n \\ s.t. \quad & \{i \in STAs : i = \min\left(r_i \in R_{DATA}\right)\} \end{aligned} \tag{5.4}$$

where $n \in APs$ and R_{DATA} is the PHY data rate. For the case of multirate WLAN studied in Section 5.5.2, the constraint ensures the STA that is receiving data frames with the lowest R_{DATA} and has the highest MACK[2] is transfered to a better quality AP first. For the case of single rate WLAN considered in this section, the 'worst' STA is simply selected as the one with the highest MACK. Otherwise, the source AP would nominate a 'best' STA as

$$\begin{aligned} \min \quad & CU_{ave} - CU_{total}^m - CU_i^n \\ s.t. \quad & CU_{ave} - CU_{total}^m - CU_i^n \geq 0, \quad m \neq n \end{aligned} \tag{5.5}$$

where $i \in STAs$, $m, n \in APs$, and CU_{ave} is the average CU of all APs. CU_i^n is the per-STA CU which includes both UL and DL flows owing to the fact that all flows in an infrastructure BSS are relayed via the AP. The constraint prevents load distribution to an

---

[2] Also, note that negative acknowledgment (NACK) is employed as an enhancement over MACK for the case of multirate WLAN.

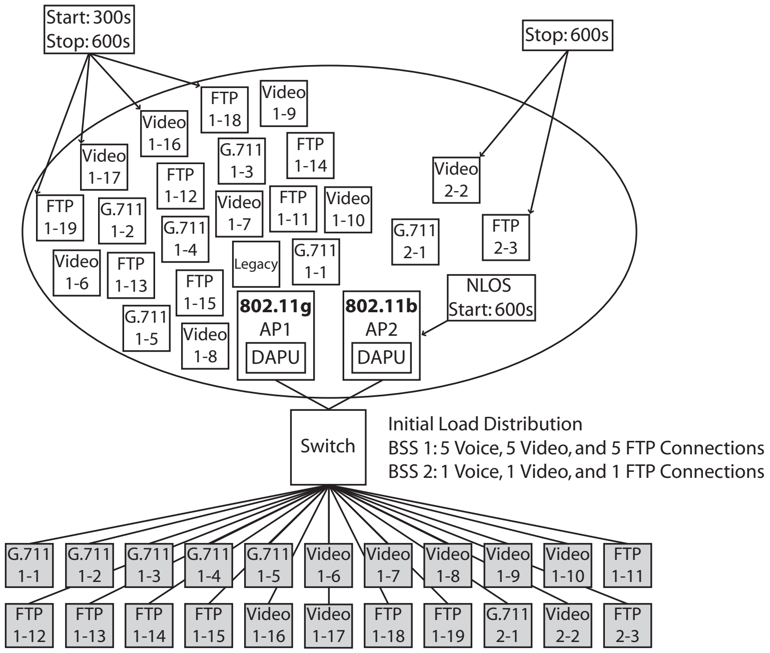
overloaded AP and overloading the target AP. Accordingly, the nominated STA will then perform handover to the optimal target AP.

5.3 Performance Evaluation of the QLO Framework

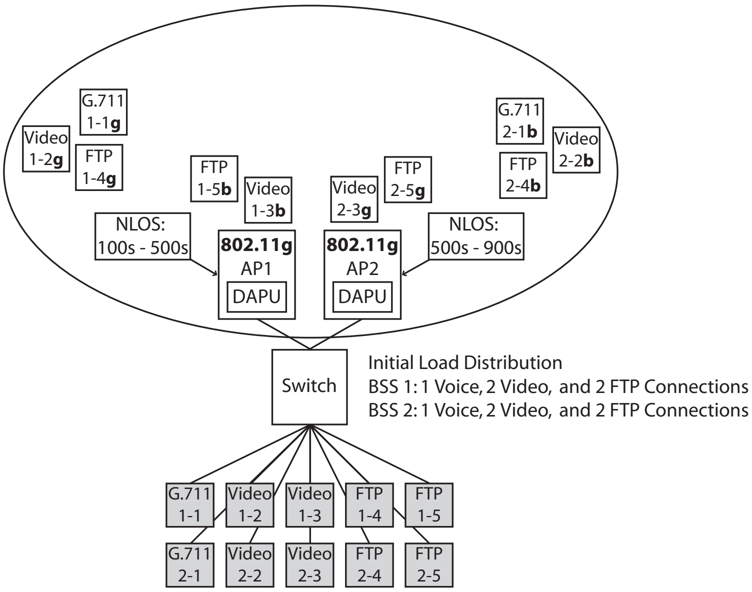
To evaluate the performance and effectiveness of the QLO framework, built on the basis of tri-domain cooperation, two separate simulation scenarios are considered. The simulation models are developed by using OPNET™ Modeler® 14.0 with Wireless Module. Modifications to the existing DCF models are performed in order to provide QoS support with service prioritization, admission control, network selection, fast handover, and load control mechanisms which are the focus of this study. The physical layer model is also enhanced to include shadow fading and the capability to simulate NLOS transmissions by varying the path loss exponent of the log-normal path loss model as described in Appendix A-2.2. The details of general simulation models can also be found in Appendix A-2.

The first scenario simulates a typical hotspot which consists of a heterogeneous multi-AP WLAN with one IEEE 802.11b AP and one IEEE 802.11g AP operating at the maximum data rates of 11 Mbps and 54 Mbps, respectively as shown in Figure 5.6(a). In this simulation, an unbalanced load of five G.711, five video, and five FTP STAs in BSS 1 while one G.711, one video, and one FTP STAs in BSS 2 is initially introduced. At time 300 s, two video and two FTP connections from BSS 1 are started. At time 600 s, these two video and two FTP connections are stopped whilst one video and one FTP connections from BSS 2 are also stopped. These discrete events generate traffic variations during the first $600s$. Furthermore, wireless channel variations are introduced in BSS 2 at time 600 s by simulating NLOS transmissions in practice.

For the second scenario, a typical hotspot, which consists of a homogeneous multi-AP WLAN with two IEEE 802.11g APs operating at the maximum data rate of 54 Mbps and heterogeneous IEEE 802.11b/g STAs, is simulated as illustrated in Figure 5.6(b).



(a) Scenario 1: Heterogeneous multi-AP WLAN with the IEEE 802.11b/g APs.



(b) Scenario 2: Homogeneous multi-AP WLAN with the IEEE 802.11g APs and heterogeneous IEEE 802.11b/g STAs.

Figure 5.6: Simulation models.

Table 5.1: Traffic generation parameters.

Traffic Type	Packet Size (Bytes)	Inter-arrival (ms)	Avg. Data Rate (Kbps)
Voice-CBR (G.711)	80	10	64
Video-CBR	1000	125	64
Data-FTP (UL)	450	120	30
Data-FTP (DL)	1500	40	300

The wireless channel variations are simulated by introducing NLOS transmission in an alternate fashion across both BSSs over a duration of 400 s each, starting from 100 s in BSS 1. No traffic variations are simulated, and therefore a similar load across both BSSs is used. Shadow fading is included for the entire simulation duration for both scenarios and the multimedia traffic sources are summarized in Table 5.1. The MSDU lifetime limit mechanism is incorporated to discard MSDUs from the transmitter queue if they exceed the MSDU lifetime before successful transmission.

5.3.1 Effect of Network Traffic and Wireless Channel Variations

The simulation results presented in both scenarios include the QoS performance of the QLO framework evaluated in terms of the QoS satisfaction factor (QSF), throughput fairness index (TFI), and QBI as defined in Appendix A-3.1. A comparative analysis between the performance of the QLO framework and the LBM, which does not consider wireless channel variations and utilizes CU as the only load metric, representative of those implemented in [99] and [101] is also examined. Figure 5.7 illustrates an overview of the modified LBM which is essentially implemented in the load control of the QLO framework. Specifically, the original load metric of throughput is being replaced by the CU as throughput is highly influenced by the data rate of STAs, rendering it ineffective as discussed in Section 4.1. Furthermore, an additional admission threshold CU_{max} is also introduced as in [94] and [96] to prevent the AP from operating under a fully loaded medium. This admission threshold will also cater for the bandwidth variability of traffic sources, especially, for VBR sources. The motivation here is to compare the effects of load distribution using the CLM and a single load metric under dissimilar wireless

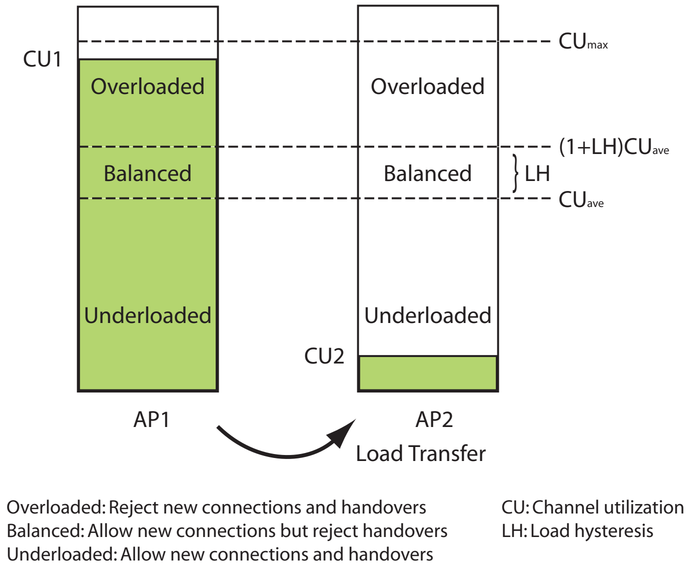
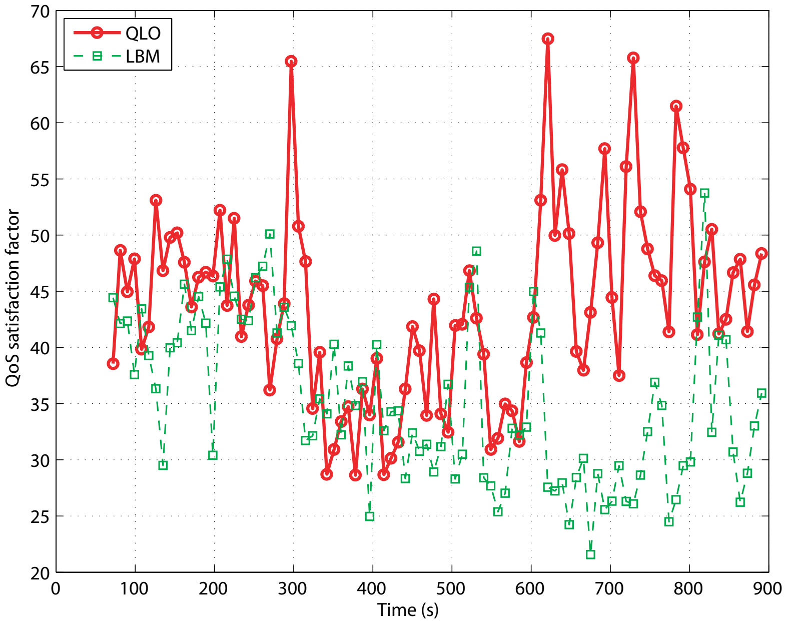

CU1
Overloaded
Balanced
Underloaded
AP1
Overloaded
Balanced
Underloaded
CU2
AP2
Load Transfer
CUmax
(1+LH)CUave
LH
CUave
Overloaded: Reject new connections and handovers
Balanced: Allow new connections but reject handovers
Underloaded: Allow new connections and handovers
CU: Channel utilization
LH: Load hysteresis

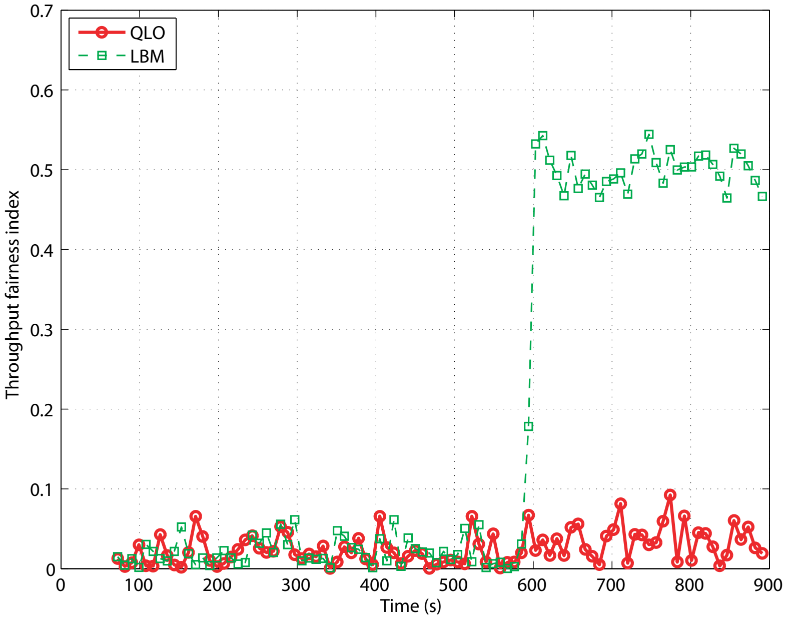
Figure 5.7: Overview on the implementation of the LBM.

channel conditions, as well as investigate the effects of proactive and reactive handover triggers.

Ideally, according to the definitions of the three key performance indicators (KPIs) in (A-7) – (A-9), the QSF should be greater than 1, the TFI should be close to 0, and the QBI should be close to 1 so as to offer QoS guarantee, throughput fairness, and QoS fairness, respectively. First, the effectiveness of QLO to support QoS-demanding multimedia services in terms of QSF of STAs is studied. From Figure 5.8(a), the average QSFs for LBM and QLO are 37.3 and 40.6, respectively for the first 600 s where only traffic variations are simulated. However, the QSF of 31.6 with LBM improves to 48.8 with QLO for the last 300 s when NLOS transmissions are introduced in BSS 2. Clearly, QLO outperforms LBM under such dissimilar wireless channel conditions and provides comparable QoS support under similar wireless channel conditions with traffic variations. Second, the performance of QLO on the throughput of STAs using the TFI is examined. From Figure 5.8(b), QLO exhibits throughput fairness with an average TFI of 0.08 when subjected to traffic variations during the first 600 s and dissimilar wireless channel conditions during the last 300 s. In contrast, LBM has an average TFI of 0.23 for the entire simulation. It



(a) Average QSF of STAs.



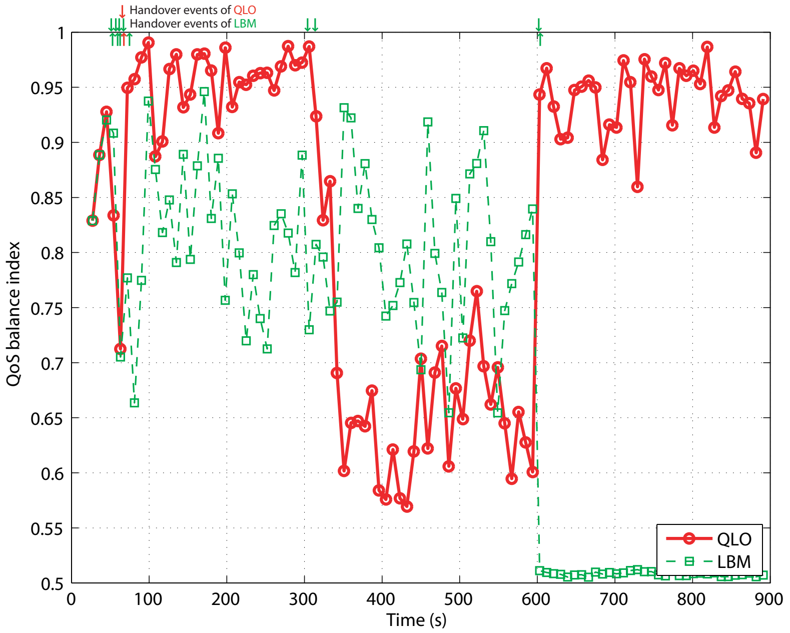
(b) Average TFI of STAs.

Figure 5.8: Scenario 1: Average QSF and TFI of STAs.

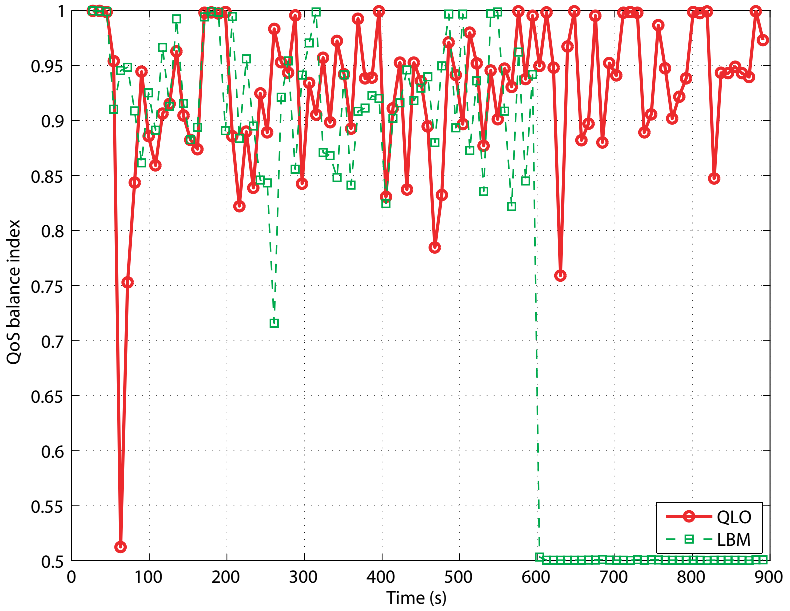
is worth noting that under disparate network conditions during the last 300 s, LBM has a high TFI of 0.5 as some STAs suffer huge deviations from their target throughput. Hence, LBM fails to provide throughput fairness.

Last, the effects of QLO on PD and PLR, which are critical QoS metrics for supporting multimedia services, in particular, RT traffic, in both APs are investigated using the QBI. The results of Figure 5.9 are analyzed in three segments corresponding to the simulated scenario. The simulation starts with an unbalanced load of fifteen STAs in BSS 1 and three STAs in BSS 2 during the first 300 s. Over this period, eight handovers are triggered by LBM in an effort to balance the load between both APs but only two handovers are triggered by QLO. The QBIs of PD for LBM and QLO are 0.71 and 0.85, respectively, and the QBIs of PLR for both LBM and QLO are similar with a value of 0.9. QLO achieves better performance as it handovers two voice STAs, which are the most aggressive source in this simulation, nominated as the 'best' STA by AP 1. Although LBM uses the same algorithm to nominate the 'best' STA, it tries to balance the load in a proactive manner. In contrast, QLO takes the reactive approach such that handover is triggered only if the PLR of the source AP exceeds 1% and when a better quality target AP exists. As a result, no handover is triggered by QLO during the next 300 s even when four additional connections are started in BSS 1, whereas two additional handovers are triggered by LBM. It is for this reason that the QBI of PD for QLO drops to 0.68 while LBM maintains relatively at the same level of 0.79. However, it is important to emphasize that QoS for QLO is not compromised which is evident from Figure 5.8(a). In fact, the QBIs of PLR for both LBM and QLO during the second 300 s are again similar with a value of 0.92. These observations are a direct consequence from the key principle of RQB, which avoids unnecessary handovers when the QoS requirements of STAs can be supported, by opportunistic yet altruistic exploitation.

In the last $300s$, NLOS transmissions in BSS 2, which essentially create disparate network conditions between both BSSs, are introduced. Under such conditions, Figure 5.9 illustrates that QLO is still able to maintain high QBIs of both PD and PLR with values of



(a) Average QBI of PD between APs.



(b) Average QBI of PLR between APs.

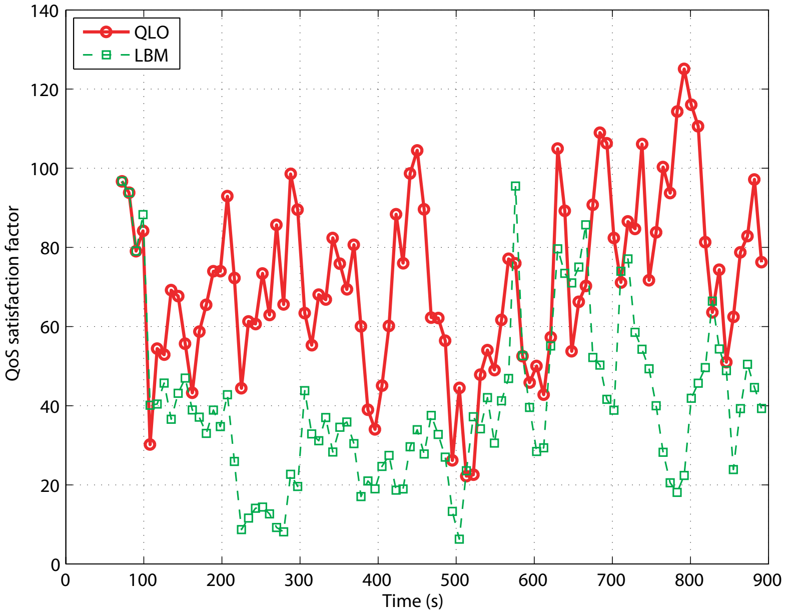
Figure 5.9: Scenario 1: Average QBI of PD and PLR between APs.

0.95 while QBIs of both PD and PLR with LBM drop to 0.5. This signifies that PD and PLR between both APs are extremely unbalanced for the case with LBM. Hence, load distribution based on purely a single load metric such as the CU will fail under diverse wireless channel conditions. As a final note, there are a total of twelve handovers triggered by LBM but only two handovers triggered by QLO. This translates to a significant 83% reduction in handovers while maintaining a QoS-balanced system with high QSF of STAs.

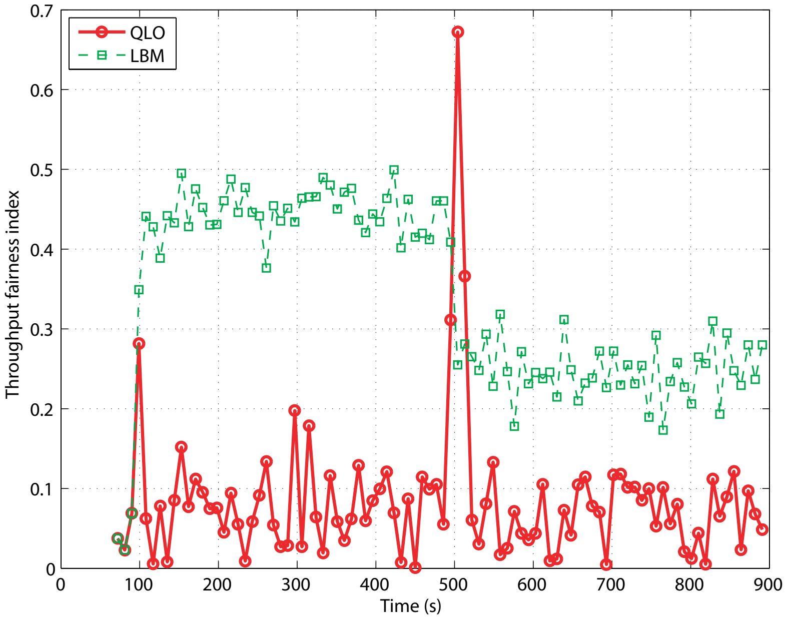
5.3.2 Effect of Wireless Channel Variations

In this study, the impact of wireless channel variations on both LBM and QLO is investigated. Figure 5.10(a) shows that the average QSF of 40.8 for LBM can be improved to 71.9 with QLO. The key reason for this improvement is due to the capability of QLO in detecting STA operating under bad wireless channel conditions where the 'worst' STA is identified as the one with the highest MACKs. Figure 5.10(b) illustrates that QLO preserves throughput fairness with a lower TFI of 0.15 as opposed to LBM of 0.38. Note that the two peaks in Figure 5.10(b) and two troughs in Figure 5.11 of QLO correspond to the adaptation time that detects and subsequently handovers the 'worst' STA to a better quality AP. Accordingly, nine handovers are triggered with QLO during these adaptation periods, whereas no handover is triggered with LBM as shown in Figure 5.11(a). This reiterates that the major pitfall of LBM, which considers CU as a single load metric, is the inability to respond to wireless channel variations. Consequently, Figure 5.11 shows that the QBIs of both PD and PLR for QLO are 0.89 and 0.87 which outperform LBM's of 0.57 and 0.56, respectively.

Clearly, QLO is highly resilient to dynamic network conditions which may arise due to traffic and wireless channel variations, thanks to the CLM in which its QoS context information captures traffic variations explicitly and wireless channel variations implicitly. In practice, many link adaptation algorithms are implemented by vendors to combat varying

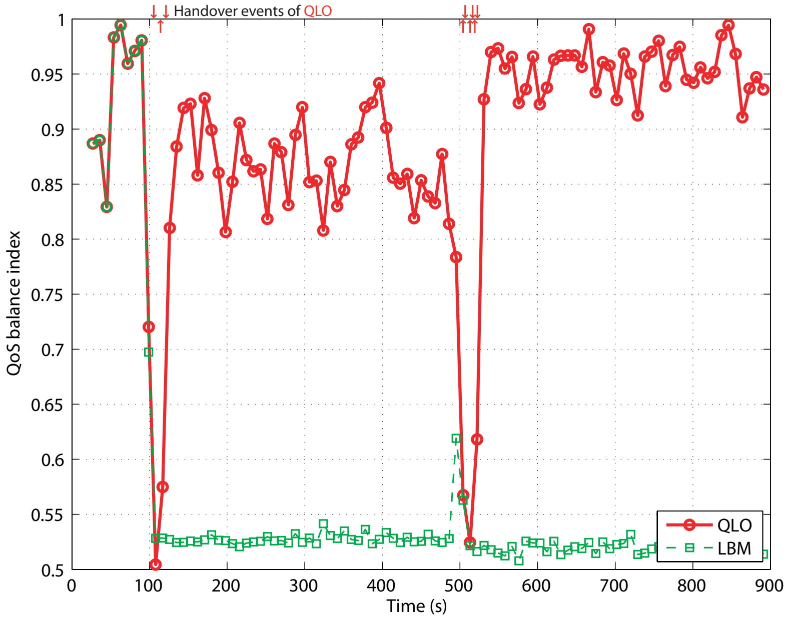


(a) Average QSF of STAs.

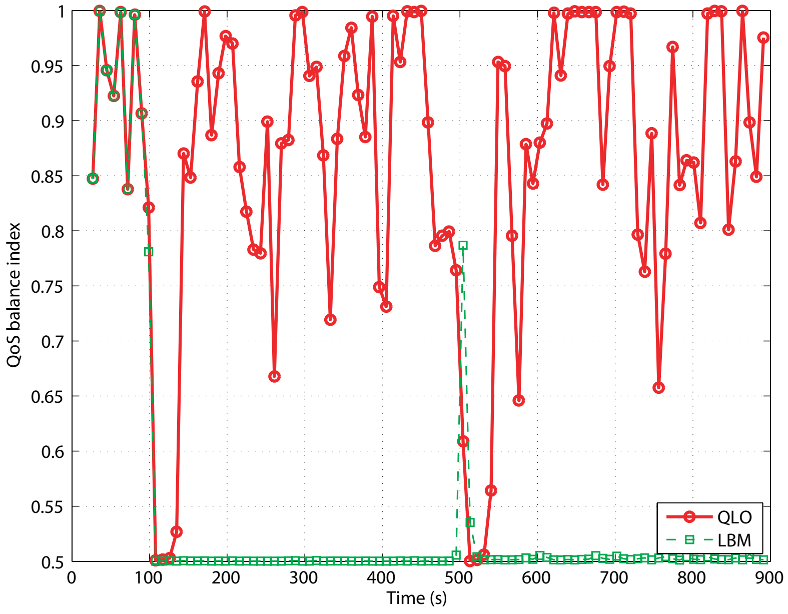


(b) Average TFI of STAs.

Figure 5.10: Scenario 2: Average QSF and TFI of STAs.



(a) Average QBI of PD between APs.



(b) Average QBI of PLR between APs.

Figure 5.11: Scenario 2: Average QBI of PD and PLR between APs.

wireless channel conditions. However, this gives rise to the rate anomaly problem which has serious implications on the perceived network capacity. Specifically, the rate anomaly problem would manifest as *capacity outages* if the arriving traffic is higher than the degraded system capacity caused by link adaptations (cf. Section 5.4). To this end, it is conjectured that the novel concept of RQB employed in the QLO framework can effectively minimize this capacity outage time by maintaining a QoS-balanced system where the 'worst' STA can be detected and transferred to a better quality AP opportunistically. As soon as the 'worst' STA is removed, it is expected that link adaptation procedures will act to increase the transmission rate, thus recovering the overall system capacity. This conjecture will be further validated in Section 5.6.

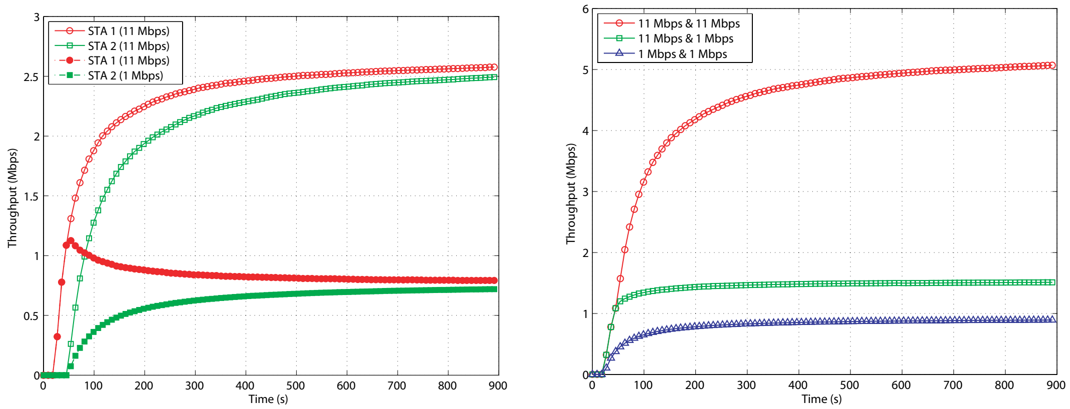
5.4 Challenges of WLAN-based Cognitive Networks

Within the research community, the IEEE 802.11 WLAN is envisioned as one of the introductory de-facto wireless access networks in future cognitive network architecture, attributed to its pervasive deployments over many diverse environments. In fact, according to the Federal Communications Commission [121], WLAN could be considered as cognitive radio since it operates with listen-before-talk access protocol together with dynamically changing transmission powers and data rates to allow more efficient spectrum use. However, the delivery of QoS-demanding multimedia applications over WLAN is not trivial, particularly, in the context of a WLAN-based cognitive network for the following reasons.

The recent IEEE 1900.1 standard [38] states that cognitive network is a composition of radio nodes subjected to cognitive functionality where the cognition process could take place in the radio, in the higher layers, or both. As previously mentioned, cognitive network [14], [36], [37] postulates the ability to optimize both user and network performances, as well as bring about better utilization of radio resources by adopting cognitive functionality, particularly, in the radio to exploit spectrum holes. Although cognitive radio [7], [8] can identify these spectrum holes, a parallel cognitive functionality is required in the higher layer to harness any effective use of these additional spectrums. Given that the frequency spectrum is typically divided into multiple channels, the ramification of dynamic spectrum access supported by cognitive radio will manifest as the main challenge of managing these dynamically available heterogeneous channels which emanate from diverse parts of the spectrum with different propagation characteristics.

On the other hand, link adaptation or adaptive modulation and coding has been widely implemented across standards such as the 3GPP, 3GPP2, IEEE 802.11a/b/g, IEEE 802.15.3, and IEEE 802.16 to provide spectrally efficient and flexible data rate access while adhering to a target error performance over wireless channels. Although link adaptation has the ability to achieve optimum throughput by matching transmission parameters to the time-varying wireless channel conditions of a realistic wireless environment, it introduces an upper bound on the maximum achievable throughput due to the reduction of transmission rate in reality. This would inevitably affect users running bandwidth intensive and delay sensitive multimedia applications where QoS guarantee is extremely important. Such impact is especially pronounced in contention-based networks such as WLAN since link adaptation is known to give rise to the rate anomaly phenomenon [50], [51].

The rate anomaly of DCF can occur when contending STAs in the same BSS, with similar wireless channel conditions, transmit packets of same size but with different data rates. Under such conditions, the DCF preserves throughput fairness such that each STA would yield approximately the same throughput, regardless of its own data rate as illustrated in Figure 5.12(a). The rate anomaly of DCF in UL transmissions is a direct consequence of throughput fairness which dictates an equal probability of channel access. In other words, slower rate STAs will disadvantage higher rate STAs by occupying more channel time to acquire approximately the same number of transmission opportunities. In fact, Figure 5.12(b) also illustrates that the aggregate throughput of multirate transmissions by two STAs using 11 Mbps and 1 Mbps is much lesser than the average of the total



(a) STA's throughput.

(b) Aggregate throughput.

Figure 5.12: Impact of the rate anomaly of DCF in UL transmissions.

throughput that could be achieved when both STAs transmit using either 11 Mbps or 1 Mbps.

While the DCF will affect channel bandwidth allocation and cause rate anomaly in the UL, the AP queueing mechanism will affect channel bandwidth allocation and similarly induce rate anomaly in the DL for an infrastructure BSS. The rate anomaly of AP queueing mechanism in DL transmissions is attributed to varying wireless channel conditions encountered at different STAs. For the case of a single FIFO queue in a typical IEEE 802.11 DCF-based AP, head-of-line (HOL) blocking will occur as soon as the wireless channel quality toward the STA of the HOL packet is degraded due to burst error [122] since the AP first performs retransmissions and eventually link adaptation to transmit that packet at a lower rate. Such prolonged HOL blocking will cause the unfair allocations of channel bandwidth and severe packet loss due to buffer overflow once the arrival rate exceeds the service rate. To overcome this form of unfairness due to location-dependent errors, many fair queueing mechanisms have been developed for wireless environments, e.g., wireless fair service, idealized wireless fair queueing algorithm, channel-condition independent fair queueing [123], and distributed weighted fair queueing [124]. These algorithms are primarily designed to consider only *DL scheduling* based on a *single rate* server to provide throughput fairness. In reality, however, the IEEE 802.11 networks are

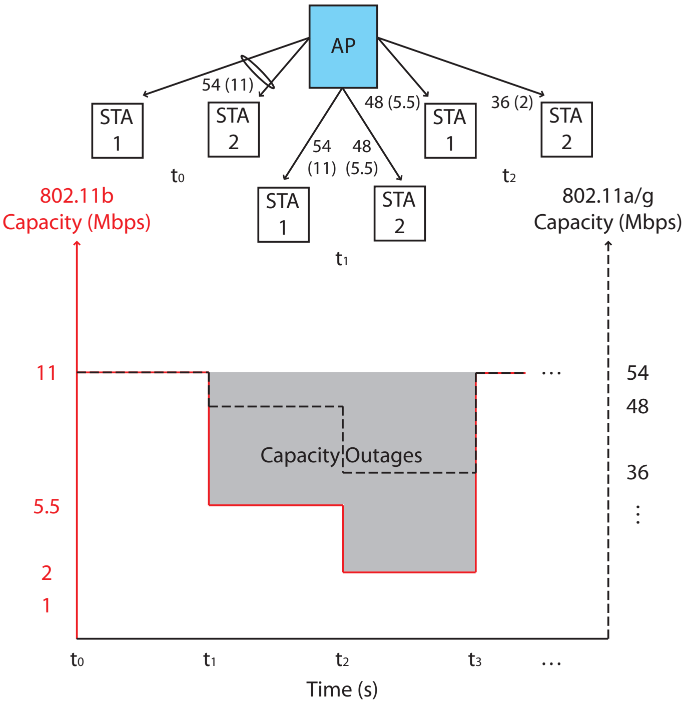


Figure 5.13: Impact of the rate anomaly of AP queueing mechanism in DL transmissions.

half-duplexed and subjected to multirate operations as a consequence of link adaptation used to combat varying wireless channel conditions at the STAs. Hence, when these fair queueing mechanisms are deployed in the AP of multirate WLAN, the rate anomaly of AP queueing mechanism will occur as exemplified in Figure 5.13 and described in the following.

At time t_0, both STAs associated to the AP experience similar wireless channel conditions and receive packets from the AP at the maximum rate of 54 (11) Mbps. At time t_1, STA 2 experiences poor wireless channel conditions and receives packets from the AP at a reduced rate of 48 (5.5) Mbps due to link adaptation. By the throughput fairness property of AP queueing mechanism, STA 1 will be disadvantaged and eventually receives packets at a rate similar to STA 2. When this happens, the effective capacity of the system will be reduced to 48 (5.5) Mbps. At time t_2, STA 2 continues to suffer wireless channel quality degradation and the effective capacity of the system will be further reduced to

36 (2) Mbps. Clearly, the rate anomaly of AP queueing mechanism will create severe capacity outages in terms of packet loss when the offered load exceeds the downgraded system capacity, particularly, when the AP performs link adaptation to combat varying wireless channel conditions experienced by the STAs. Notice that such capacity outages will be even more significant for the IEEE 802.11b WLAN, as indicated by the data rates in parentheses, due to its coarser link adaptation steps.

Apparently, rate anomaly will be unavoidable in a multirate WLAN-based cognitive network since link adaptation is widely employed to combat varying wireless channel conditions which could occur as a consequence of: (i) channel impairments in hotspot and indoor environments arising from frequent NLOS transmissions caused by structures and obstacles; and (ii) dynamic spectrum access supported by cognitive radio, resulting in heterogeneous channels operation. The impact of rate anomaly will escalate for WLAN-based cognitive network since cognitive radio can opportunistically access diverse channels which are subjected to very different amount of frequency-dependent path loss, multipath effects, and attenuations. Specifically, the negative impact of rate anomaly will be significant when two or more STAs contend with diverse data rates *and* during heavy network load situations. Moreover, this impact would be magnified in the coexistence of the IEEE 802.11b and IEEE 802.11g WLANs as the former has coarser link adaptation steps. It is also worth to mention that the achievable throughput is not easily predictable if different data rates traverse both UL and DL, even though all STAs use the DCF and the AP employs a fair queueing mechanism [51]. Hence, a serious implication of rate anomaly is that it will *dilute* the benefits derived from any link adaptation techniques to exploit the tradeoff between date rate and BER under varying wireless channel conditions.

5.5 Quad-Domain Cooperation

Numerous solutions have been proposed to address rate anomaly prevalent in multirate environment. Some notable works in [51], [106], and [125] address this problem from a temporal fairness perspective in order to overcome the throughput fairness of DCF. Tan and Guttag [51] are the first to tackle the rate anomaly problem by advocating the concept of temporal fairness where each contending STA receives an equal share of the channel occupancy time. They have highlighted that such time-based fairness has an important baseline throughput property, which is not available from the throughput fairness of the DCF, and this could be used to mitigate the rate anomaly of multirate WLAN. Tinnirello and Choi [106] have shown that the EDCA channel access mechanism in the forthcoming IEEE 802.11e standard can also provide temporal fairness by using TXOP to maintain equal channel occupancy time for all contending STAs. However, the authors show that there is a tradeoff between achieving temporal fairness and overall system bandwidth efficiency as the equalization of channel occupancy time requires fragmentation which introduces significant overheads. Most importantly, the study also reveals that temporal fairness can be guaranteed only under uniform TXOP settings. However, this would be difficult to realize in practice as STAs can select different fragmentation sizes without a fixed threshold, as opposed to the DCF fragmentation rule. Hence, the challenges in QoS provisioning for multimedia traffic delivery over the DCF remain as an open research issue.

Recently, Joshi *et al.* [125] present a distributed time fair carrier sense multiple access scheme to mitigate the rate anomaly problem of DCF. The authors use the baseline property suggested in [51] to estimate a target throughput for each contending STA. The key idea is that STAs would adjust their minimum CW to meet the estimated baseline throughput which has the intrinsic property of providing temporal fairness. However, they make several strong assumptions that limit pragmatic implementation. First, the authors assume that every STA has a priori knowledge of the number of contending STAs which they concede is non-trivial, especially, in multirate WLAN. Second, in their proposal, each STA assumes that transmission rate, packet size, and packet error rate experienced by all other STAs are the same as itself. However, this scheme will fail to guarantee airtime fairness

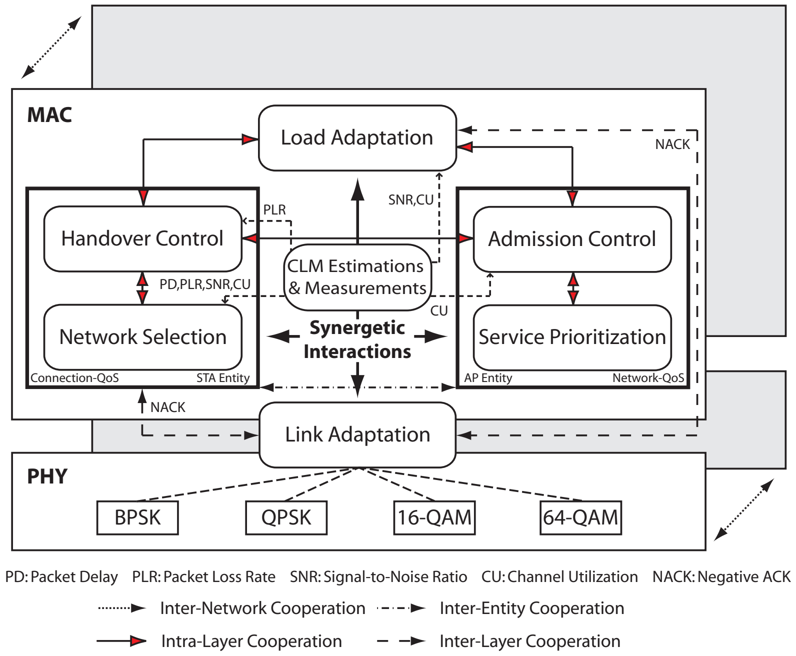


Figure 5.14: Quad-domain cooperation: The distributed LAS framework.

when traffic of different packet sizes exists, in reality, due to the fact that the baseline throughput is dependent on packet sizes [51].

5.5.1 Inter-Layer Cooperation

This thesis takes a dramatically novel approach to mitigate rate anomaly from a QoS perspective which is first proposed in [118] and later extended in [126]. Particularly, the QLO framework for a single rate WLAN in Section 5.2.2 is extended to devise a harmonized solution for multimedia service delivery in a dynamic multirate WLAN-based cognitive network by introducing the notion of LAS to mitigate rate anomaly. The distributed LAS framework which supports quad-domain cooperation is illustrated in Figure 5.14. It leverages on an additional inter-layer cooperation between the PHY and MAC layer to enable synergies between: (i) link adaptation and connection-QoS entity; and (ii) link adaptation and load adaptation such that their benefits can be exploited on-demand.

Without loss of generality, the key idea is to exploit heterogeneity within such multi-AP cognitive network, whenever possible, through the existing tri-domain cooperation, to promote a QoS-balanced system by the redistribution of load in an opportunistic yet altruistic manner. By further inducing synergetic interactions between link adaptation and load adaptation, the QoS requirements of multimedia STAs can be better satisfied and not be penalized unnecessarily due to rate anomaly. Figure 5.15 illustrates that STA with degrading QoS can be identified and transferred to a better quality AP in an effort to maintain a QoS-balanced system through the LAS. As soon as the STA with degrading QoS is removed, link adaptation will increase the transmission rates of both the STA and AP, which in turn recover the QoS and overall system capacity, respectively. In fact, such synergetic interactions between link adaptation and load adaptation on-demand is an important aspect of optimized handover that is also raised in the recent IEEE 802.21 standard [16].

To the best of the author's knowledge, this is the first attempt to address rate anomaly in a multirate WLAN-based cognitive network, where the serious implication of rate anomaly arising from link adaptation becomes imperative, from a single unifying generalized CCRRM architecture. The primary contributions differ from the related works in three significant ways: (i) the notion of RQB which exhibits an additional intrinsic property of mitigating rate anomaly is corroborated as a feasible criterion to quantify the state of balance in a multirate WLAN-based cognitive network; (ii) the CLM, which is augmented from the QLO framework, consists of PD, PLR, SNR, CU, and NACK for enhanced adaptation to dynamic network conditions prevalent in a multirate WLAN-based cognitive network; and (iii) the LAS, which is compatible with both DCF and EDCA access mechanisms, enables multimedia traffic delivery over a multirate WLAN-based cognitive network within a single unifying QoS framework.

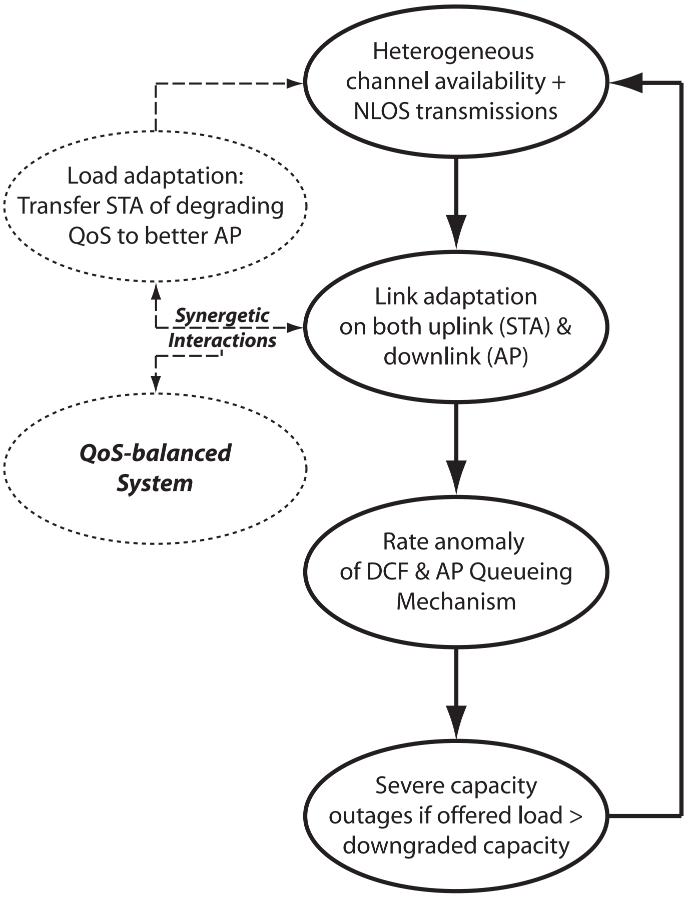


Figure 5.15: LAS: A QoS perspective. Solid lines depict the implication of rate anomaly arising from link adaptation. Dashed lines depict synergetic interactions between link adaptation and load adaptation to mitigate rate anomaly and maintain a QoS-balanced system.

5.5.2 Load Adaptation Strategy Framework

The underlying principle of the LAS framework first proposed in [126] is also based on network-assisted discovery, in which the source AP broadcasts the measurement report of neighboring APs together with its own, over the TONA handover architecture as presented in Section 3.2. The QoS broadcast (cf. Section 3.2.3) contains the QoS context information of PD, PLR, CU, and SNR, which is global to both AP and STAs. Although it has been shown in Sections 5.3.1 and 5.3.2 that the QoS context information of PD and PLR can capture wireless channel variations implicitly, the SNR is employed here so that it can be used to make explicit decisions concerning the wireless channel conditions proactively. E.g., STA could select an AP according to its SNR requirements on top of its QoS requirements and the AP load conditions. In addition, the AP could make use of the SNR information to effectuate load distribution more proactively when dealing with multirate WLAN. For the case of a multi-RAT environment with heterogeneous wireless networks, the SNR can be normalized by using the Shannon capacity formula to relate the maximum achievable data rate from different technologies as in [127].

An additional NACK, which is a local information to both AP and STAs, is included with the QoS context information to form the CLM. Note that this CLM is an enhancement over the one employed in the QLO framework. The NACK is used as input by: (i) the AP to detect STAs with degrading QoS; and (ii) the respective AP and STAs to invoke link adaptation procedures. The idea of NACK, in general, is similar to MACK as explained in Section 5.2.2, except for the fact that MACK could also occur due to collisions. It has been pointed out in [128] that link adaptation procedures such as automatic rate fallback which depend on MACK will suffer throughput degradation and prolonged collision time, determined by the slower STA, when transmission failure is due to collisions rather than link errors. Hence, the idea of NACK is used instead for distinguishing between link errors and collisions. This is an enhancement over MACK for two reasons: (i) link adaptation procedures will not be falsely triggered to reduce data rate; and (ii) STA will not be unnecessarily identified as the ‘worst’ STA to perform load adaptation during collisions in which exponential backoff is more appropriate.

Figure 5.16 illustrates the thematic connections of the LAS framework. Accordingly, the PD and PLR are short-term data used by STA to make handover decision, whereas the additional SNR information together with the CU are medium-term data that serve as constraints for network selection and load adaptation. The handover information discovery and conditioning by using the technology abstraction and link cognition module of the generalized CCRRM architecture have been described in Chapter 3. The measurement-based network selection relies on the greedy approach in which the optimal AP with the highest network quality probability will be selected based on the short-term data of PD and PLR.

On the other hand, the load adaptation plays a central role by serving two main functions. First, it identifies the ‘worst’ STA with the highest NACKs among STAs with degrading QoS for unplanned handover as in (5.4). Second, it attempts to perform RQB by selecting the ‘best’ STA for planned handover as in (5.5). Note that the ‘worst’ STA has priority over the ‘best’ STA in order to mitigate rate anomaly by reducing the capacity outage time. It is important to note that both ‘best’ and ‘worst’ STAs correspond to the *most aggressive* traffic source associated with the source AP. The ‘best’ STA will be the traffic source with the highest CU that can fit into the available capacity of the target AP. The ‘worst’ STA will be the traffic source with the highest NACKs, which also corresponds to the traffic source with the highest CU since the probability of NACKs increases with higher traffic arrival rate. This *highest channel utilization first out (HCUFO)* policy, also based on the greedy approach, is attractive as it can essentially reduce the number of unnecessary handovers and result in the long-term uniform distribution of aggressive traffic sources over the multi-AP cognitive network, which is beneficial from a load distribution perspective. Such greedy approaches are adopted due to the fact that obtaining an optimal allocation of STAs to the available APs such that the allocation maximizes the overall composite capacity is a combinatorial problem which is non-deterministic polynomial-time (NP)-hard [129], [130]. Collectively, these greedy approaches contrive the classical 'balls into bins' heuristic which provide an intuitively simple way to jointly optimize overall composite capacity while mitigating rate anomaly and providing statistical QoS guarantee. Note that the HCUFO policy together with RQB are part of the RRM policy described in Section 3.2.3 by issuing the target group bitmask of (x0100000) and the RRM policy bitmask of (xxxx1000).

Similar to the iLB scheme and QLO framework, the design philosophy of the LAS framework is again based on the key principle of RQB (cf. Section 4.2) to exploit heterogeneity within a multi-AP cognitive network in an opportunistic yet altruistic manner. Accordingly, the LAS is a cognitive functionality of the higher layer that resides in the MAC. More specifically, the LAS, which is extended from the QLO framework, will also provide synergetic interactions between network-QoS and connection-QoS entities to nominate candidate STAs for handover based on the CLM as explained in Section 5.2.2. This facilitates joint optimization within the MAC layer and between network-terminal entities to perform distributed RRM decision in which STAs make the final handover decision. For completeness, the network-QoS entity consists of both service prioritization and admission control to deal with different user service profiles while the connection-QoS entity consists of both network selection and handover control to deal with dynamic network conditions associated with network congestions, channel impairments, and dynamic spectrum access. More importantly, the LAS is responsible for synergetic interactions between the PHY and MAC layer to exploit the benefits of both link adaptation and load adaptation on-demand. To be more specific, load adaptation through VHO will be invoked when opportunistic yet altruistic exploitation is possible, otherwise link adaptation will be invoked. The pseudo codes describing the algorithms of the LAS framework based on quad-domain cooperation in both AP and STA are given in Algorithm 5.1 and Algorithm 5.2, respectively.

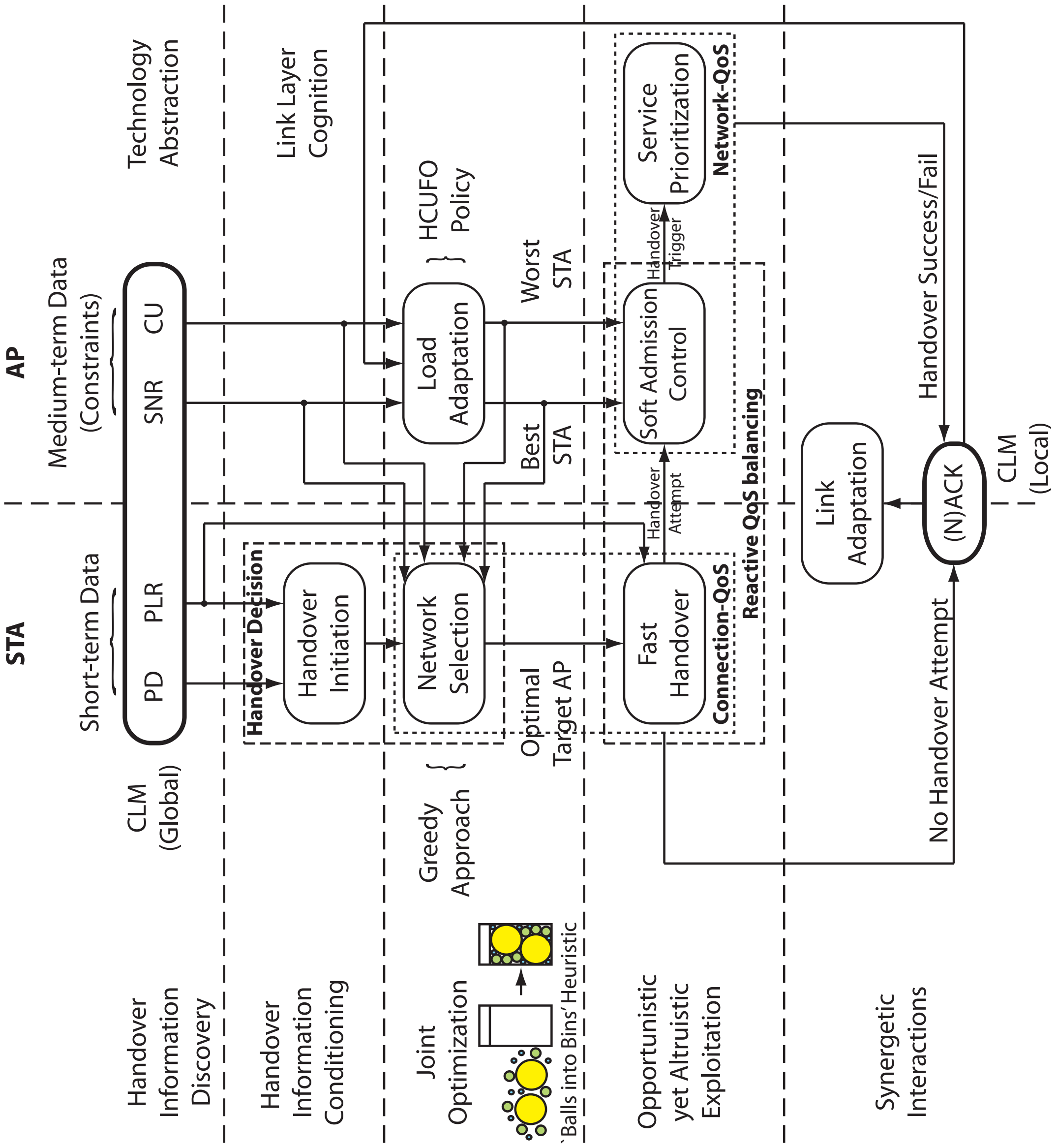


Figure 5.16: An overview on the thematic connections of the LAS framework.

Algorithm 5.1 LAS algorithm in an AP.

```
  On Reception of Measurement Report from APC
Require: Periodic update of measurement report in each AP every QoS broadcast interval
  AP extract QoS context information from measurement report
  /* Compute total CU in each AP */
  $CU^n_{total} \leftarrow \sum_{i \in STAs} CU^n_i, \quad CU^n_i = \sum_{k \in Links} CU^n_{i,k}, \quad n \in APs$
  /* Compute average CU across APs */
  $CU_{ave} \leftarrow \frac{1}{n} \sum_n CU^n_{total}$
  /* Determine available capacity in target APs for load distribution */
  $CU^m_\Delta \leftarrow CU_{ave} - CU^m_{total}, \quad m \in APs, \quad m \neq n$
  On Link Adaptation
  if NACK_RCVD then
    $NACK^n_i \leftarrow NACK^n_i + 1$
  else // ACK_RCVD
    $ACK^n_i \leftarrow ACK^n_i + 1$
  end if
  On Load Adaptation
  for all $i \in STAs$ do
    if $CU^n_{total} > CU_{ave}$ or $SNR_n < SNR_{threshold}$ then
      /* Prevent load distribution to overloaded AP & overloading target AP */
Require:      $CU^m_\Delta - CU^n_i \geq 0, \quad m \neq n$
      /* Nominate 'best' STA for handover */
      $STA^n_{best} \leftarrow \min \quad CU^m_\Delta - CU^n_i$
    end if
    /* Transfer STA with the lowest PHY data rate $R_{DATA}$ and highest NACK */
Require:    $\{i \in STAs : i = \min(r_i \in R_{DATA})\}$
    /* Nominate 'worst' STA for handover */
    $STA^n_{worst} \leftarrow \max \quad NACK^n_i$
  end for
  On Admission Control
  if $CU^n_i + CU^n_{total} < CU_{max}$ then
    /* Invoke SERVICE_PRIORITIZATION for admitted connection */
  end if
```

Algorithm 5.2 LAS algorithm in a STA.

```
On Reception of Beacon from AP
Require: Periodic update of beacon information in each STA every QoS broadcast interval
STA extract QoS context information, $T_{LHO}$, and RRM policy from beacon frame
On Initial Access
for all $i \in$ initial access STAs do
    /* Invoke NETWORK_SELECTION */
    if $QoS_n \geq QoS_{req,i}$ then
        /* Association with AP upon admission */
    end if
end for
On Synergetic Interactions between the PHY and MAC layer
for all $i \in$ handover STAs do
    /* Invoke LOAD_ADAPTATION */
    if $QoS_m > QoS_n$ and $PLR_n > 1\%$ and STA nominated, $m \neq n$ then
        /* Invoke HANDOVER_CONTROL */
        while $T_{current} - T_{LHO} < T_{stability}$ do
            /* Wait for stability period to elapse */
        end while
        if WORST_STA and BEST_STA are true then
            /* ‘Worst’ STA has priority */
        end if
        /* Reassociation with new AP upon admission */
    else // Load adaptation conditions not met
        /* Invoke LINK_ADAPTATION */
    end if
end for
```

5.6 Performance Evaluation of the LAS Framework

To demonstrate the effectiveness of the LAS framework, built on the basis of quad-domain cooperation, simulation models are developed by using OPNET™ Modeler® 14.0 with Wireless Module. Modifications to the DCF models developed for the QLO framework are performed to include link adaptation and load adaptation which are the focus of this study. Moreover, an additional AP is modeled here as compared to the performance evaluation of the QLO framework which is based on two APs. Furthermore, multipath based on the exponential channel model as described in Appendix A-2.2 is considered, in addition to the log-normal path loss model for capturing different propagation characteristics as a result of NLOS transmissions and/or dynamic spectrum access.

A typical hotspot which consists of a homogeneous multirate multi-AP WLAN-based cognitive network with three IEEE 802.11g APs, each operating at an initial data rate of 54 Mbps is simulated as shown in Figure 5.17. In this simulation, a balanced load of three voice, three video, and three FTP STAs in each BSS is considered. The wireless channel variations in each BSS are simulated by introducing NLOS transmissions and the effects of dynamic spectrum access. Specifically, the wireless channel variations are introduced according to Table 5.2 where binary representations of 0 to 7 are used to simulate 2^3 possible states, resulting in either 'high' or 'low' SNR in each AP per 100 s interval. Shadow fading and multipath are included for the entire simulation duration, and multimedia traffic sources are simulated according to Table 5.1. The MSDU lifetime limit mechanism is also incorporated to discard MSDUs from the transmitter queue if they exceed the MSDU lifetime before successful transmission. More details on the general simulation models are available in Appendix A-2.

Without loss of generality, the QoS performance of the LAS framework is evaluated in terms of the QSF, TFI, and QBI, as in the case of the QLO framework, defined in Appendix A-3.1. In addition, the aggregate throughput and source of packet loss in a multi-AP cognitive network are presented. A comparative study between the perfor-

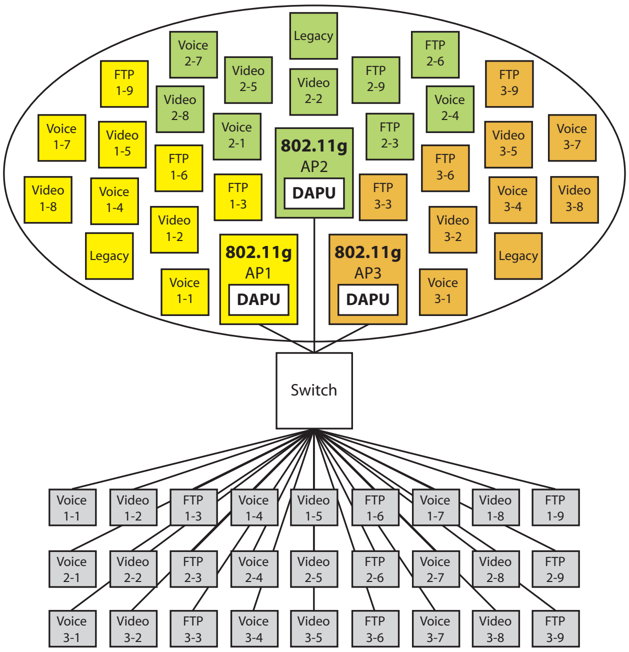


Figure 5.17: Simulation model of a homogeneous multirate multi-AP WLAN-based cognitive network with the IEEE 802.11g APs subjected to wireless channel variations.

Table 5.2: Wireless channel variations.

State	SNR		
	AP1	AP2	AP3
0	high	high	high
1	high	high	low
2	high	low	high
3	high	low	low
4	low	high	high
5	low	high	low
6	low	low	high
7	low	low	low

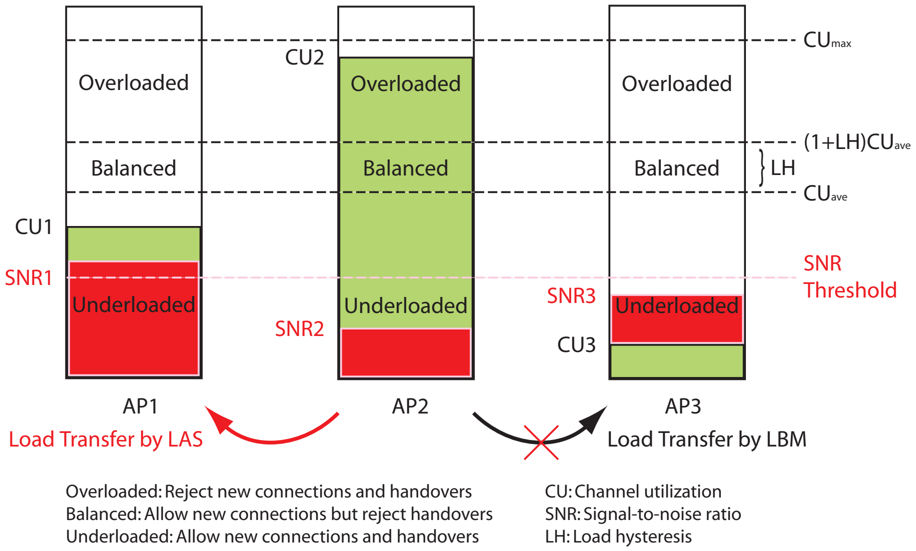


Figure 5.18: Overview on the implementation of the LBM vs. LAS framework.

mance of the LAS framework and the typical LBM implementation in [101], which does not consider wireless channel variations and utilizes the CU as the only load metric, is also investigated. Figure 5.18 illustrates an overview on the implementation of the LBM and LAS framework. In this case, an additional SNR metric is explicitly employed in the LAS framework together with the CU to function as constraints for load distribution as discussed in Section 5.5.2 (cf. Figure 5.16). Ideally, according to the definitions of the three KPIs in (A-7) – (A-9), the QSF should be greater than 1, the TFI should be close to 0, and the QBI should be close to 1 so as to offer QoS guarantee, throughput fairness, and QoS fairness, respectively. The results are analyzed in eight states, which correspond to the simulated scenario, starting from 100 s (0 – 100 s is the warm-up period).

From Figure 5.19, LAS has an average QSF of greater than 1 except for state seven. LBM also has an average QSF of greater than 1 except for state seven and at the beginning of states four and six. In general, LBM has higher average QSF than LAS except for state seven when all the APs have low SNR. This counterintuitive result can be explained from the design philosophy of RQB to avoid unnecessary handovers when the QoS requirements of STAs can be supported. Since the traffic load distribution is the same with 68%

load per AP, LBM does not trigger any handovers as the load hysteresis δ is set to 10% and the CU of all three APs is deemed balanced. However, note that when the load hysteresis is zero, LBM has similar results in terms of average QSF and aggregate throughput but a massive 40% increase in handovers as compared to LAS. Although the load hysteresis could prevent unnecessary handovers, it is static, and hence it is not responsive to wireless channel variations. Evidently, the major pitfall of using the CU as a single load metric in LBM is the inability to manage wireless channel variations which could arise due to NLOS transmissions and/or dynamic spectrum access in a WLAN-based cognitive network. On the other hand, LAS enables the handover of STAs with degrading QoS to a better quality AP in an opportunistic yet altruistic manner. Accordingly, the load of a good quality AP may be filled up to the admission threshold of 87%. Thus, the average QSF of LAS under such conditions will be lower than LBM. However, it is important to emphasize that the QoS requirements of STAs are not compromised as the average QSF of LAS is still greater than 1, thanks to its altruistic design. On the contrary, LAS has higher average QSF than LBM during state seven even when all APs have low SNR since one of the APs has only 33% load while the remaining two APs have about 85% load each. It follows that the AP which is lightly loaded is not affected by low SNR. Hence, all the associated STAs can still meet their QoS requirements, which result in higher average QSF.

From Figure 5.20, LAS achieves higher QBI than LBM for all states except for state seven where LBM has higher QBI as the QSFs of all STAs are equally *bad*. This observation is in direct contrast with that of Figure 5.19, which suggests that tradeoffs exist between the average QSF and QBI. For every increase in the QBI achieved by LAS, there is a corresponding decrease in the average QSF as compared to LBM. In other words, LAS *trades* the QSF for the QBI in order to maintain a QoS-balanced system. Although LBM has higher average QSF, it fails to provide any QoS fairness which means that there is a huge disparity in the QoS or throughput between STAs of the same service class. Moreover, Figure 5.21 illustrates that the QBI of STAs' QSF per service class

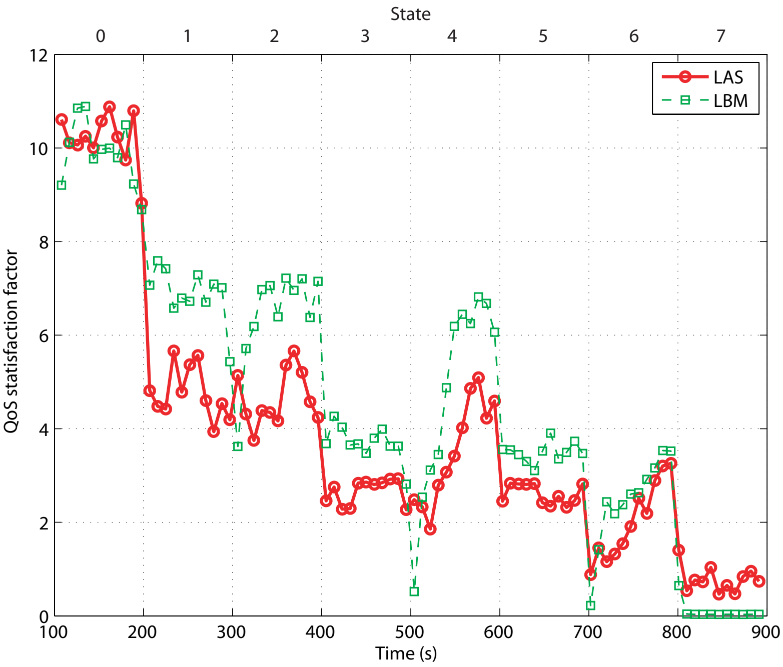


Figure 5.19: Average QSF of STAs.

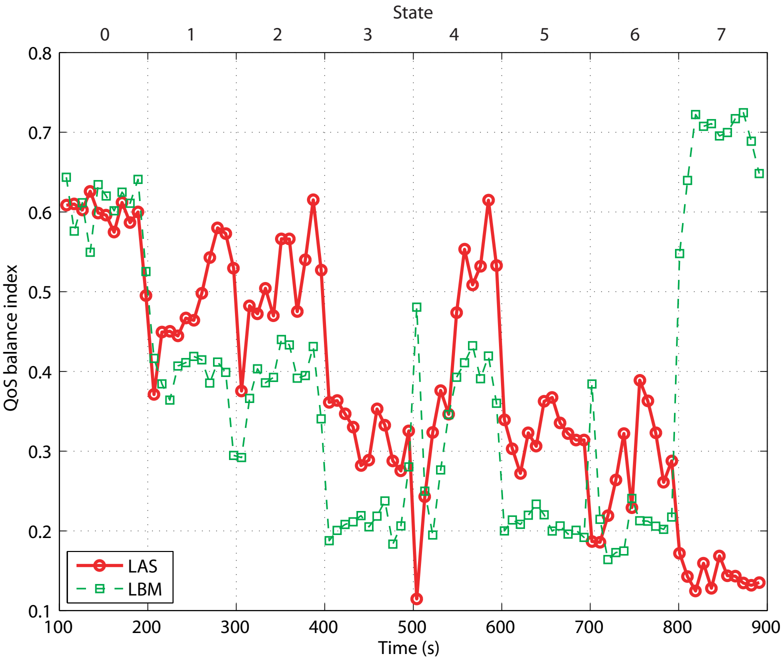
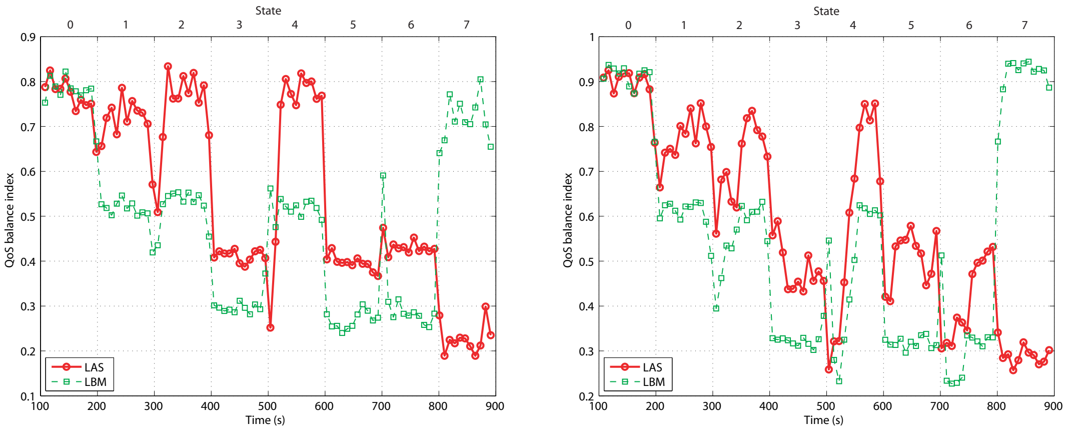
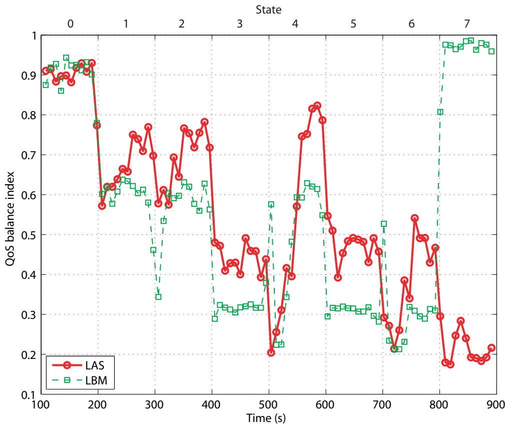


Figure 5.20: Average QBI of STAs' QSF.



(a) Voice STAs.

(b) Video STAs.



(c) Data STAs.

Figure 5.21: Average QBI of STAs' QSF for different service classes.

shows similar trends to Figure 5.20. Specifically, it shows that voice STAs yield the best improvement in the QBI over each state duration, followed by video STAs and then data STAs. Such strict QoS prioritization is a desirable outcome of employing the strict priority queueing technique to provide service prioritization in the DCF, which the EDCA cannot guarantee as explained in Section 5.1.

From Figure 5.22, LAS attains the highest aggregate throughput as compared to LBM and DCF across all states. Note that the DCF serves as a baseline for this comparative study since both LAS and LBM are essentially DCF-based. Apparently, both LBM and DCF suffer marked throughput degradation as a consequence of rate anomaly in the DL which causes HOL blocking at the AP's FIFO queue. This subsequently extends to buffer overflow and excessive RT packets being dropped due to MSDU lifetime expiry for the case of LBM as depicted in Figure 5.23. Note that the MSDU lifetime expiry mechanism is not implemented in the DCF as specified by the IEEE 802.11 standard, thus no packets are lost due to the expiry of MSDU lifetime. In this simulation, rate anomaly in the DL has a more dominant effect than the UL since an infrastructure BSS WLAN is DL limited and the DCF has only a single FIFO queue in the MAC layer. Thus, it can be concluded that although LBM (with link adaptation) yields better throughput than DCF (without link adaptation) under varying wireless channel conditions, the throughput enhancement is *rate-limited* by link adaptation. In fact, Figure 5.23 also illustrates that although link adaptation reduces the amount of packet loss in the PHY due to retry limit exceeded, it results in significant packet loss due to buffer overflow and MSDU lifetime expiry in the MAC layer. It is worth noting that this dilution of link adaptation gains will become more substantial with the IEEE 802.11b or mixed IEEE 802.11b/g WLANs as explained in Section 5.4 (cf. Figure 5.13).

Another interesting observation is that LAS has a throughput gain of 20% over LBM in states one, two, and four where $2/3$ of the APs have high SNR. On the other hand, LAS has a throughput gain of 12.5% over LBM in states three, five, and six where only $1/3$ of the APs have high SNR. Similarly, Figure 5.24 illustrates that LAS obtains the highest improvement in states one, two, and four followed by states three, five, and six. It is worth noting that LBM with higher TFI on average implies that some STAs suffer larger deviations from their target throughput, and hence LBM is unlikely to provide any throughput fairness. As a matter of fact, these observations are consistent with the QBI but inversely related to the QSF as LAS trades the QSF for the QBI. Obviously, these also suggest that the throughput and QoS performance gains of LAS will scale with the heterogeneity of increasing number of APs as the probability of all APs with low SNR at the same time is remote. Given that multi-AP (multi-RAT) deployments in excess of three APs (RATs) will become a reality in future wireless networks, it is argued in this

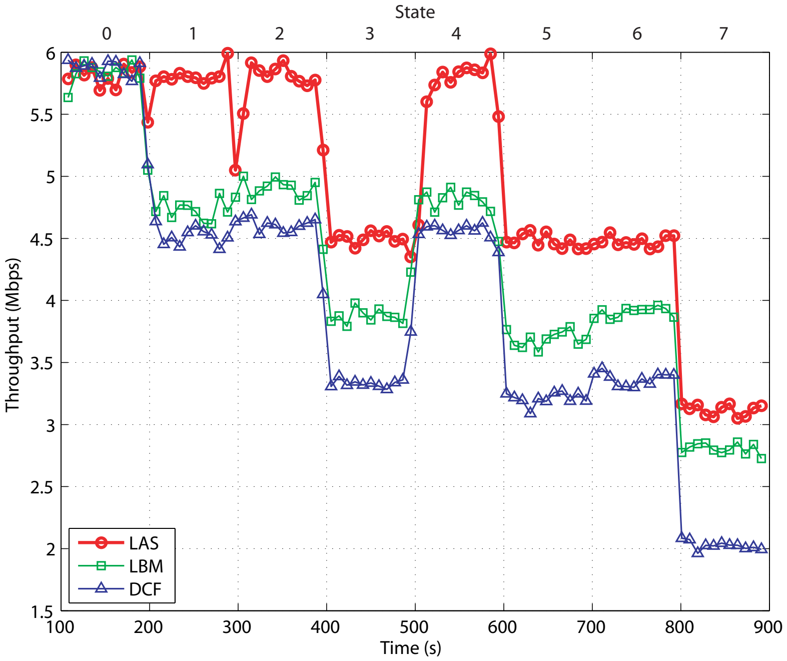


Figure 5.22: Average aggregate throughput in a multi-AP cognitive network.

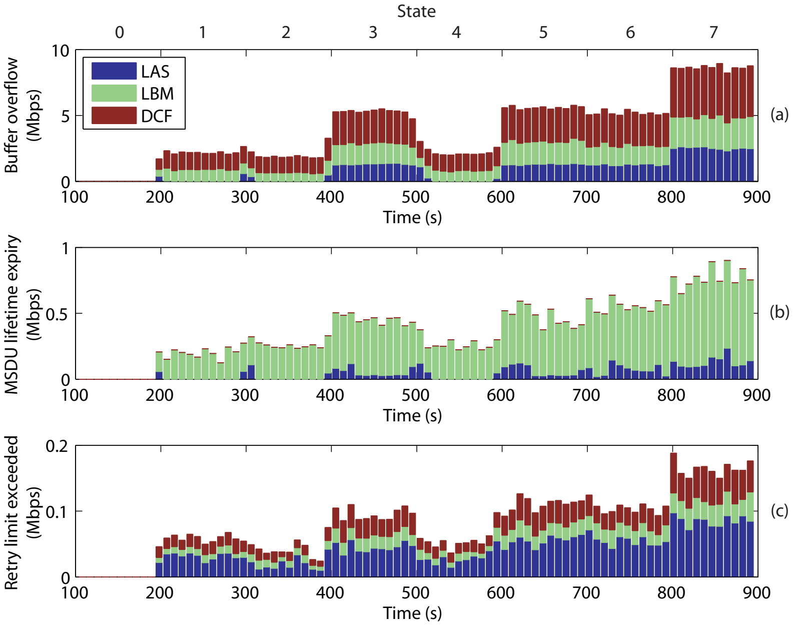


Figure 5.23: Average packet loss due to (a) buffer overflow, (b) MSDU lifetime expiry, and (c) retry limit exceeded in a multi-AP cognitive network.

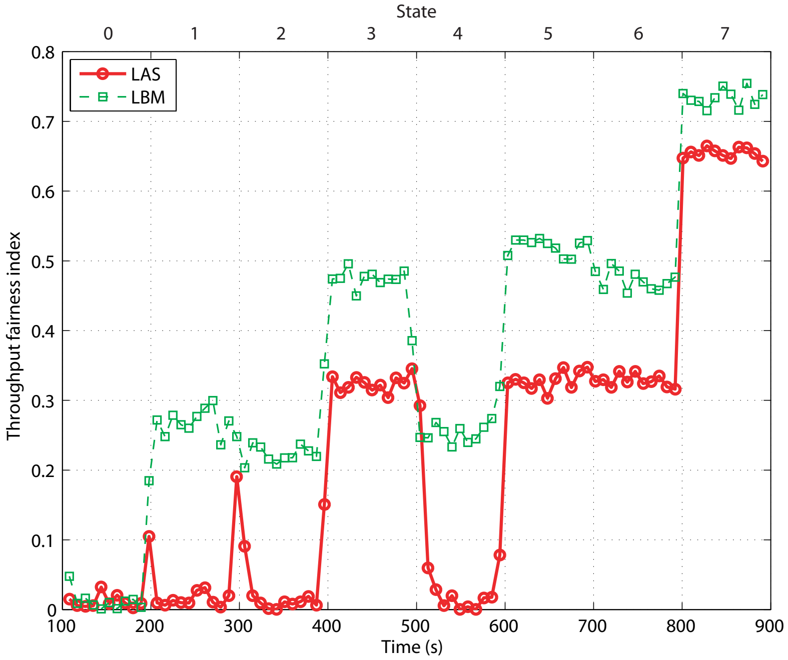


Figure 5.24: Average TFI of STAs.

thesis that heterogeneity should be exploited, whenever possible, to harness higher overall system capacity.

Clearly, these results corroborate that by maintaining a QoS-balanced system, LAS provides better utilization of radio resources through the redistribution of load in an opportunistic yet altruistic manner. In fact, RQB has exhibited intrinsic properties of mitigating rate anomaly in multirate environments, providing statistical QoS guarantee, and maximizing overall system capacity. These validate the conjecture in Section 5.3.2 that the capacity outage time can be minimized by maintaining a QoS-balanced system.

5.7 Chapter Summary

This chapter has explored two different flavors of multi-domain cooperation conceivable within the generalized CCRRM architecture. First, a novel QLO framework based on tri-domain cooperation has been developed to give a unifying treatment in QoS provisioning for multimedia traffic delivery over the widely deployed single rate DCF-based WLANs. Second, a novel LAS framework based on quad-domain cooperation has been developed by incorporating link adaptation with the QLO framework to devise a harmonized solution for a multirate multi-AP WLAN-based cognitive network. The key finding in this chapter has unveiled that load distribution with a single load metric such as the CU will fail under diverse wireless channel conditions prevalent in hotspot deployments and indoor propagation environments. By introducing the CLM and facilitating the cooperative exchange of QoS context information, it has been demonstrated that both QLO and LAS frameworks can provide adaptation to dynamic network conditions, which encompass traffic and wireless channel variations, in a self-adjusting manner to maintain a QoS-balanced system. Additionally, it has been shown that rate anomaly arising from link adaptation, which is used to combat varying wireless channel conditions, can also be mitigated by maintaining a QoS-balanced system without any elaborated modifications. It is worth mentioning that the QLO and LAS frameworks are compatible with both DCF and EDCA channel access mechanisms. Particularly, the possibility to augment the legacy DCF in providing a unifying QoS solution has significant advantages over the EDCA, owing to its prevalence.

This chapter has also highlighted that RQB has the salient intrinsic properties of: (i) mitigating rate anomaly in multirate environments; (ii) providing statistical QoS guarantee for multimedia traffic; (iii) precluding unnecessary handovers; and (iv) maximizing system capacity through the better use of radio resources, which collectively attains the end-to-end goal of promoting a QoS-balanced system. Hence, the notion of QoS balance has been advocated in this thesis to serve as the criterion for quantifying the state of balance in a multi-AP WLAN (multi-RAT environment) where network conditions are highly unpredictable. Furthermore, it has been argued in this thesis that the heterogeneity of multi-AP WLAN (multi-RAT environment) should be exploited, whenever possible, to orchestrate better utilization of radio resources.

In the next chapter, an elegant unified analytical model is developed to obtain the key performance metrics of MAC delay, PLR, and throughput efficiency for the IEEE 802.11 DCF infrastructure BSS. The analytical model integrates both Markov chain model and finite queueing model to capture non-saturation operating conditions. Additionally, it considers non-homogeneous conditions by modeling asymmetric traffic load between the AP and its associated STAs, heterogeneous flows between STAs, and heterogeneous wireless channel conditions between BSSs. These key performance metrics will serve as bounds for reliable capacity analysis from which a model-based PQB algorithm is developed.

CHAPTER 6

PERFORMANCE ANALYSIS OF THE IEEE 802.11 INFRASTRUCTURE BSS

To this end, an array of comprehensive simulation studies have demonstrated that the generalized CCRRM architecture is able to achieve the end-to-end goal of maintaining a QoS-balanced system, thanks to the measurement-based RQB algorithms. In particular, the novel concept of RQB has several intrinsic properties, which are favorable for the coordination of better radio resource utilization in future wireless networks, as shown in the previous chapters. In order to benchmark the performance of the RQB algorithm to gain further engineering insights, a unified analytical model will be developed in this chapter to obtain the key performance metrics of MAC delay, PLR, and throughput efficiency which serve as bounds for reliable capacity analysis of the IEEE 802.11 DCF infrastructure BSS. This unified analytical model will then be used to develop a model-based dynamic load distribution algorithm to function as a baseline for comparative studies with the measurement-based dynamic load distribution algorithms in Chapter 7, one of which is the RQB algorithm.

The throughput analysis of the IEEE 802.11 networks under saturation conditions is first introduced by Bianchi [131] and has since been extensively studied in the literature for an IBSS, also known as ad hoc networks, to support data communications with relaxed delay constraints. With the increasing popularity of RT services, e.g., VoWLAN, the QoS performance analysis of metrics such as MAC delay and PLR is becoming more important. The key reason is that these QoS metrics are particularly useful in the capacity analysis of delay-sensitive VoWLAN since they can be utilized as upper bounds when designing network control mechanism, e.g., admission or load control. In reality, STAs have to operate under non-saturation conditions in order to support delay-sensitive voice traffic and achieve maximum throughput [93]. Consequently, Bianchi's model which assumes saturation conditions cannot be directly applied to derive these QoS metrics for VoWLAN. To be more specific, the proper capacity analysis of VoWLAN lies in the ability to model the transition from the non-saturation to saturation mode (and vice-versa) which is critical for admission control as the quality of all voice connections will be compromised when the network capacity is exceeded. Furthermore, bulk of the existing IEEE 802.11 deployments in hotspots, enterprises, and campuses are configured as an infrastructure BSS with the basic access scheme of the DCF where STAs are associated with an AP. Although there is a plethora of analytical models developed for an IBSS, not many are devoted to an infrastructure BSS.

In this chapter, a simple, elegant unified analytical model is proposed to analyze the performance of the IEEE 802.11 DCF infrastructure BSS VoWLAN. This modeling approach follows closely to the works reported in [132] which incorporates retransmission limit to Bianchi's model and [133] which integrates Bianchi's model with standard queueing models to derive the QoS metrics of MAC delay and PLR, as well as throughput efficiency. Collectively, these are known as the performance metrics of an AP. The performance metrics of an AP are of particular interest as the AP relays all traffic to and from WLAN, and consequently it will be the capacity bottleneck of an infrastructure BSS. Hence, the AP typically operates in the saturation mode while STAs generally operate in the non-saturation mode due to the fact that the AP has much higher load than its associated STAs. In addition, the work presented in [134] is incorporated to augment the analytical model for operating under both ideal and error-prone channel conditions. Furthermore, the importance of modeling backoff freezing, i.e., freezing of backoff counter when medium is busy, as in [135] and [136] for an infrastructure BSS, and the consequences if ignored or improperly modeled are exhaustively discussed.

This chapter is outlined as follows. Section 6.1 gives a comprehensive overview of the related works in the performance analysis of the IEEE 802.11 networks and discusses the modeling approach of the unified analytical model. Section 6.2 presents the mathematical analysis which composes of the Markov model analysis, average MAC service time analysis, and queueing model analysis to achieve the key performance metrics of MAC delay, PLR, and throughput efficiency for the AP of an infrastructure BSS. Section 6.3 validates the proposed analytical model with OPNET simulations. Section 6.4 investigates how the performance of the unified analytical model is influenced by various traffic sources, PHYs, and data rates. Additionally, the effect of backoff freezing for an infrastructure BSS and IBSS is established, and the consequences if ignored or improperly modeled are revealed. Finally, Section 6.5 provides the conclusions of this chapter.

6.1 Overview on the Performance Analysis of the IEEE 802.11 Networks

The two-dimensional Markov chain model first proposed by Bianchi in [131] evaluates the saturation throughput of the IEEE 802.11 IBSS with the basic access scheme of the DCF under the assumptions of infinite retransmissions and ideal channel conditions. Subsequently, there are many enhancements to Bianchi's model. However, most of them are carried out independently to address a specific issue. Wu *et al.* [137] and later Chatzimisios [132] modify Bianchi's model by introducing a limit on the number of retransmissions, i.e., introducing a maximum number of backoff stages which enables new performance metrics such as packet dropping probability and average time to drop a packet to be derived and studied. Chatzimisios *et al.* [138] also improve Wu's model to account for transmission failures but only for data frames. Ni *et al.* [134] then amend Chatzimisios's model to consider transmission failures in both data and ACK frames. Bianchi and Tinnirello [139] later provide a new approach to derive the performance metrics of throughput and delay by relying on elementary conditional probability instead of the original Markov chain so that it can address generic backoff procedures.

Bianchi's model is accurate but implicitly assumes that the backoff counter is decremented at the beginning of a slot time, regardless of whether medium is busy or idle, which does not conform to the IEEE 802.11 standard. Ziouva and Antonakopoulos [135] first notice this discrepancy and introduce backoff decrement probability to account for the freezing of backoff counter when medium is busy so that the backoff counter is decremented only during idle slots. Subsequently, Xiao [140], Ergen [141], Engelstad and Osterbo [142], and Yang *et al.* [143] adopt backoff freezing in their analytical models as in [135]. However, the authors of [132] and [140] find that Ziouva's model is still not entirely compliant to the IEEE 802.11 standard as it includes an additional non-backoff state which allows STAs to transmit without entering the backoff stage. After omitting the non-backoff stage, Foh and Tantra [136] show that the unconditional backoff decrement probability of Ziouva's model causes inaccuracies in saturation throughput, particularly, when contention is increased. Recently, Tinnirello *et al.* [144] acknowledge these open issues in the modeling of backoff freezing and its resumption process, and hence introduce an augmented Markov chain to address backoff freezing properly in an IBSS under saturation conditions.

On the other hand, the performance analysis under non-saturation conditions has been receiving much attention in recent times. Most of the reported works have extended Bianchi's model to include: (i) additional states; (ii) post-backoff states based on bufferless assumption; or (iii) the probability of non-empty buffer which is derived from standard queueing models as a suitable scaling factor of saturation conditions. Ergen [141] extends Bianchi's model by introducing additional idle states to model non-saturation scenarios, which are geometrically distributed with parameter λ. However, according to Malone *et al.* [145], Ergen's model does not allow for the packet inter-arrival and IEEE 802.11's post backoff period to overlap, and the relationship between λ and the system load is not given. Instead, Malone *et al.* model non-saturation conditions by incorporating post-backoff states under the bufferless network assumption. In addition, they consider heterogeneous STAs with different arrival rates but same packet lengths. Cantieni *et al.* [146] also account for post-backoff states and consider heterogeneous STAs in their generalized model to support different modulation rates for addressing the rate anomaly problem of DCF. Dao and Malaney [147] consider the probability of immediate transmission when medium is idle after a distributed (coordination function) interframe space (DIFS) duration has elapsed, in addition to post-backoff states and the probability of a new packet arrival immediately after post-backoff. They also point out that the post-backoff of Malone's model is performed only when the queue is empty, whereas the IEEE 802.11 standard mandates post-backoff to be performed after the end of each transmission. Moreover, the probability of a new packet arrival immediately after post-back off in Malone's model holds only for the bufferless assumption which is not realistic in practice. However, the authors do not consider finite retransmission limit in their model. Zhao *et al.* [148] show that although Malone's model provides the most accurate collision probability prediction, it suffers from poor mean MAC delay accuracy. Recently, Ghaboosi *et al.* [149] and Liu *et al.* [150] introduce a new paradigm to integrate the DCF contention resolution and queueing processes into a single model, which comes at the expense of increased computational complexity.

In a different light, the standard queueing models have been used in conjunction with the random backoff process analytical model, to analyze non-saturation performance, based on: (i) Bianchi's Markov model; (ii) Tay and Chua's [151] non-Markovian model; or (iii) Markovian renewal-reward model of Kumar *et al.* [152] which dispense from analyzing the Markov chain as in [131]. Accordingly, Zhai *et al.* [133] integrate Bianchi's model with the $M/M/1/K$ and $M/G/1/K$ models to give non-saturation throughput, delay, and loss bounds. On the other hand, Tickoo and Sikdar [153] extend Tay and Chua's model and consider individual STA queues as the $G/G/1$ model to allow for generic arrival process. Cai *et al.* [154] first modify Tickoo's model to analyze the asymmetric traffic situation of an infrastructure BSS. Medepalli and Tobagi [155] model the random backoff process using a renewal-reward model, which is similar to that of Kumar's model, and integrate it with (i) the $M/G/1/PS$ model to model the DCF from a system-centric perspective; and (ii) the $G/G/1$ model to represent user's queue from a user centric perspective. Zhao *et al.* [148] extend Kumar's model to obtain the performance metrics of throughput and MAC delay under the Poisson traffic and bufferless assumptions.

To this end, most of the existing throughput analyses have been dedicated to an IBSS, and often, under saturation conditions. As previously explained, such analyses cannot be directly applied to an infrastructure BSS VoWLAN as STAs have to operate under non-saturation conditions in order to support delay-sensitive voice applications. Although backoff freezing in the DCF and EDCA has been largely studied for an IBSS under the assumptions of saturation conditions and homogeneous STAs, it has not received equal attention for a finite load infrastructure BSS. The remainder of this chapter concentrates on the performance analysis of an infrastructure BSS VoWLAN operating under non-homogeneous conditions[1] and spanning across the non-saturation to saturation modes. Furthermore, the importance of backoff freezing for an infrastructure BSS, especially, during the transition from the non-saturation to saturation mode will be highlighted.

6.1.1 Modeling Approach: Obtaining Performance Metrics

The measure of effectiveness to a given process such as packet transmission is often associated with either the time a packet spends in the queue or the total time a packet spends in the system, i.e., service time plus queue delay. Depending on the system under investigation, one may be of more interest than the other. E.g., although service time of a packet suffices when evaluating the MAC layer performances, it is the total delay that

---

[1] The term non-homogeneous conditions is used throughout this thesis to refer to: (i) asymmetric traffic load between an AP and its associated STAs of an infrastructure BSS; (ii) heterogeneous STAs in which traffic load are generated by STAs with different voice codecs; and (iii) heterogeneous wireless channel conditions due to channel impairments such as NLOS transmission, propagation characteristics, thermal noises, and interferences from other radio sources.

would determine the QoS of end-users when dealing with packets sent from a higher layer application, e.g., VoIP. Most of the existing works analyze the time that a packet spends in contention for medium access until it is successfully received, also known as the MAC service time, without considering the time spend in the queue. This is possible since these analytical models are derived under the assumption of saturation conditions which allows the abstraction of queueing dynamics from their analysis [145]. However, in reality, the IEEE 802.11 networks do not typically operate under saturation conditions. E.g., most of the data traffic such as web browsing and email are bursty in nature. Moreover, VoIP traffic operates at relatively low packetization intervals and normally with silence suppression of significant silence periods. Hence, it is imperative to model these non-saturation behaviors in order to analyze the load conditions at which the network will become saturated.

In recent years, a vast amount of analytical models have been developed to account for the non-saturation operations of STAs, following the seminal work of [93] which reveals that the IEEE 802.11 networks operate optimally only under non-saturation conditions. However, most of these analytical models [141], [142], [147], [148], and [153] are designed for an IBSS under the assumption of homogeneous STAs with same traffic flow, albeit, [145] considers two classes of heterogeneous STAs in their work. Apart from [143], [146], [154], [156], [157], [158], [159], and [160], not many analytical models are developed for an infrastructure BSS. Out of these works, only [146], [154], [156], [158], and [159] consider the non-saturation conditions of STAs. However, [146], [154], and [156] are based on either the $M/G/1$ or $G/G/1$ model which may not be appropriate in practice due to the assumption of infinite queue length. In addition, the expressions in [156] for deriving the collision probability, probability of busy medium, and probability of successful transmission are not properly generalized to account for the heterogeneity between the AP and STAs in which the AP is the capacity bottleneck. The adoption of infinite queueing model is motivated by the fact that finite queueing models, other than the $M/M/1/k$ model, require the explicit knowledge of service time distribution. However, the DCF contention process results in an unknown complex service time distribution where only approximation of its first moment can be readily obtained and then applied to infinite queueing models [146]. It is important to note that the assumption of infinite queue implicitly limits the analysis to only the non-saturation mode since the $G/G/1$ model is stable only when traffic intensity $\rho < 1$. In other words, although these models aim to provide non-saturation analysis, they fail to capture the transition from the non-saturation to saturation mode (and vice-versa) which is of prime importance in the capacity analysis of VoWLAN.

Although the model in [158] considers heterogeneous STAs under non-saturation conditions, it is based on a three-dimensional continuous-time Markov chain with higher computation complexity. Furthermore, the authors provide only numerical results which lack the validation of numeric accuracy with simulation results. The most notable work is that of [159] which utilizes the saturation analysis of [131] and [152] to model a finite load infrastructure BSS for both transmission control protocol (TCP) and VoIP applications. In particular, the authors approximate the voice packet arrivals with a binomial distribution where the probability that a voice call generates a packet in an interval of l slots is modeled as $p_l = 1 - (1 - \lambda)^l$. This acts as a scaling factor to the saturation attempt probability to account for the non-saturation conditions of STAs. It is worth noting that this concept is similar to that of [133], [148], [153], and [154]. The binomial approximation works well in the non-saturation mode where the queue of STAs is typically empty, but it underestimates the attempt rates in the saturation mode as the number of STAs increases. Although the authors give the capacity estimation for both homogeneous and heterogeneous voice codecs, they do not explicitly derive the performance metrics such as MAC delay, PLR, and throughput efficiency. Furthermore, they exclude VBR traffic sources and error-prone channel conditions from their analysis. Moreover, most of the above-mentioned analytical models for an infrastructure BSS, except for [156], do not consider backoff freezing which has a significant influence on these performance metrics after the onset of saturation.

The motivations for the modeling approach in what follows are based on a rigorous survey of existing analytical models which have been independently developed to address different issues. From the discussions of these prior works, it is apparent that there are many intricacies involved in developing an analytical model to account for both non-saturation and saturation modes under non-homogeneous conditions. More importantly, the transition from the non-saturation to saturation modes is of particular interest in the capacity analysis of VoWLAN as the QoS of existing voice connections will be compromised as soon as the AP saturates. As previously explained, the total time that a VoIP packet spends in the MAC layer before it is successfully received, also known as the MAC delay, is crucial for the perceived QoS of end-users. The MAC delay includes: (i) the time duration between when a packet is first inserted into the transmission queue and when it becomes the HOL packet, i.e., queueing delay; and (ii) the time duration between when the HOL packet starts its contention process for medium access and when it is successfully received, i.e., MAC service time. According to the works reported in [132] and [139], a packet which is dropped after exceeding the maximum retry limit does not contribute to the MAC service time computation since it is not successfully received. Furthermore, the effect of a finite queue which is commonly employed in practice needs to be accounted. Hence, a packet could be dropped either: (i) *before* entering the contention process for medium access due to a full queue; or (ii) *during* the contention process for medium access after exceeding the maximum retry limit. The fraction of the total packets that is dropped before and during the contention process is known as the PLR.

In essence, the performance analysis of the IEEE 802.11 DCF infrastructure BSS can be classified into three regions as illustrated in Figure 6.1. It shows a simulation result reflecting the departure rate (or throughput) of an AP with different number of homogeneous STAs contending for medium access using equal arrival rates. In the first region, all the generated traffic load is successfully transmitted as the medium is in the non-saturation mode. This non-saturation region is of little interest since both MAC delay and PLR will be very low, and the QoS requirements of STAs will be effortlessly met. In the second

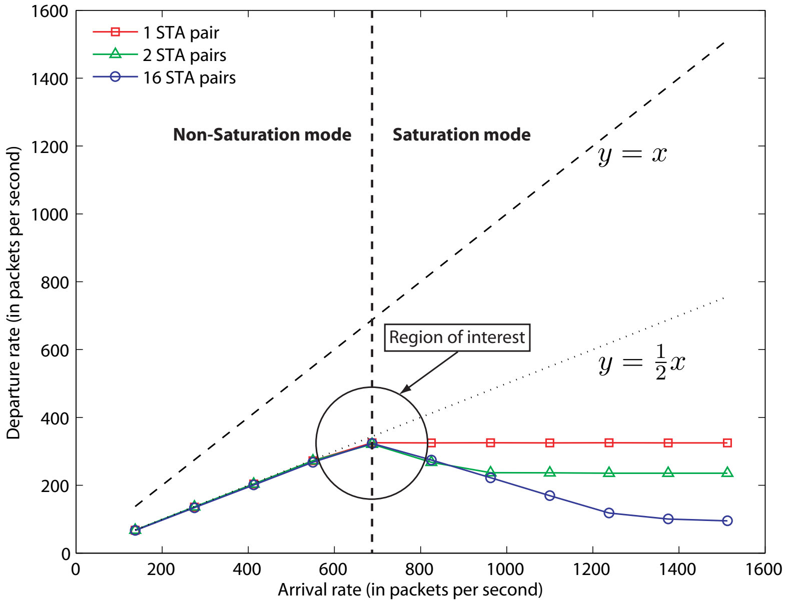


Figure 6.1: Region of capacity analysis: Transition from the non-saturation to saturation mode of an AP in an infrastructure BSS.

region, the saturation throughput of the AP varies with the different number of contending STAs. The analysis in this saturation region is of no practical use to VoIP traffic since the MAC delay and PLR constraints are violated, and thus no realistic communication is possible. In the third region, the throughput of the AP falls below the dotted line $y = \frac{1}{2}x$, as traffic load increases. This signifies that the AP is transiting into the saturation mode and only a fraction of generated load is successfully transmitted as the queue is always full. Note that for an infrastructure BSS VoWLAN with n STA pairs engaging in 2-way voice communications, the AP transmits $\frac{1}{2}$ of the voice traffic in the DL while each STA transmits only $\frac{1}{2n}$ of the voice traffic in the UL. Thus, the throughput of the AP will increase linearly along the dotted line $y = \frac{1}{2}x$ while the aggregate throughput will increase along the dashed line $y = x$, until the maximum throughput is reached in both cases. Given that the AP and STAs have the *same* priority to medium access with the DCF, the AP which has a significantly higher load will become the capacity bottleneck.

It is now clear that the analysis of this transition region is of great importance since accurate performance metrics obtained within this regime will be useful for admission control and capacity analysis in VoWLAN. Particularly, when the generated traffic load exceeds the network capacity, collision probability will increase and the frame service rate will correspondingly reduce, which in turn increases the queue length and MAC delay. As a result, the QoS performance of STAs will be immediately affected by the significant MAC delay due to the backlog queue. This will eventually cause excessive PLR when the AP starts dropping packets before they have a chance to enter the contention process for medium access. Hence, the knowledge of network capacity in VoWLAN is critical because it can predict the upper bound of admissible traffic load that the network can support with acceptable QoS for delay-sensitive voice traffic, especially, under non-homogeneous conditions.

6.2 Mathematical Analysis

In this section, a unified analytical model is proposed for a more realistic performance analysis of the IEEE 802.11 DCF infrastructure BSS in terms of MAC delay, PLR, and throughput efficiency under non-homogeneous conditions as illustrated in Figure 6.2. The random backoff process is first modeled using a discrete time Markov chain which is solved numerically to obtain the transmission and its failure probabilities. Subsequently, closed-form expressions for the average MAC service time are derived from the random backoff process based on the transmission failure probability. This average MAC service time, i.e., $1/\mu$ is then used in conjunction with the $M/M/1/K$ model to obtain the average MAC delay, PLR, and throughput efficiency for each STA and the AP. Although the traffic generated by each STA is assumed to follow a Poisson process, it does not dilute the performance accuracy of VoIP traffic with ON-OFF periods which is typically characterized as a Markov modulated Poisson process [161].

The key idea of the modeling approach is motivated by the work reported in [155] and extended from the work described in [133]. In the former, the authors give an alternative system-centric approach to analyze the total delay of the IEEE 802.11b infrastructure

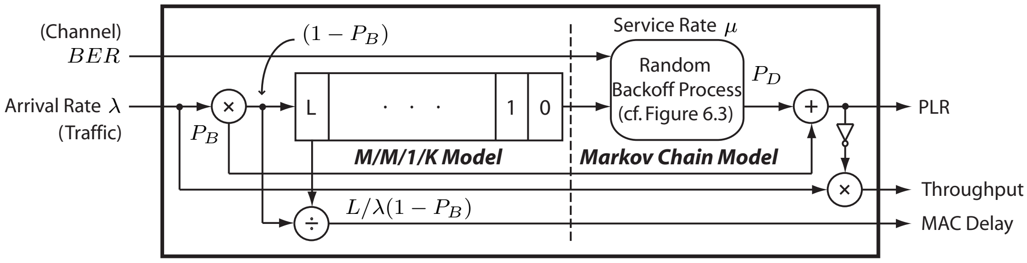


Figure 6.2: Unified analytical model: Markov chain model in conjunction with finite queueing model.

BSS by modeling the DCF as a central server due to its throughput fairness property in which contending STAs access the channel resources in a round robin manner. They also reveal that for Poisson arrivals, the resource sharing model reduces to a $M/G/1/PS$ system where the average queueing delay and service time are actually the same as the $M/M/1$ system, regardless of service time distribution [162]. In the latter, the authors engage the finite queueing models to obtain the performance metrics of MAC delay, PLR, and throughput for an IBSS. They show that although the $M/G/1/K$ model provides better accuracy than the $M/M/1/K$ model, in general, they do not exhibit significant differences which are in good agreement with the findings of [155] for the infinite queue case.

In order to obtain the key performance metrics of MAC delay, PLR, and throughput efficiency for the IEEE 802.11 DCF infrastructure BSS VoWLAN under non-homogeneous conditions, several extensions to the existing analytical models are necessary. To be more specific, Zhai's model [133] is extended to reflect the asymmetric load situation of VoWLAN as in [154]. In addition, traffic variability between WLAN STAs is introduced by considering heterogeneous voice codecs of different packetization intervals and packet lengths. Furthermore, wireless channel variability between BSSs is considered by factoring in transmission failures in both MAC data frame and ACK frame as in [134]. More importantly, backoff freezing during the times when medium is busy is modeled according to [135] without including the non-backoff stage. To the best of the author's knowledge, there is no prior analytical model that offers the performance analysis from a unified perspective in order to pragmatically capture non-homogeneous operating conditions which span across both non-saturation and saturation modes. The key contributions of this chapter are twofold. First, the proposed analytical model collectively accounts for: (i) asymmetric traffic load between the AP and its associated STAs of an infrastructure BSS; (ii) transition from the non-saturation to saturation mode (and vice-versa) from the AP perspective; (iii) heterogeneous traffic flows between STAs; and (iv) heterogeneous wireless channel conditions between BSSs of a multi-AP hotspot scenario, all in a single unifying framework. Second, extensive analyses and simulations have unveiled that the improper modeling or ignorant of backoff freezing for an infrastructure BSS will result in overly conservative bounds which will lead to low network utilization when deployed as admission control, particularly, in heavy load scenarios.

6.2.1 Markov Chain Model Analysis

The analysis in what follows is based on similar assumptions as in [131], [132], [133], and [134] to model the random backoff process of the IEEE 802.11 infrastructure BSS under non-homogeneous conditions with the basic access scheme of the DCF and finite retransmission limit. In order to consider the non-homogeneous conditions of varying error-prone channel and traffic load in the same analysis, a new transmission failure (or unsuccessful transmission) probability P_f is introduced, in addition to the conditional collision probability P_c. More specifically, P_c is the probability of a collision as seen by a packet being transmitted on the medium. It is observed that P_c is also the probability that there is at least one packet transmission in the medium among other $(n-1)$ STAs as seen by the tagged STA[2]. It is known that with the DCF, a transmitter cannot distinguish between a collision and frame error based on the positive ACK since ACK frame will not be received in both cases. In particular, a collision occurs when two or more STAs simultaneously transmit while a frame error occurs when either a MAC data frame or ACK frame is corrupted due to channel noises. As a result, the occurrence of either a collision or frame

---

[2] The term tagged STA is used throughout this chapter to refer to a particular STA under consideration.

error event is treated in the same manner by the DCF as an unsuccessful transmission, and therefore retransmission procedures will be invoked. Hence, P_f is the probability of unsuccessful transmission, in which collision and frame error are considered as two independent events, as seen by a packet being transmitted on the medium. Similarly, the corruption of MAC data frame and ACK frame as a result of channel noises are regarded as two independent events where the occurrence of either event will constitute a frame error. Again, this is due to the fact that ACK frame will not be received by the transmitter in both cases and retransmission procedures will be invoked.

The key assumption of this analysis is that the collision probability P_c and transmission failure probability P_f of a packet transmission remain constant and are independent of the number of previous retransmissions. It is intuitive that this assumption yields better accuracy with the increasing CW size W and number of STAs n. This simplifying assumption is justified in [163] where an exhaustive analysis of the Markov chain associated with the back-off process of the contending STAs is performed without making use of the decoupling assumption rooted in the work of [131] and other related works such as [152]. The freezing of backoff counter when medium is busy is also modeled as in [135]. It will be shown later that backoff freezing has important implications on the performance metrics of an infrastructure BSS, particularly, in the region of interest. The key assumptions in this analysis are summarized as follows:

- Collision probability P_c and transmission failure probability P_f of a packet transmission remain constant, and they are independent of the number of previous retransmissions.

- An AWGN wireless channel is considered. Thus, each bit has the same bit error probability, and bit errors are i.i.d. over the entire frame.

- Link adaptation and the effects of distance are ignored. Hence, STAs have fixed PHY data rate and the same BER, respectively.

- No hidden terminals are considered (cf. Appendix A-2.4 for details). Therefore, collisions will occur only in MAC data frames but not in ACK frames.

Modeling Packet Transmission Probability

A discrete time Markov chain in Figure 6.3 is used to study the random backoff behavior of any STAs by modeling it as a two-dimensional process $\{s(t), b(t)\}$ where $s(t)$ and $b(t)$ are stochastic processes representing the backoff stage and backoff time counter, respectively of the tagged STA at time t. The CW size of different backoff stages $i \in [0, m]$ can be expressed as

$$W_i = \begin{cases} 2^i W_0 = 2^i \left(CW_{\min} + 1\right), & i \in [0, m'] \\ 2^{m'} W_0 = CW_{\max} + 1, & i \in [m', m] \end{cases} \tag{6.1}$$

where W_i is the current CW size, $CW_{\min}$ is the minimum CW size, W_0 is the initial CW size, m' is the maximum CW increasing factor, $CW_{\max}$ is the maximum CW size, and m is the retry limit which is also the maximum backoff stage. The value of W_i depends on the number of transmission failures encountered by a packet (cf. Figure A-3 of Appendix A-2). From the discrete time Markov chain, the only non-null one-step transition probabilities[3] are

$$\begin{cases} P\{i, k\,|i, k+1\} = 1 - P_c, & k \in [0, W_i - 2], i \in [0, m] \\ P\{i, k\,|i, k\} = P_c, & k \in [1, W_i - 1], i \in [0, m] \\ P\{0, k\,|i, 0\} = (1 - P_f)/W_0, & k \in [0, W_0 - 1], i \in [0, m-1] \\ P\{i, k\,|i-1, 0\} = P_f/W_i, & k \in [0, W_i - 1], i \in [1, m] \\ P\{0, k\,|m, 0\} = 1/W_0, & k \in [0, W_0 - 1] \end{cases} . \tag{6.2}$$

---

[3] Short notation is adopted as in [131] to express the transition probabilities:

$$P\{i_1, k_1\,|i_0, k_0\} = P\{s(t+1) = i_1, b(t+1) = k_1\,|s(t) = i_0, b(t) = k_0\}$$

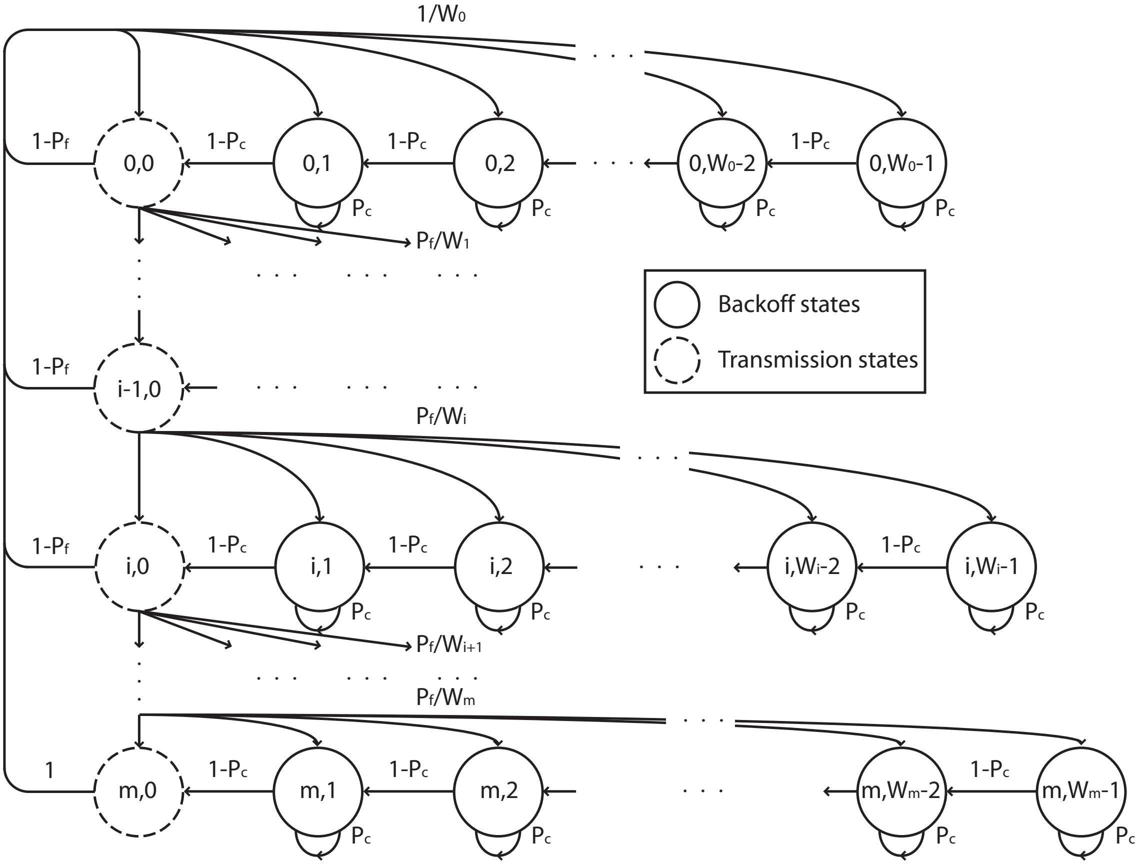


Figure 6.3: Discrete time Markov chain transition diagram.

These five transition probabilities account for: (i) the decrement of backoff timer when medium is idle; (ii) the freezing of backoff timer when medium is busy; (iii) the backoff timer that will always start from backoff stage 0 after a successful transmission; (iv) the backoff timer that starts from a new backoff stage of increasing order after an unsuccessful transmission which could be due to either a collision or frame error; (v) the CW that will always reset and the backoff timer that will always start from backoff stage 0 when maximum retransmission limit is reached, regardless of whether the transmission is successful or not.

Similar to [131], let $b_{i,k} = \lim_{t\to\infty} P\left\{s\left(t\right) = i, b\left(t\right) = k\right\}$, $i \in \left[0, m\right], k \in \left[0, W_i - 1\right]$ be the stationary distribution of the Markov chain. First, it is noted that

$$b_{i-1,0}.P_f = b_{i,0} \Rightarrow b_{i,0} = P_f^i b_{0,0}, \quad i \in \left[0, m\right]. \tag{6.3}$$

Owing to chain regularities, for each $k \in [1, W_i - 1]$, it is

$$b_{i,k} = \frac{W_i - k}{W_i\left(1 - P_c\right)}. \begin{cases} \left(1 - P_f\right)\sum_{j=0}^{m-1} b_{j,0} + b_{m,0}, & i = 0 \\ P_f.b_{i-1,0}, & i \in [1, m] \end{cases}. \tag{6.4}$$

By means of (6.3) and realizing the fact that $\left(1 - P_f\right)\sum_{j=0}^{m-1} b_{j,0} + b_{m,0} = b_{0,0}$, (6.4) can be simplified to

$$b_{i,k} = \frac{W_i - k}{W_i\left(1 - P_c\right)} b_{i,0} + b_{i,0}, \quad k \in \left[0, W_i - 1\right], i \in \left[0, m\right]. \tag{6.5}$$

Finally, $b_{0,0}$ can be determined by imposing the normalization condition together with (6.1), (6.3), and (6.5) as

$$\begin{aligned} 1 &= \sum_{i=0}^{m}\sum_{k=0}^{W_i-1} b_{i,k} = \sum_{i=0}^{m} b_{i,0} \sum_{k=1}^{W_i+1} \frac{W_i - k}{W_i\left(1 - P_c\right)} + 1 \\ &= \sum_{i=0}^{m} b_{i,0}\frac{W_i - 1}{2\left(1 - P_c\right)} + 1 = \sum_{i=0}^{m} b_{i,0}\frac{W_i + 1 - 2P_c}{2\left(1 - P_c\right)}, \end{aligned} \tag{6.6}$$

from which

$$b_{0,0} = \begin{cases} \frac{2\left(1-P_c\right)\left(1-2P_f\right)\left(1-P_f\right)}{\Theta}, & m \le m' \\ \frac{2\left(1-P_c\right)\left(1-2P_f\right)\left(1-P_f\right)}{\Phi}, & m > m' \end{cases} \tag{6.7}$$

where

$$\begin{aligned} \Theta &= W_0\left(1 - \left(2P_f\right)^{m+1}\right)\left(1 - P_f\right) + \left(1 - 2P_c\right)\left(1 - P_f^{m+1}\right)\left(1 - 2P_f\right), \\ \Phi &= W_0\left(1 - \left(2P_f\right)^{m'+1}\right)\left(1 - P_f\right) + \left(1 - 2P_c\right)\left(1 - P_f^{m+1}\right)\left(1 - 2P_f\right) \\ &+ 2^{m'} W_0\left(P_f^{m'+1}\right)\left(1 - P_f^{m-m'}\right)\left(1 - 2P_f\right). \end{aligned} \tag{6.8}$$

Since transmission occurs when backoff time counter reaches zero (in the transmission states), regardless of backoff stage, the probability of transmission τ that a STA transmits in a randomly chosen slot time on the condition that the STA has packets to transmit can be derived as

$$\tau = \sum_{i=0}^{m} b_{i,0} = \frac{1 - P_f^{m+1}}{1 - P_f} b_{0,0}, \tag{6.9}$$

which can be simplified to

$$\tau = \begin{cases} \frac{2(1-P_c)\left(1-2P_f\right)\left(1-P_f^{m+1}\right)}{\Theta}, & m \leq m' \\ \frac{2(1-P_c)\left(1-2P_f\right)\left(1-P_f^{m+1}\right)}{\Phi}, & m > m' \end{cases}. \tag{6.10}$$

Note that for $P_c = 0$, (6.10) reduces to the model of [134] which does not consider backoff freezing. From (6.10), the probability of transmission τ depends on the collision probability P_c and transmission failure probability P_f which are still unknown.

Now, consider the case of n STAs where the per-STA quantities are subscripted with the STA label $a = 1, \ldots, n$. To compute P_{c_a}, each packet transmitted by the tagged STA is assumed to have a constant and independent collision probability. Accordingly, the probability that medium is idle as seen by the tagged STA is

$$1 - P_{c_a} = \prod_{b \neq a} \left[1 - \left(1 - P_{0_b}\right) \tau_b\right] \tag{6.11}$$

where P_{c_a} is the collision probability as seen by the tagged STA. τ_b is the packet transmission probability that other STAs transmit in a randomly chosen slot time given that they have packets to transmit. $1 - P_{0_b}$ is the probability that other STAs have a non-empty queue by assuming that they can be modeled as a finite queue as in [133]. Essentially, $1 - P_{0_b}$ functions as a scaling factor of τ_b in the saturation mode by assuming that τ_b in the non-saturation mode is proportional to $1 - P_{0_b}$. The subscripts a and b reflect the non-homogeneous network model [145] where the traffic generated by each STA and wireless channel conditions between BSSs may be different, and the fact that the AP of an infrastructure BSS has much higher traffic load than its associated STAs. In other words, (6.11) implies that when STAs are heterogeneous, their collision probabilities will be different unless they have equal transmission probabilities. As a result, (6.11) reduces to $1 - p = (1 - \tau)^{n-1}$ for the case of homogeneous STAs in the saturation mode presented in [131], [132], and [134], and $1 - p = [1 - (1 - P_0)\,\tau]^{n-1}$ for the case of homogeneous STAs in the non-saturation mode shown in [133].

Similarly, to compute P_{f_a}, each packet transmitted by the tagged STA is assumed to have a constant and independent failure probability. A transmission failure is deemed to occur when either a collision or frame error happens by assuming collision and frame error as two independent events. It then follows that the transmission failure probability as seen by the tagged STA is

$$P_{f_a} = 1 - (1 - P_{c_a})\,(1 - FER)\,. \tag{6.12}$$

In addition, the data frame error FER^{data} and ACK frame error FER^{ack} are assumed as two independent events where FER, the frame error rate (FER) of either a MAC data frame or an ACK frame, is given by

$$FER = 1 - \left(1 - FER^{data}\right)\left(1 - FER^{ack}\right). \tag{6.13}$$

Given that the bit errors are uniformly distributed over the entire frame, FER^{data} and FER^{ack} can be calculated as

$$\begin{cases} FER^{data} = 1 - (1 - BER)^{L_{DATA}} \\ FER^{ack} = 1 - (1 - BER)^{L_{ACK}} \end{cases} \tag{6.14}$$

where the probability of bit error BER for different modulation schemes can be obtained from OPNET's empirical modulation curves [164] or found by using the Q-function of the distance between signal points in the constellation diagram which is extensively discussed in [165] and reported in [134]. Note that STAs have the same FER as a result of the same BER and (6.24).

For n STAs, (6.10) gives an expression for the per-STA transmission probability τ_a where $a = 1, \ldots, n$ is the STA label. Hence, (6.10) and (6.12) form $2n$ coupled non-linear equations which can be solved numerically by fixed point iteration technique [166] for $P_{f_1}, \ldots, P_{f_n}$ and $\tau_1, \ldots, \tau_n$.

6.2.2 Average MAC Service Time Analysis

The MAC service time is defined in [133] as a non-negative random variable T_S. Accordingly, the probability generating function of T_S is

$$P_{T_S}(Z) = \sum_{i=0}^{\infty} p_i Z^{t_{si}} = p_0 Z^{t_{s0}} + p_1 Z^{t_{s1}} + p_2 Z^{t_{s2}} + \ldots \tag{6.15}$$

which has a discrete probability of p_i for T_S where t_{si} is the unit of one-bit transmission time or the smallest system clock. The motivation for deriving the probability generating function is to conduct subsequent queueing analysis based on the *average* MAC service time $E[T_S]$. However, the computation of $E[T_S]$, in this case, involves finding the first derivative of a fairly complex probability generating function that is computationally expensive. This technique of computing the average MAC service time is first proposed in [167], then in [133], and subsequently adopted in [142], [153], and [168]. Here, a simpler approach of deriving the average MAC service time with closed-form expressions is presented.

First, it is observed that the duration of each backoff state in the Markov chain is a random variable. More specifically, each backoff state could be occupied by one of the five virtual events with the corresponding time slot duration of: (i) successful transmission T_{s_a}; (ii) unsuccessful transmission with ACK frame error $T_{e_a}^{ack}$; (iii) unsuccessful transmission with collision T_{c_a}; (iv) unsuccessful transmission with data frame error $T_{e_a}^{data}$; and (v) idle slot T_{idle}, according to a discrete and non-uniform slotted time scale. Although this analysis considers the basic access scheme of the DCF, it can be easily extended to incorporate the four-way handshake procedure of the RTS/CTS mechanism. It is important to note

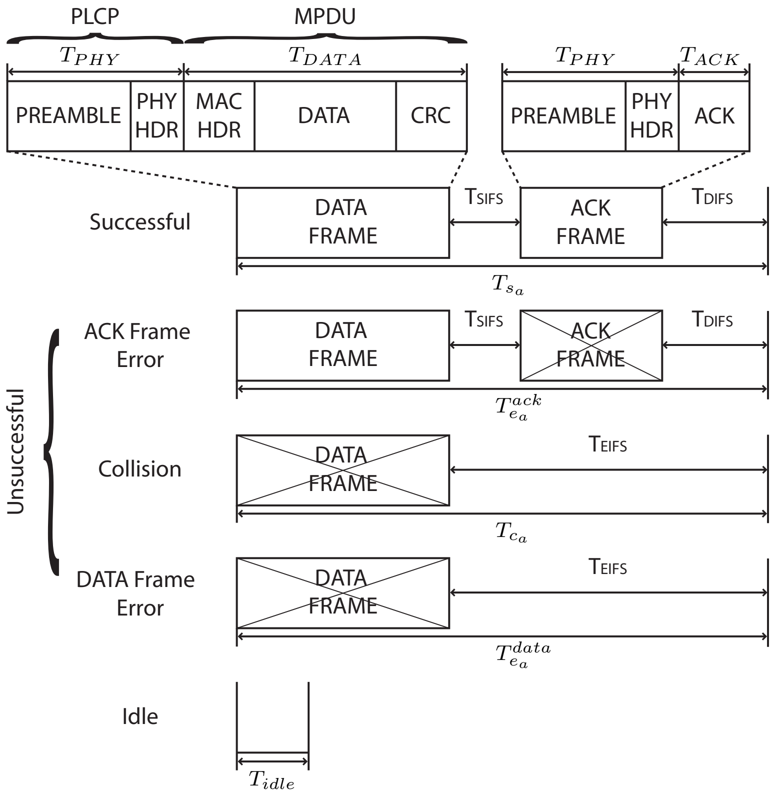


Figure 6.4: Channel states duration.

that voice frames are typically transmitted using the basic access scheme for reducing overheads due to their small payload size as in [154] and [159]. Furthermore, one voice packet corresponds to one MAC frame without link layer fragmentation. Accordingly, the five different time slot durations as depicted in Figure 6.4 for basic access scheme are

$$\begin{cases} T_{s_a} = 2T_{PHY} + T_{DATA_a} + 2\delta + T_{SIFS} + T_{ACK} + T_{DIFS} \\ T_{e_a}^{ack} = T_{s_a} \\ T_{c_a} = T_{PHY} + T_{DATA_a} + \delta + T_{EIFS} \\ T_{e_a}^{data} = T_{c_a} \\ T_{idle} = \sigma \end{cases} \tag{6.16}$$

where

$$T_{EIFS} = T_{SIFS} + T_{PHY} + T_{ACK} + \delta + T_{DIFS}. \tag{6.17}$$

T_{PHY} is the duration of physical layer convergence procedure (PLCP) overheads, T_{DATA_a} is the expected time taken by the tagged STA to transmit a data frame including MAC overheads, δ is the propagation delay, and σ is a PHY-dependent slot time. Note that σ, T_{SIFS}, T_{DIFS}, and T_{EIFS} are defined in the IEEE 802.11 standard. From (6.16), $T_{e_a}^{data} = T_{c_a}$ for the DCF as transmitter cannot differentiate between collisions and data frame errors with positive ACK. On the other hand, $T_{e_a}^{ack} = T_{s_a}$ as other STAs can still correctly decode the duration field from the successfully received data frame even though the ACK frame is in error.

In order to compute the expected length of backoff slot time, it is necessary to determine the probabilities that correspond to the five different time slot durations. The probability of an idle slot is defined in (6.11). The probability of a successful transmission in a time slot occurs only when one STA transmits an error-free data frame, as well as an error-free ACK frame. It is immediate from (6.11) and (6.13) that the probability of a successful transmission as seen by the tagged STA is

$$P_{s_a} = \sum_{c \neq a} (1 - P_{0_c})\tau_c \prod_{b \neq a,c} \left[1 - (1 - P_{0_b})\,\tau_b\right](1 - FER)\,, \tag{6.18}$$

and it immediately follows that the probability of a collision in a time slot is

$$P_{col_a} = P_{c_a} - P_{s_a} - P_{e_a}^{data} - P_{e_a}^{ack}. \tag{6.19}$$

Next, the probability of a data frame error in a time slot occurs only when one STA transmits and the data frame is in error which is seen by a tagged STA as

$$P_{e_a}^{data} = P_{s_a}.FER^{data}. \tag{6.20}$$

The probability of an ACK frame error in a time slot occurs when a data frame is successfully transmitted but the ACK frame is in error which is seen by the tagged STA as

$$P_{e_a}^{ack} = P_{s_a}\left(1 - FER^{data}\right) FER^{ack}. \tag{6.21}$$

Now, the expected length of a backoff slot time can be expressed as

$$E\left[slot_a\right] = \left(1 - P_{c_a}\right)\sigma + T_{s_a}\left(P_{s_a} + P_{e_a}^{ack}\right) + T_{c_a}\left(P_{col_a} + P_{e_a}^{data}\right). \tag{6.22}$$

In the first approximation, it is noted that $E\left[slot_a\right]$ can be rewritten as

$$E\left[slot_a\right] = \left(1 - P_{c_a}\right)\sigma + T_c P_{c_a} \tag{6.23}$$

if T_{s_a} and T_{c_a} of the tagged STA in (6.22) are equal. This will hold for STAs with homogeneous traffic flows operating with the basic access scheme of the DCF without the RTS/CTS mechanism in an IBSS. However, VoWLAN is typically configured as an infrastructure BSS in a wireline-to-wireless topology where BSS consists of one AP, $N-1$ WLAN STAs, and $N-1$ ethernet STAs which are connected through a wireline backbone. In such a scenario, the traffic load flowing through the AP is $N-1$ times that of a WLAN STA when considering 2-way voice conversations between WLAN and ethernet STAs. As a matter of fact, the AP transmits half of the voice traffic to WLAN STAs. Therefore, L_{DATA} of the tagged STA can be reasonably approximated as the weighted mean of l different packet sizes in an infrastructure BSS in order to consider STAs with heterogeneous traffic flows. By symmetry, the STA label subscript of T_{DATA_a} is dropped such that

$$T_{DATA} = \frac{8L_{DATA}}{R_{DATA}}, \quad L_{DATA} = \left(\frac{\sum_l \lambda_l PLEN_l}{\sum_l \lambda_l} + L_{MAChdr}\right) \tag{6.24}$$

where R_{DATA} is the PHY data rate and L_{MAChdr} is the size of the MAC header. λ_l and $PLEN_l$ are the arrival rate and packet length of l different packet sizes in an infrastructure

BSS, respectively. Note that (6.24) also implies that the STA label subscript of the time slot durations in (6.16) can be omitted.

Once the expected length of backoff slot time is known, the average MAC service time is computed in two parts, viz., the expected time spent in the backoff states and expected time spent in the transmission states, according to the discrete time Markov chain transition diagram illustrated in Figure 6.3. First, the expected number of backoff states $b_{i,k}$, where $i \in [0,m], k \in [1, W_i - 1]$, encountered by the tagged STA before its packet arrives at stage i can be expressed as

$$\begin{aligned} E\left[BO_a\right] &= \sum_{i=0}^{m} p_{f_a}^{i}.\frac{W_i - 1}{2} \\ &= \begin{cases} \frac{W_0}{2}\left(\frac{1-\left(2P_{f_a}\right)^{m+1}}{1-2P_{f_a}}\right) - \frac{1}{2}\left(\frac{1-P_{f_a}^{m+1}}{1-P_{f_a}}\right), & m \le m' \\ \frac{W_0}{2}\left(\frac{1-\left(2P_{f_a}\right)^{m'+1}}{1-2P_{f_a}} + \frac{2^{m'}P_{f_a}^{m'+1}\left(1-P_{f_a}^{m-m'}\right)}{1-P_{f_a}}\right) - \frac{1}{2}\left(\frac{1-P_{f_a}^{m+1}}{1-P_{f_a}}\right), & m > m' \end{cases}. \end{aligned} \tag{6.25}$$

Owing to the fact that a packet is dropped when it experiences another collision after reaching the last backoff stage m, i.e., after $m+1$ collisions, the expected number of backoff states $b_{i,k}$, where $i \in [0,m], k \in [1, W_i - 1]$, encountered by the tagged STA before its packet is dropped can be written as

$$\begin{aligned} E\left[BO_{drop}\right] &= \sum_{i=0}^{m} \frac{W_i - 1}{2} \\ &= \begin{cases} \frac{W_0\left(2^{m+1}-1\right)-(m+1)}{2}, & m \le m' \\ \frac{W_0\left(2^{m'+1}-1\right)+W_0 2^{m'}(m-m')-(m+1)}{2}, & m > m' \end{cases}. \end{aligned} \tag{6.26}$$

It then follows that the expected time spent by the tagged STA in the backoff states $b_{i,k}$, where $i \in [0,m], k \in [1, W_i - 1]$, conditioned on successful packet delivery is

$$E\left[T_{BO_a}\right] = \left(\frac{E\left[BO_a\right] - P_{f_a}^{m+1}.E\left[BO_{drop}\right]}{1 - P_{f_a}^{m+1}}\right) E\left[slot_a\right] \tag{6.27}$$

where $p_{f_a}^{m+1}$ is the probability that the tagged STA's packet is dropped after exceeding its retry limit, and $1 - p_{f_a}^{m+1}$ is the probability that the tagged STA's packet is not dropped. In other words, expression (6.27) gives the expected time spent in the backoff states only for packets that are successfully received at the destination, whereas packets dropped due to retry limit do not contribute to the average MAC service time computation as in [132] and [139]. Similarly, the expected time spent in the transmission states $b_{i,0}$, where $i \in [0, m]$, conditioned on successful packet delivery by modeling the number of transmissions per packet of the tagged STA as geometrically distributed with the probability of success $1 - P_{f_a}$ can be expressed as

$$\begin{aligned} E\left[T_{TX_a}\right] &= \left[\left(1 - P_{f_a}\right) T_s + P_{f_a}\left(1 - P_{f_a}\right)\left(T_c + T_s\right) + \dots \right. \\ &\quad + P_{f_a}^{m-1}\left(1 - P_{f_a}\right)\left(mT_c - T_c + T_s\right) + \dots \\ &\quad \left. + P_{f_a}^{m}\left(1 - P_{f_a}\right)\left(mT_c + T_s\right)\right] \frac{1}{1 - P_{f_a}^{m+1}} \\ &= T_s + T_c\left[\frac{P_{f_a}}{\left(1 - P_{f_a}\right)\left(1 - P_{f_a}^{m+1}\right)}\left(1 - (m+1) P_{f_a}^{m} + mP_{f_a}^{m+1}\right)\right]. \end{aligned} \tag{6.28}$$

Without loss of generality, $E\left[T_{TX_a}\right]$ can be rewritten as

$$E\left[T_{TX_a}\right] = T_c\left[1 + \left(\frac{P_{f_a}}{\left(1 - P_{f_a}\right)\left(1 - P_{f_a}^{m+1}\right)}\left(1 - (m+1) P_{f_a}^{m} + mP_{f_a}^{m+1}\right)\right)\right] \tag{6.29}$$

which is immediate from (6.16) and (6.24). Finally, the closed-form of the average MAC service time can be expressed as the total amount of time spent by the tagged STA in both the backoff and transmission states given by

$$E\left[T_{S_a}\right] = E\left[T_{BO_a}\right] + E\left[T_{TX_a}\right] \tag{6.30}$$

without the need to differentiate a fairly complex probability generating function. This is the main simplification which achieves computational efficiency for pragmatic implementations. Moreover, the average MAC service time suffices for the analysis of theoretical queueing models such as the $M/M/1/K$ and $M/G/1/K$ models. Furthermore, it is noted that expression (6.30) is consistent with the one found in [160].

6.2.3 Queueing Model Analysis

The importance of queueing delay in ensuring good QoS for RT applications has been explained in Section 6.1.1. In a realistic networking scenario, most of the MAC frames will carry higher layer packets such as TCP/IP or real-time transport protocol (RTP)/user datagram protocol (UDP)/IP in their payload for NRT or RT applications, respectively. These applications are typically sensitive to the end-to-end delay and queue characteristics such as average queue length, MAC delay, queue blocking probability, and throughput. Thus, it will be imperative to analyze the queueing model in order to obtain such performance metrics for admission control and capacity analysis in VoWLAN.

In general, the queueing model is characterized by its arrival process, service time distribution, and queue discipline. This analysis assumes that packets, which arrive at each STA and consequently the AP, follow a Poisson process where the average number of packets arriving per unit time is λ. Each packet then requires $\frac{1}{\mu}$ time units of MAC service on average where the average MAC service time spent during contention process for medium access has been derived in Section 6.2.2. Note that Poisson arrivals have been demonstrated in [169] to give a reasonable approximation. Furthermore, it has been shown in [133] that an exponential distribution is a good approximation of the MAC service time in the non-saturation mode. Apparently, the packet transmission process at each STA and the AP can be modeled as a single server, FIFO queue with finite length K. More specifically, under the assumptions of Poisson arrivals and exponential service time, the queue of each STA and the AP can be analyzed by using the $M/M/1/K$ model where the steady state probabilities are readily obtained from [170] as

$$P_0 = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}}, & \rho \neq 1 \\ \frac{1}{K+1}, & \rho = 1 \end{cases},$$

$$P_n = \rho^n P_0, \qquad n \in [0, K], \tag{6.31}$$

which are stable even for $\rho > 1$. The average queue length is given by

$$L_q = \begin{cases} \frac{\rho}{1-\rho} - \frac{\rho\left(K\rho^K+1\right)}{1-\rho^{K+1}}, & \rho \neq 1 \\ \frac{K(K-1)}{2(K+1)}, & \rho = 1 \end{cases}. \tag{6.32}$$

It is known that the average number of packets in the system of a single server queue ($G/G/1$) is $L = L_q + \lambda/\mu$ where λ/μ is the expected number of packets in service in steady state, also known as the offered load. However, this relation and the Little's formula need to be adjusted for the finite $M/M/1/K$ model by a factor of $(1 - P_B)$ where P_B is a fraction of the arrivals that does not join the system when the queue is full. Accordingly, the average number of packets in the system and MAC delay by relations from the Little's formula are given by

$$\begin{cases} L = L_q + \frac{\lambda(1-P_B)}{\mu} \\ W = \frac{L}{\lambda(1-P_B)} \end{cases}, \quad P_B = P_K. \tag{6.33}$$

In order to consider heterogeneous traffic flows between STAs, it is assumed that the queue of the AP can store K packets, independent of their sizes. Such a logical buffer can be achieved easily by using virtual memory mapping as in [171] and [172], which has become a reality with the recent advances of high performance network processors [173].

Again, consider the case of n STAs where the per-STA quantities are subscripted with the STA label $a = 1, \ldots, n$. The PLR_a of the tagged STA is then computed by assuming that the probability of blocking P_{B_a} and the probability of packet drop due to retry limit P_{D_a}

are two independent events as

$$PLR_a = 1 - (1 - P_{B_a})(1 - P_{D_a}), \quad P_{D_a} = P_{f_a}^{m+1} \tag{6.34}$$

where P_{f_a} is the transmission failure probability from (6.12) that occurs according to a Bernoulli process. It is now trivial to compute the throughput efficiency (or normalized throughput) of each STA and the AP by

$$\bar{S}_a = \frac{8L_{DATA}}{R_{DATA}}.\begin{cases} \lambda_a (1 - PLR_a), & a \in [1, N-1] \\ \sum_{b=1}^{N-1} \lambda_b (1 - PLR_a), & a = N \end{cases} \tag{6.35}$$

where R_{DATA} is the PHY data rate. Note that for the case of $M/M/1$ model where $K \to \infty$, the steady state probabilities, average queue length, packets in the system, and MAC delay reduce to

$$\begin{cases} P_0 = 1 - \rho \\ P_n = \rho^n P_0 \\ L_q = \frac{\rho^2}{1-\rho} \\ L = \frac{\rho}{1-\rho} \\ W = \frac{L}{\lambda} \end{cases}, \quad \rho < 1. \tag{6.36}$$

As a result, the throughput efficiency (or normalized throughput) of each STA and the AP also reduces to

$$\bar{S}_a = \frac{8L_{DATA}}{R_{DATA}}.\begin{cases} \lambda_a \left(1 - P_{f_a}^{m+1}\right), & a \in [1, N-1] \\ \sum_{b=1}^{N-1} \lambda_b \left(1 - P_{f_a}^{m+1}\right), & a = N \end{cases}. \tag{6.37}$$

The expressions (6.31) – (6.35) are of key importance since they relate traffic intensity ρ (function of arrival rate and service rate) and wireless channel conditions (function of service rate) to the key performance metrics of MAC delay, PLR, and throughput efficiency. These performance metrics are crucial for proper admission control and provide insights into the capacity analysis of VoWLAN so that its saturation point can be accurately predicted.

6.3 Model Validation

The unified analytical model presented in the previous sections is validated by comparing numerical and simulation results. The analytical model is PHY independent and can be applied to any IEEE 802.11 PHYs by using the appropriate set of PHY parameters. In what follows, the numerical and simulation results are obtained based on the system parameters in Table 6.1 specified for the high rate direct sequence spread spectrum using the long preamble and header (HR/DSSS) and extended rate PHY using orthogonal frequency division multiplexing modulation (ERP-OFDM) PHYs in the IEEE 802.11 standard. The simulation models are developed by using OPNET™ Modeler® 14.5 with Wireless Module discrete event simulator. The simulations of the IEEE 802.11 carrier sense multiple access with collision avoidance (CSMA/CA) protocol are performed under both ideal and error-prone AWGN wireless channel conditions.

It is clear from previous discussions that any performance analysis of the IEEE 802.11 networks requires approximations in order to be analytically tractable. The tradeoffs between simplicity and accuracy would be meaningful only if these approximations do not result in unacceptable numerical inaccuracies. The performance metrics for the case of error-prone channel are evaluated at the worst case BER of 10^{-5}. This corresponds to the packet error rate (PER) threshold of 1% which is a standard measure of robustness used by the IEEE 802.11 committee to determine the SNR requirements during the selection of standards. Note that the SNR requirement is defined as the SNR required to maintain a PER of 1% with a 1000 byte packet. The IEEE 802.11 standard supports a wide array of PHYs. E.g., the IEEE 802.11 direct sequence spread spectrum (DSSS) PHY uses differential binary phase shift keying and differential quadrature phase shift keying modulations with Barker code where each code word is encoded with 1 bit or 2 bits for the correspond-

Table 6.1: System parameters of the IEEE 802.11b and 802.11g PHYs.

System Parameters	Notations	802.11b (HR/DSSS)	802.11g (ERP-OFDM)
Slot time	σ	20 μs	**9** (only ERP STAs), 20 μs
SIFS duration	T_{SIFS}	10 μs	10 μs
DIFS duration	T_{DIFS}	50 μs	28 μs
Propagation delay	δ	1 μs	$<<$ 1 μs
PLCP preamble duration	$T_{PLCPpre}$	144 μs (long) 72 μs (short)	16 μs
PLCP header duration	$T_{PLCPhdr}$	48 μs (long) 24 μs (short)	4 μs
Total PLCP overheads duration	T_{PHY}	192 μs (long)	20 μs
Symbol duration	T_{SYM}	NA	4 μs
Signal extension duration	T_{SIGEXT}	NA	6 μs
Service field size	L_{SER}	NA	16 bits
Tail field size	L_{TAIL}	NA	6 bits
Data bits per OFDM symbol	N_{DBPS}	NA	24 bits
PHY data rate	R_{DATA}	1, 2, 5.5, **11** Mbps	**6**, 9, 12, 18, 24, 36, 48, 54 Mbps
PHY control rate	R_{CON}	**1**, 2 Mbps	**6**, 12, 24 Mbps
MAC header size including 32 bit FCS	L_{MAChdr}	28 bytes	
MAC payload size	L_{PLD}	60, 64, 120, 1000 bytes	
MAC data frame size	L_{DATA}	$L_{MAChdr} + L_{PLD}$	
MAC ACK frame size	L_{ACK}	112 bits	
Minimum CW size	CW_{min}	31	
Maximum CW size	CW_{max}	1023	
Maximum CW increasing factor	m'	5	
Retry limit (Maximum backoff stage)	m	6	
Bit error rate	BER	0, 10^{-5}	

ing data rates of 1 Mbps and 2 Mbps, respectively. The IEEE 802.11b HR/DSSS PHY also employs differential binary phase shift keying and differential quadrature phase shift keying modulations but with complementary code keying where sophisticated mathematical transforms allow the use of a few 8-bit sequences to encode 4 or 8 bits per code word to achieve data rates of 5.5 Mbps and 11 Mbps, respectively. The IEEE 802.11g ERP-OFDM PHY, which is based on the IEEE 802.11a orthogonal frequency division multiplexing (OFDM) PHY, utilizes binary phase shift keying (BPSK), quadrature phase shift keying (QPSK), 16-quadrature amplitude modulation (QAM), or 64-QAM and convolutional coder with the coding rate of 1/2 or 3/4 to deliver data rates of 6, 9, 12, 18, 24, 36, 48, and 54 Mbps. The works reported in [174] and [175] illustrate the PER vs. SNR for different types of modulation and coding scheme. It is evident from these PER vs. SNR plots that BER of 10^{-5} is a reasonable upper bound in order to achieve PER $\leq 2\%$ in an AWGN channel for all the modulation and coding schemes considered in this thesis. Moreover, the widespread use of link adaptation mechanism in practice will typically limit PER $\leq 2\%$ to ensure robust wireless communications.

Figure 6.5 and Figure 6.6 give the comparison of the numerical results obtained from the analytical model with the simulation results obtained from the OPNET simulator under ideal and error-prone channel conditions, respectively. In both cases, there is a good agreement between the analysis and simulations, which confirms the accuracy of modeling assumptions, particularly, during the transition from the non-saturation to saturation mode for the key performance metrics of MAC delay, PLR, and throughput efficiency. The overestimation of the collision probability and MAC service time in the non-saturation region is due to the fact that both post-backoff and the possibility of immediate transmission after medium has been idle for a DIFS duration are not modeled in the Markov chain, which is originally designed by Bianchi under the saturation assumption. In other words, the Markov chain model assumes saturation condition and neglects both the possibility of packets arriving to an empty queue and the possibility of arriving packets to access the medium after it has been idle for a DIFS duration. Consequently, these result in a higher collision probability and longer MAC service time. However, as discussed in Section 6.1.1, the MAC delay and PLR incurred in this non-saturation region will be insignificant, and all offered load will be successfully transmitted. Although such overestimations could be corrected by incorporating the works of [145] and [147] at the expense of additional complexity, there will be negligible improvement in terms of accuracy for the purpose of admission control and capacity analysis. Moreover, the approach of using saturation analysis to determine the transmission probability and later incorporating it as a scaling factor to model a finite load infrastructure BSS WLAN has been recently validated in [159], which corresponds to the approach of the proposed unified analytical model.

Apart from the above observations, the analysis has also captured a number of important characteristics which will now be discussed. First, a linear relationship between the offered load and throughput exists under the non-saturation mode in which the throughput increases with the offered load along the $y = \frac{1}{2}x$ line (cf. Figure 6.1 of Section 6.1.1). Second, the maximum throughput is reached before saturation in both analysis and simulation when the number of STA pair is more than one. Furthermore, the point where the maximum throughput occurs is relatively insensitive to the number of STA pairs, but rather it is dependent on the offered load. This is intuitive in the non-saturation mode because most of the STAs' queue would be empty, given that all packets that arrive are immediately served by the MAC, and would not contend for medium access. Third, the saturation throughput during high offered load has similar behavior to that of Bianchi's model. Specifically, the saturation throughput decreases as the number of STA pairs increases for small initial CW sizes of 8, 16, 32, and 64. The decrease of saturation throughput with an increasing number of STA pairs is again intuitive because the collision probability increases with the number of STA pairs. Although the collision probability is dependent on the number of STA pairs in the saturation mode, it is relatively insensitive to the number of STA pairs and increases with the offered load in the non-saturation mode. Fourth, the transition from the non-saturation to saturation mode where a marked increase in the col-

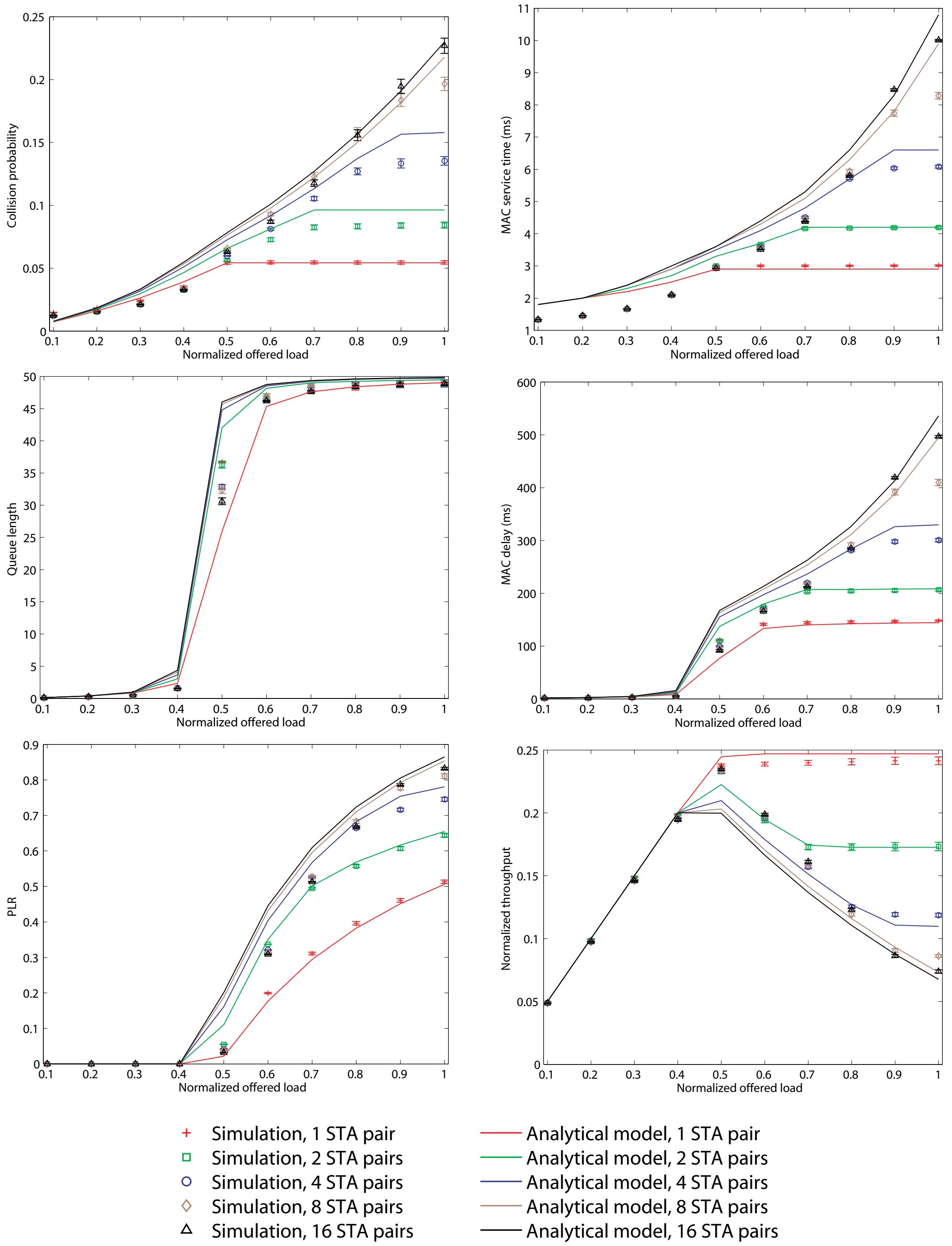


Figure 6.5: Model validation: Analysis vs. OPNET simulation for homogeneous CBR traffic source, HR/DSSS PHY @ 11 Mbps, L_{PLD} = 1000 bytes, and BER = 0 with different number of STAs and varying traffic arrival rates.

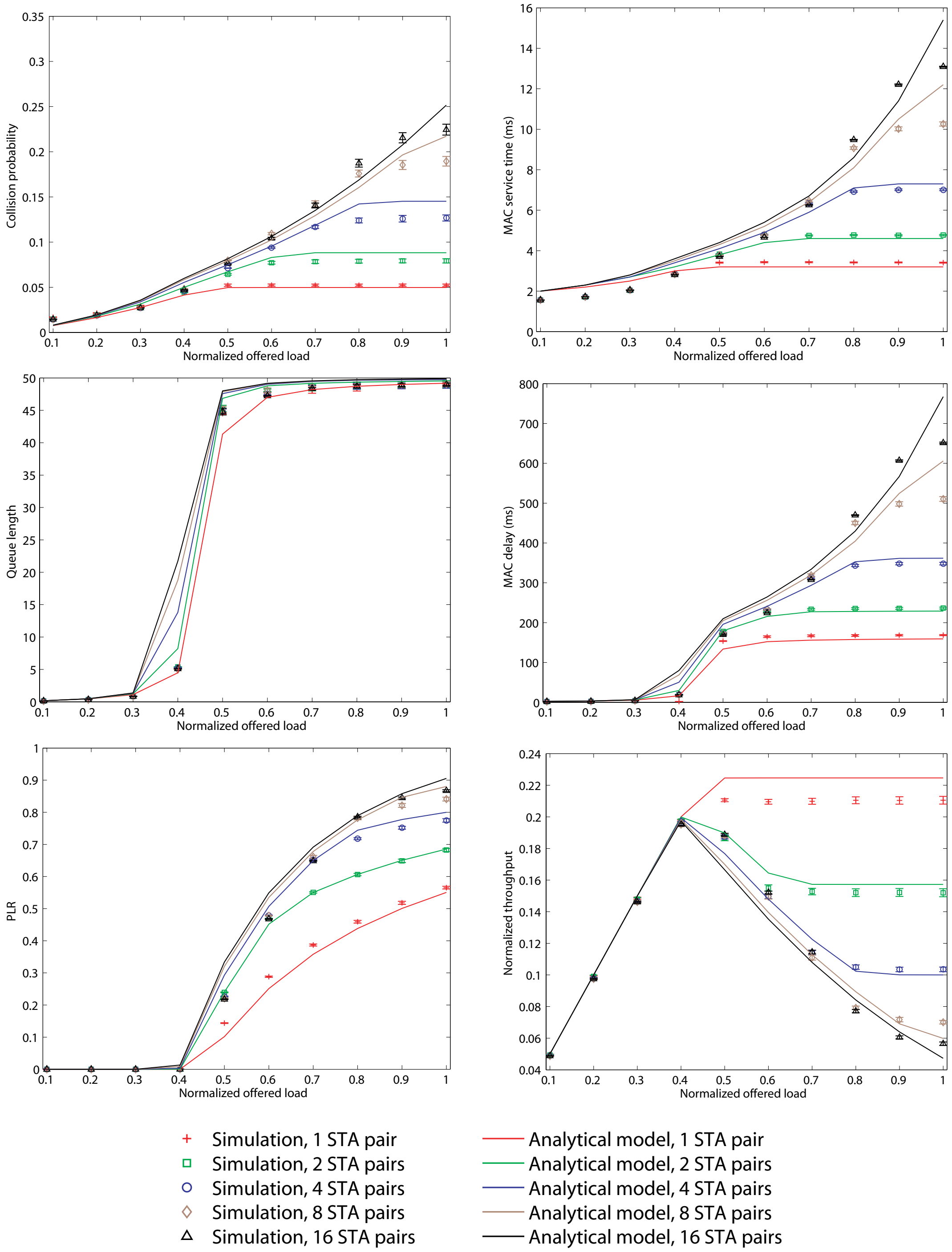


Figure 6.6: Model validation: Analysis vs. OPNET simulation for homogeneous CBR traffic source, HR/DSSS PHY @ 11 Mbps, L_{PLD} = 1000 bytes, and BER = 10^{-5} with different number of STAs and varying traffic arrival rates.

lision probability, MAC service time, queue length, MAC delay, and PLR, as well as a corresponding decrease in the throughput are successfully captured.

It is clear that the MAC service time is dependent on the collision probability. Therefore, it exhibits similar trends that correlate very well to the collision probability over the different range of offered load. On the other hand, all the other performance metrics derived from the queueing analysis, viz., queue length, MAC delay, PLR, and throughput efficiency are dependent on the MAC service time. Considering all facts, the Markov chain analysis provides an upper bound of the average MAC service time. Consequently, the queueing analysis that is based on the average MAC service time gives the upper bounds of the average queue length, MAC delay, and PLR, whereas it gives the lower bound of the throughput efficiency. Collectively, these are desirable for reliable admission control and capacity analysis.

6.4 Performance Evaluation

This section presents the performance evaluation of the analytical model developed in Section 6.2 with some of the notable, pertinent analytical models and OPNET simulation. The comparison between these pertinent and proposed analytical models is summarized in Table 6.2. In the case of ideal channel, the models of Chatzimisios, Zhai, and Xiao are used for comparison. Note that Xiao's model is originally catered for service differentiation in the IEEE 802.11e EDCA. It is essentially similar to Chatzimisios's model when considering only a single traffic class. The collision probability expression of these models is replaced with the generalized form of (6.11) in order to evaluate the performance metrics for a finite load infrastructure BSS where asymmetric load between the AP and its associated STAs exists. On the other hand, Cai's model, which is based on the non-Markovian random backoff model and $G/G/1$ queueing model, is also included for comparison under ideal channel conditions. In the case of error-prone channel, Ni's model which is also similar to Chatzimisios's model is used for performance comparison. This

Table 6.2: Comparison between the pertinent and proposed analytical models.

Models	BSS Types	Backoff Processes	Queueing Models	Backoff Freezing	Traffic Conditions	Channel Conditions	STA Types
Chatzimisios [132]	IBSS	Markovian	No	No	Saturation	Ideal, Error-prone	Homogeneous
Zhai [133]	IBSS	Markovian	$M/M/1/K$, $M/G/1/K$	Yes	Non-saturation, Saturation	Ideal	Homogeneous
Xiao [140]	IBSS	Markovian	No	Yes	Saturation	Ideal	Heterogeneous
Cai [154]	Infrastructure BSS	Non-Markovian	$G/G/1$	No	Non-saturation	Ideal	Homogeneous
Ni [134]	IBSS	Markovian	No	No	Saturation	Ideal, Error-prone	Homogeneous
Proposed (cf. Section 6.2)	Infrastructure BSS	Markovian	$M/M/1/K$	Yes	Non-saturation, Saturation	Ideal, Error-prone	Heterogeneous

is due to the fact that Ni's model considers errors in both data and ACK frames while Chatzimisios's model considers only errors in data frames. In order to have a uniform comparison across all the considered performance metrics, the models of Chatzimisios, Xiao, and Ni are augmented with the $M/M/1/K$ queueing model. In all cases, the performance evaluations in the following are based on the system parameters found in Table 6.1.

6.4.1 Effect of Traffic Sources, Physical Layers, and Data Rates

Two types of RT traffic which are generated from either CBR or VBR traffic sources (cf. Appendix A-2.3 for details) are considered in this chapter. Since the IEEE 802.11 standard specifies a variety of PHYs and data rates, it is also interesting to investigate the performance of the proposed analytical model using different PHYs and data rates. In the following evaluations, unless otherwise stated, the PHY data rate of 11 Mbps (6 Mbps) and PHY control rate of 1 Mbps (6 Mbps) are used for the HR/DSSS (ERP-OFDM) PHY, respectively. In addition, a slot time of 9 μs is used for the case of ERP-OFDM PHY as only ERP-enabled STAs are simulated.

Figure 6.7 and Figure 6.8 show the comparison between different analyses vs. OPNET simulation for the CBR traffic sources and HR/DSSS PHY, under both ideal and error-prone channel, respectively. Figure 6.9 and Figure 6.10 show the comparison between different analyses vs. OPNET simulation for the VBR traffic sources and ERP-OFDM

PHY, under both ideal and error-prone channel, respectively. In all cases, the proposed model offers the best performance in terms of accuracy between the analysis and simulations, whereas the models of Chatzimisios, Zhai, Xiao, and Ni are found to exhibit inaccuracies in the collision probability in the saturation region. As a result of the overestimation of collision probability, the corresponding MAC service time, queue length, MAC delay, and PLR are also overestimated whilst the throughput efficiency is underestimated in the saturation region. In particular, Zhai's model suffers from the largest approximation error. This phenomenon will be more pronounced for an infrastructure BSS since the AP is the capacity bottleneck owing to the asymmetric traffic load between the AP and its associated STAs. To demonstrate this fact, Figure 6.11 plots the numeric and simulation results of throughput efficiency against the normalized offered load for the AP and STA. The PLR is also superimposed to illustrate its relationship with the throughput efficiency. From the results, it is clear that the AP is the capacity bottleneck since the PLR increases and the throughput efficiency drops below the dotted normalized offered load line when the number of STA pairs increases beyond thirty. This signifies that the AP is transiting into the saturation mode, and therefore it starts discarding packets as the MAC delay increases and the queue becomes full. On the other hand, the STA remains in the non-saturation mode where the PLR is zero and the throughput efficiency equals the normalized offered load. Such an observation can also be found in [154] and [159]. In addition, this phenomenon is directly associated to whether and how backoff freezing is modeled, which will be deferred to Section 6.4.2 for a thorough discussion.

On the other hand, Cai's model which assumes an infinite queue length fails to predict the transition from the non-saturation to saturation mode as the $G/G/1$ model is stable only when traffic intensity $\rho < 1$. As a result, this model has limited practical use in load distribution algorithms since the spare capacity cannot be effectively optimized without the proper knowledge of the optimal operating point or, in other words, the transition region. Moreover, it is observed that the collision probability which is one of the key parameters has poor accuracy, especially, under high offered load. This is due to the fact that Cai's

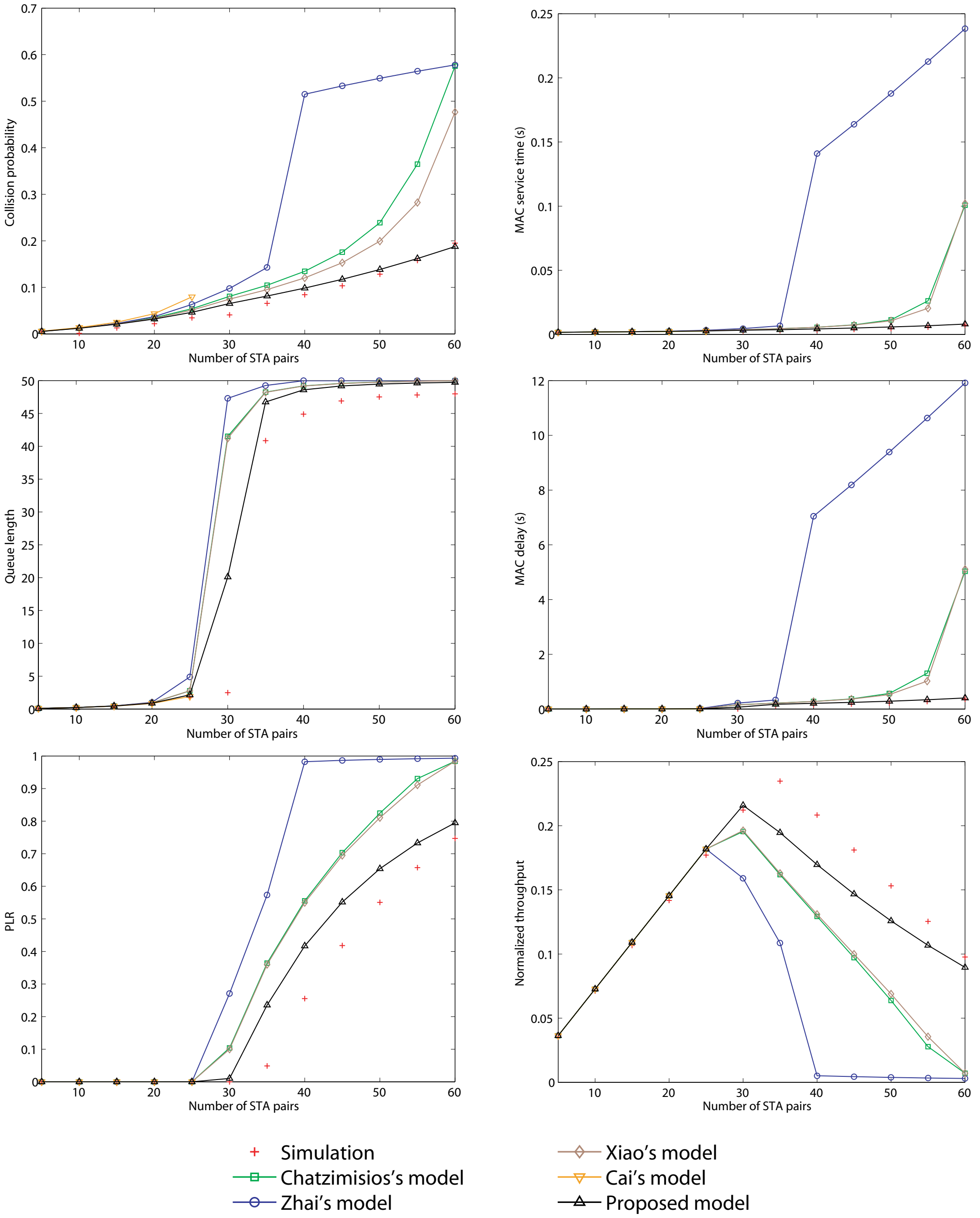


Figure 6.7: Performance evaluation: Comparison of different analyses vs. OPNET simulation for homogeneous CBR traffic source, HR/DSSS PHY @ 11 Mbps, λ = 10 frames/sec, L_{PLD} = 1000 bytes, and BER = 0.

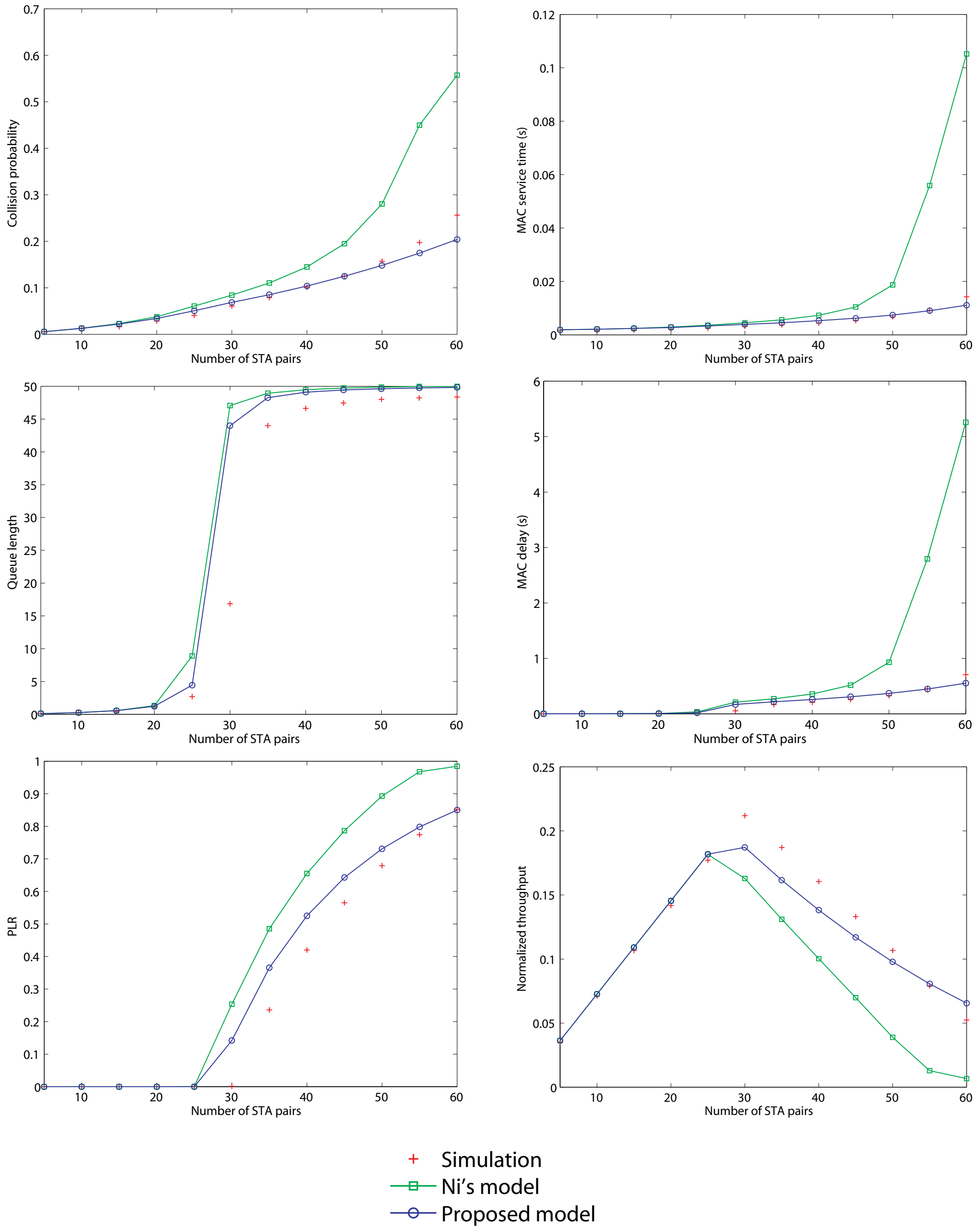


Figure 6.8: Performance evaluation: Comparison of different analyses vs. OPNET simulation for homogeneous CBR traffic source, HR/DSSS PHY @ 11 Mbps, λ = 10 frames/sec, L_{PLD} = 1000 bytes, and BER = 10^{-5}.

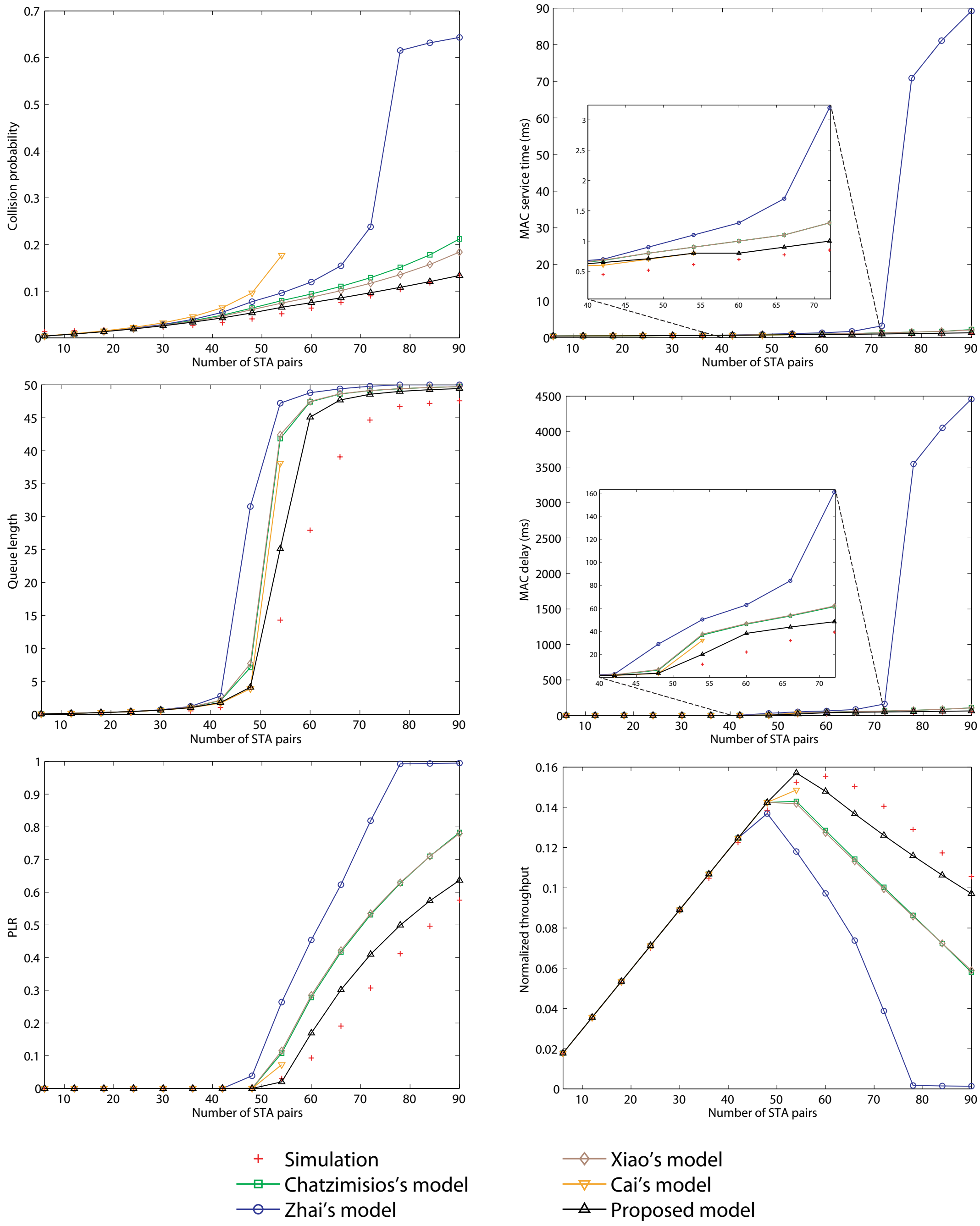


Figure 6.9: Performance evaluation: Comparison of different analyses vs. OPNET simulation for heterogeneous VBR traffic source, ERP-OFDM PHY @ 6 Mbps, λ = 50, 33.3, 100 frames/sec, L_{PLD} = 60, 64, 120 bytes, respectively, and BER = 0.

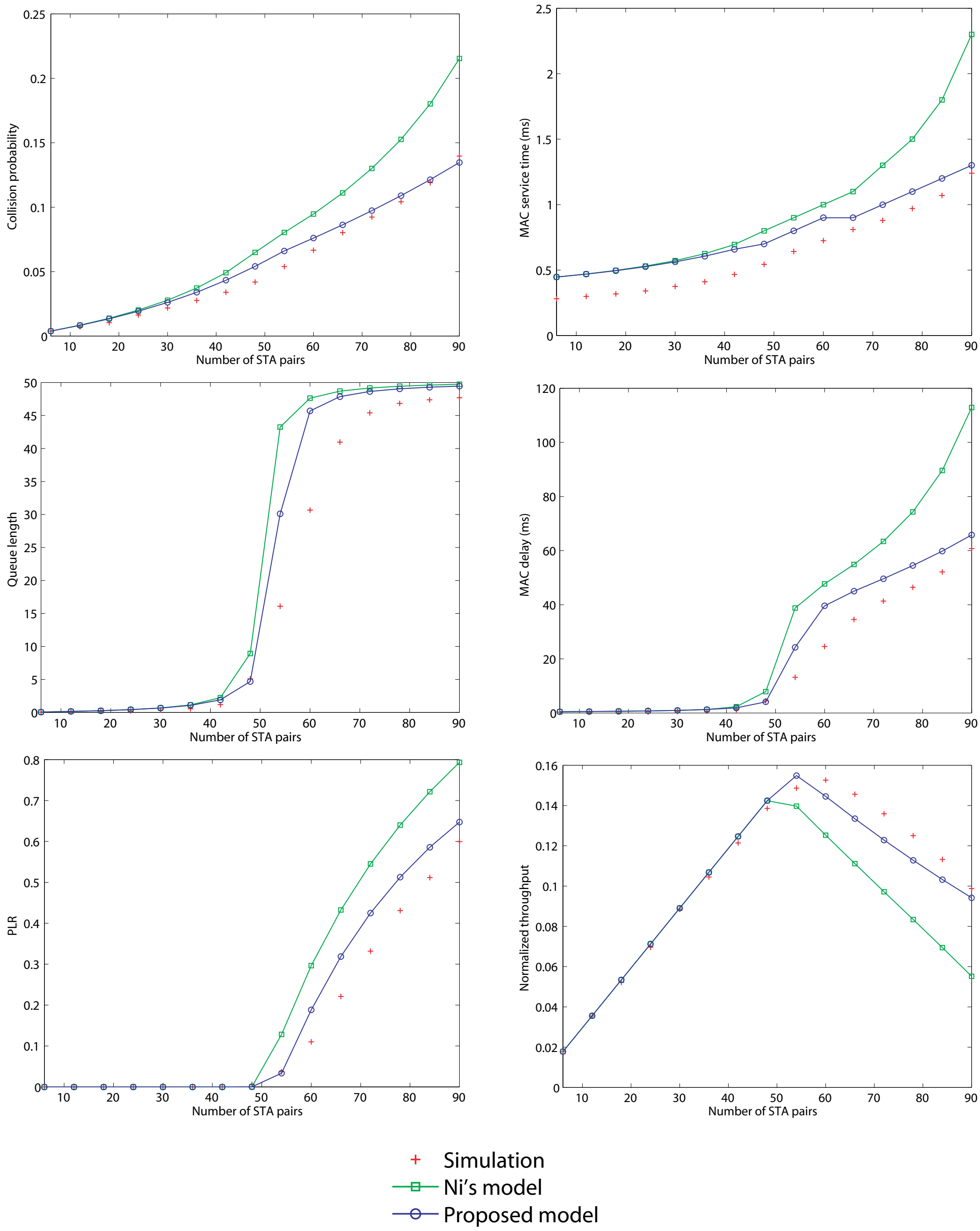


Figure 6.10: Performance evaluation: Comparison of different analyses vs. OPNET simulation for heterogeneous VBR traffic source, ERP-OFDM PHY @ 6 Mbps, λ = 50, 33.3, 100 frames/sec, L_{PLD} = 60, 64, 120 bytes, respectively, and BER = 10^{-5}.

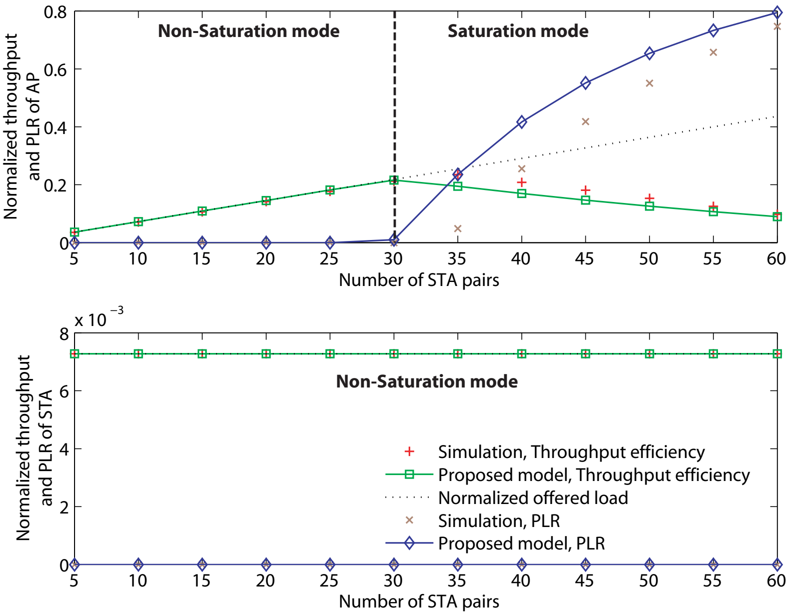


Figure 6.11: Performance evaluation: Asymmetric load between the AP and STA for homogeneous CBR traffic source, HR/DSSS PHY @ 11 Mbps, λ = 10 frames/sec, L_{PLD} = 1000 bytes, and BER = 0.

model analyzes the random backoff process of the DCF using the non-Markovian model where several heuristics are made without mathematical validations. Furthermore, Cai's model is originally designed for voice capacity analysis based on homogeneous STAs with CBR traffic sources. Although the authors argue that using the CBR model gives a tighter upper bound of admissible traffic load to provision guaranteed QoS, it will inevitably result in low network utilization which is undesirable for network operators when bursty VBR traffic sources are prevalent in practice. Additionally, while voice packet is sensitive to delay, it can essentially withstand a loss of up to 2%. Hence, predictive QoS is typically preferred over guaranteed QoS, particularly, in broadband WLANs where a high degree of statistical multiplexing is commonplace. The reason is because predictive QoS will introduce more flexibility in the admission controller, which in turn yields higher network utilization, by exploiting spare capacity opportunistically at the expense of occasional violations [176]. Some examples of load distribution algorithms which provision predictive QoS can be found in Section 7.4. Along the same line, the overestimation of the collision probability noticed in the models of Chatzimisios, Zhai, Xiao, and Ni will ultimately result in low network utilization when employed as admission or load control.

6.4.2 Effect of Backoff Freezing

The effect of backoff freezing for an infrastructure BSS and IBSS, and the consequences if backoff freezing is not considered or improperly modeled in an infrastructure BSS will be examined in this section. From Section 6.4.1, inaccuracies in the collision probability are observed from the models of Chatzimisios, Zhai, Xiao, and Ni. The fundamental reason of this phenomenon is due to whether and how backoff freezing is modeled. Note that Chatzimisios's model does not account for backoff freezing and is used as a benchmark for comparison. It is clear that Zhai's model has the greatest inaccuracy. In fact, the authors notice the overestimation of the MAC service time in the saturation mode, but they suggest that the Markov chain may not have captured all the protocol implementations. They further conclude that the overestimation is a reasonable upper bound of their simulation results. However, the reason for this discrepancy is because backoff freezing is accounted in the signal transfer function of their generalized state transition diagram, which models the backoff decrement process, instead of in the Markov chain used to derive the transmission probability of a packet. Accordingly, the backoff decrement process in Zhai's model can be shown as

$$H_d\left(Z\right) = \frac{\left(1 - P_c\right) Z^{\sigma}}{\left[1 - P_s S_t\left(Z\right) - \left(P_c - P_s\right) C_t\left(Z\right)\right]} \tag{6.38}$$

where the expected length of a backoff slot time can be found by differentiating (6.38) to give

$$\frac{dH_d\left(Z\right)}{dZ}|_{Z=1} = \sigma + \frac{P_s T_s + \left(P_c - P_s\right) T_c}{1 - P_c}. \tag{6.39}$$

However, the expected length of a backoff slot time found in Chatzimisios's model is

$$E\left[slot\right] = \left(1 - P_c\right)\sigma + T_s P_s + T_c\left(P_c - P_s\right). \tag{6.40}$$

Therefore, it is clear that an additional factor of $1/1 - P_c$ in (6.39), i.e., the backoff decrement process of Zhai's model actually accounts for backoff freezing. This is because the expected number of times that the tagged STA needs to sense the channel until it becomes idle is geometrically distributed with probability of success $1 - P_c$ such that

$$\begin{aligned} E\left[sense\right] =& \left(1 - P_c\right) + 2P_c\left(1 - P_c\right) \\ & + 3P_c^2\left(1 - P_c\right) + \ldots \\ =& \frac{1}{1 - P_c}. \end{aligned} \tag{6.41}$$

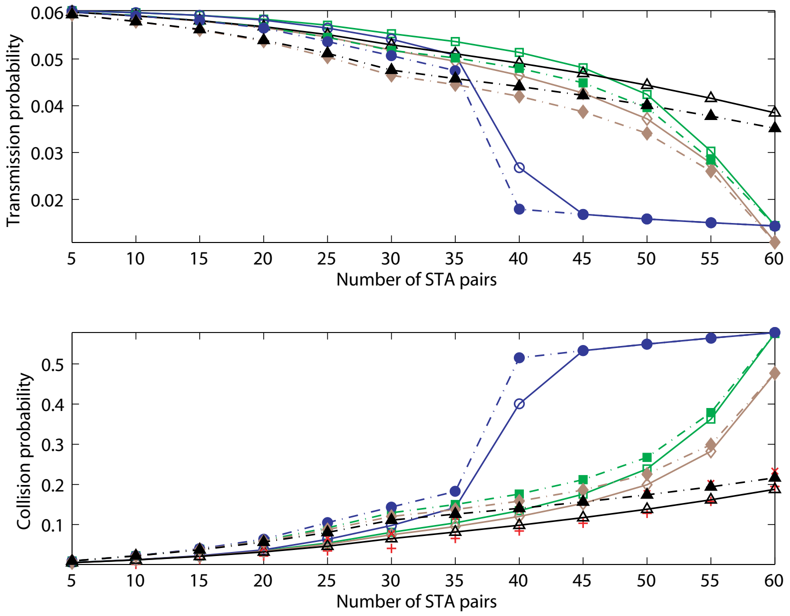
However, it is evident from Figures 6.7 – 6.10 that modeling backoff freezing in the backoff decrement process explicitly results in huge overestimation of the collision probability, especially, in the saturation region. This can be easily justified by observing expression (6.11), and the fact that the probability of a non-empty queue increases with the increasing MAC service time. Accordingly, modeling backoff freezing in the backoff decrement process actually increases the MAC service time and consequently the collision probability of AP instead of reducing the transmission probability of STAs and collision probability of AP as expected. The reason for this counterintuitive behavior is attributed to the fact that backoff freezing in the backoff decrement process induces a *positive feedback* phenomenon. Specifically, a slight increase in the MAC service time results in an increase of the collision probability, which in turn causes the MAC service time to increase further. This catastrophic effect will manifest in the AP and STA to cause serious implications for the queueing model analysis, which directly depends on the average MAC service time, as the AP transits into the saturation mode. In particular, the key performance metrics of MAC delay and PLR will be overly overestimated while throughput efficiency will be grossly underestimated. Hence, backoff freezing should not be independently modeled in the backoff decrement process. This observation is consistent to the works reported in [132] and [168], which do not account for backoff freezing in the backoff decrement process for their saturation delay analyses.

On the other hand, Xiao's model considers backoff freezing in both the backoff decrement process and Markov chain. The results in Figure 6.7 and Figure 6.9 show that the collision probability and its related QoS metric such as the MAC delay are overestimated. This finding is consistent with the work in [142], which is based on Xiao's model, that reveals that the MAC delay is also overestimated in the saturation region. In fact, it can be observed that most of the performance metrics obtained from Xiao's model share similar results to that of Chatzimisios's model which does not consider backoff freezing. The only exception is the collision probability which yields a slightly lower value as compared to Chatzimisios's model when the number of STA pairs increases. This outcome is intuitive as the transmission probability of STAs is expected to reduce with backoff freezing, owing to the fact that STAs now spend a longer time in each backoff state. Again, by the relation of expression (6.11), the decrease of transmission probability of STAs will then cause the collision probability of AP to decrease correspondingly. When the collision probability of AP decreases, the MAC service time, MAC delay, and PLR of AP will also decrease whilst the throughput efficiency will increase. This is the desired effect of backoff freezing.

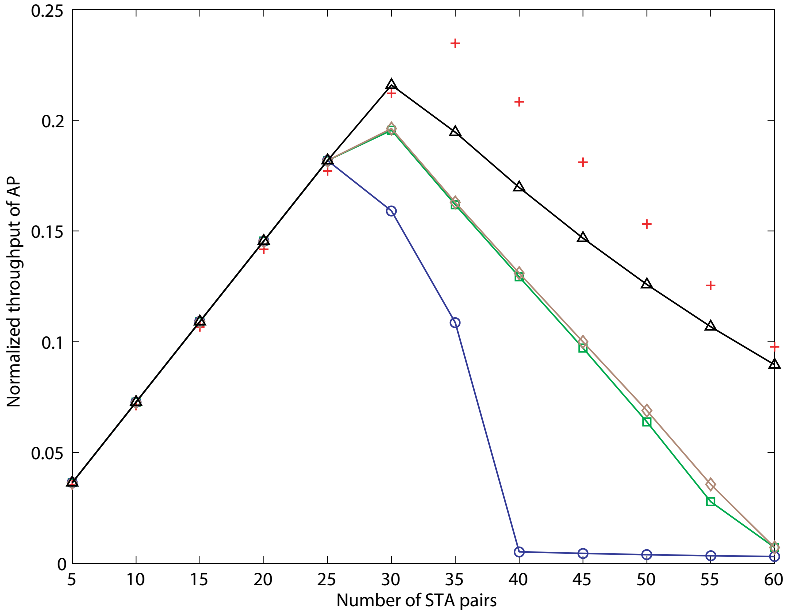
However, when backoff freezing is also accounted in the backoff decrement process, the average time spent in each backoff state, i.e., the MAC service time is immediately increased by a factor of $1/1 - P_c$ as previously explained. Thus, the additional backoff freezing in the backoff decrement process inadvertently negates the effect of backoff freezing in the Markov chain which has decreased the MAC service time initially. As a result, the performance metrics derived by Xiao's model is by large the same as Chatzimisios's model, except for the collision probability of AP which has decreased slightly. Hence, backoff freezing should be modeled only in the Markov chain as in the proposed model where the numerical results are found to be in high agreement with the simulation results. In fact, this has been noticed and recognized in [156] and [177] where backoff freezing is considered in the Markov chain but not in the backoff decrement process.

In order to prove that the inaccuracy is due to the overestimation of the collision probability of STAs, Figure 6.12 examines the relationships between the transmission probability and collision probability of both AP and STA as a function of the number of STA pairs. For the case without backoff freezing, it is noted from Chatzimisios's model that when the number of STA pairs increases, more packet collisions occur as a result of higher collision probability. On the other hand, the transmission probability reduces as STAs spend more time in the backoff states, owing to the higher collision probability. In the case of an infrastructure BSS with asymmetric traffic load between the AP and its associated STAs, the situation is quite different as the AP saturates much earlier than its associated STAs. Particularly, it should also be noticed that the collision probability (transmission probability) of AP is generally lower (higher) than its associated STAs. Moreover, it is obvious from expression (6.11) that the probability of empty queue of AP $P_{0_{AP}}$ and the transmission probability of AP τ_{AP} will increase with backoff freezing. On the contrary, the probability of empty queue of STAs $P_{0_{STA}} \sim 1$ as they operate in the non-saturation mode most of the time while the transmission probability of STAs τ_{STA} reduces with backoff freezing. As a result, the collision probability of STAs $P_{c_{STA}}$ has a smaller value and remains relatively constant as the number of STA pairs increases because $P_{c_{STA}}$ is dominated by the decrease of τ_{STA} whilst the increase of $P_{0_{AP}}$ and τ_{AP} have counterbalancing effect.

This phenomenon is observed in Figure 6.12(a) where the collision probability of the proposed model provides very good accuracy which is in almost exact agreement with the simulation results. On the contrary, the models of Chatzimisios, Zhai, and Xiao overestimate the collision probability and underestimate the transmission probability as previously explained. It is now clear that the collision probability (transmission probability) of STAs and AP will be overestimated (underestimated) in the saturation region when backoff freezing is either not considered or properly modeled. Furthermore, from Figure 6.12(b) it is clear that the consequence of these negligences results in an unrealistic lower bound of throughput efficiency. This is undesirable for admission or load control designed to provision predictive QoS because spare capacity which cannot be effectively



(a) Transmission and collision probabilities.



(b) Throughput of AP.

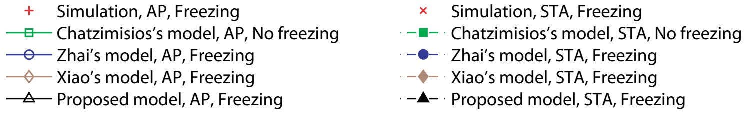


Figure 6.12: Effect of freezing for an infrastructure BSS under non-saturation condition with homogeneous CBR traffic source, HR/DSSS PHY @ 11 Mbps, $\lambda = 10$ frames/sec, $L_{PLD} = 1000$ bytes, and BER = 0.

optimized will result in low network utilization. Collectively, these results accentuate the importance of careful backoff freezing modeling for an infrastructure BSS.

Most of the existing analytical models for an IBSS are based on the seminal work of Bianchi which is developed under the assumption of homogeneous STAs, saturation conditions, and does not consider the effect of backoff freezing. Recently, the work of [144] has revisited the modeling of backoff freezing for an IBSS by introducing *anomalous slots* which occur right after a channel busy period. The authors point out that Ziouva's backoff freezing [135] will result in the underestimation of the collision probability and consequently overestimation of the throughput efficiency. Although their argument holds for the case of an IBSS with homogeneous STAs under saturation conditions, it does not necessarily apply for the case of a finite load infrastructure BSS with asymmetric traffic load between the AP and its associated STAs. In fact, it has been shown in Section 6.4 that Ziouva's backoff freezing which has been implemented in the proposed model achieves good accuracy. In what follows, a qualitative explanation is provided along the same line as in [144] to highlight the impact of Ziouva's backoff freezing on both IBSS and infrastructure BSS.

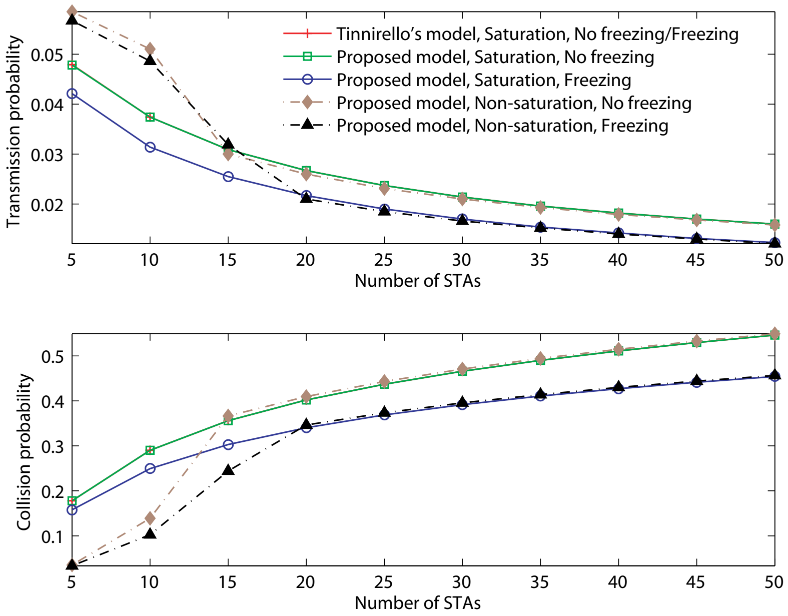
Ziouva's backoff freezing is modeled by modifying the backoff decrement probability of the Markov chain to $1 - Pc$ which corresponds to the probability of an idle slot instead of 1 as originally defined by Bianchi. Suppose that N is the number of slots representing channel activities for a given observation interval and A is the number of busy slots. The probability of a busy slot P_b for Bianchi's model without backoff freezing and Ziouva's model with backoff freezing can then be estimated as $\hat{P}_b^{Bianchi} = A/N$ and $\hat{P}_b^{Ziouva} = A/\left(N + A\right)$, respectively. In the case of an IBSS with n STAs attempting to access the medium, it can be assumed that $\hat{P}_b = 1 - \left(1 - \hat{\tau}\right)^n \simeq 1 - \left(1 - \hat{\tau}\right)^{n-1} = \hat{p}$ where $\hat{p}$ is the collision probability as seen by the tagged STA. Hence, the collision probability of Ziouva's model will be lower than Bianchi's model for an IBSS with homogeneous STAs. The basis of this illustration which is used in [144] for the context of an IBSS will now be extended to consider an infrastructure BSS.

In the case of an infrastructure BSS with n STAs attempting to access the medium, it can be assumed that $\hat{P}_{b_i} = 1 - \prod_{i=1}^{n} \left[1 - \left(1 - \hat{P}_{o_i}\right) \hat{\tau}_i\right] \approx 1 - \prod_{j \neq i} \left[1 - \left(1 - \hat{P}_{o_j}\right) \hat{\tau}_j\right] = \hat{p}_i$ where $\hat{p}_i$ is the collision probability as seen by the tagged STA. It is important to note that the expression for the collision probability is changed to a generalized form to account for the asymmetric traffic load between the AP and its associated STAs. Owing to the asymmetric traffic load, P_b from an AP perspective for Bianchi's model and Ziouva's model can be estimated as $\hat{P}_{b_{AP}}^{Bianchi} = A_{STA}/N$ and $\hat{P}_{b_{AP}}^{Ziouva} = A_{STA}/\left(N + A_{STA}\right)$, respectively. Since STAs typically operate in the non-saturation mode, $A_{STA} << A_{AP}$ and $N \geq A_{STA} + A_{AP}$, therefore, $\hat{P}_{b_{AP}}^{Bianchi} \simeq \hat{P}_{b_{AP}}^{Ziouva}$. Similarly, P_b from the STA perspective for Bianchi's model and Ziouva's model can be estimated as $\hat{P}_{b_{STA}}^{Bianchi} = \left(A_{AP} + A_{STA}\right)/N$ and $\hat{P}_{b_{STA}}^{Ziouva} = \left(A_{AP} + A_{STA}\right)/\left(N + A_{AP} + A_{STA}\right)$, respectively. Thus, the collision probability of STAs in Bianchi's model will now be higher than Ziouva's model as $A_{AP} >> A_{STA}$. This is consistent with the observation in Figure 6.12(a) where Ziouva's backoff freezing implemented in the proposed model will result in a lower collision probability than Chatzimisios's model which does not consider backoff freezing.

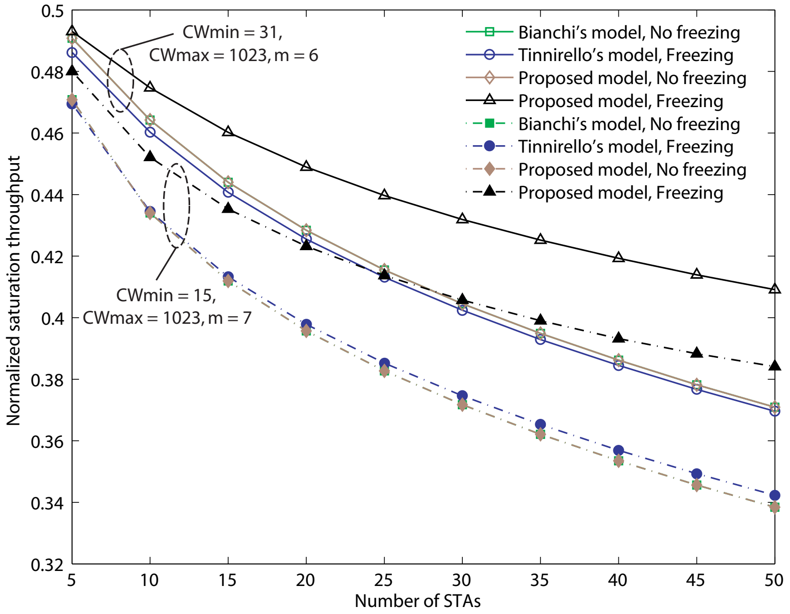
Figure 6.13 then investigates the effect of backoff freezing for an IBSS under both saturation and non-saturation conditions in order to prove the correctness of the proposed model. It is observed from Figure 6.13(a) that the proposed model which is based on Ziouva's backoff freezing results in lower transmission probability and collision probability as compared to the case without backoff freezing, regardless of traffic load. This is intuitive because STAs will spend longer time in backoff states after backoff freezing is incorporated. Hence, it is expected that the transmission probability and consequently the collision probability of the (homogeneous) STAs will be reduced. It is worth noting that this relationship holds only for an IBSS with homogeneous STAs, which is very different from the relationship for an infrastructure BSS with asymmetric load as shown in Figure 6.12(a). Particularly, the collision probability of the proposed model is lower while its transmission probability is higher as compared to the case without backoff freezing (Chatzimisios's model) for both AP and STA as previously clarified.

It is also noticed that under saturation conditions, Tinnirello's model coincides with the proposed model without backoff freezing. In fact, the transmission probability and collision probability of Tinnirello's model remains the same, regardless of whether backoff freezing is considered. This is due to the fact that backoff freezing is accounted by recognizing the existence of anomalous slots, in which their probability of being accessed is much lower as compared to normal slots, after a successful transmission or collision event. Furthermore, the transmission probability and collision probability under non-saturation conditions converge with the transmission probability and collision probability under saturation conditions, regardless of whether backoff freezing is considered. This convergence signifies the correctness of the proposed model which effectively captures both non-saturation and saturation conditions. Particularly, the intersection of both non-saturation and saturation curves can be used as a coarse estimate of capacity analysis. E.g., in the case without backoff freezing, the intersection occurs at the point when there are fifteen STAs, corresponding to a CU of 1.21. Similarly, in the case with backoff freezing, the intersection occurs at the point when there are twenty STAs which is a clear overestimation. Such an overestimation of system capacity or saturation throughput is a direct consequence of Ziouva's backoff freezing which underestimates the collision probability as pointed out by [144]. However, it must be stressed that such an observation holds only for an IBSS with homogeneous STAs under saturation conditions.

Figure 6.13(b) illustrates the effect of different backoff freezing model on the saturation throughput of an IBSS. More specifically, the overestimation of saturation throughput with Ziouva's backoff freezing increases when the contention level is high which could result from either increasing STA or decreasing CW size. This result is consistent with the study in [144]. Another interesting observation is that when the CW size decreases, Tinnirello's model gives higher saturation throughput than Bianchi's model. However, when the CW size increases, Tinnirello's model results in lower saturation throughput than



(a) Transmission and collision probabilities.



(b) Saturation throughput.

Figure 6.13: Effect of freezing for an IBSS under saturation and non-saturation conditions with homogeneous CBR traffic source, HR/DSSS PHY @ 11 Mbps, λ_{nonSAT} = 50 frames/sec, L_{PLD} = 1000 bytes, and BER = 0.

Bianchi's model. Although the difference is fairly small, this is not explicitly pointed out in [144]. When comparing between Bianchi's model without backoff freezing and Tinnirello's model with backoff freezing, it is obvious that backoff freezing has no significant impact on the throughput efficiency. Hence, this explains why most of the existing analytical models developed for an IBSS appear to be accurate, regardless of whether backoff freezing is considered. However, the impact of this modeling inaccuracy has been shown in Figure 6.12 to become more pronounced in the case of an infrastructure BSS. Therefore, careful modeling of backoff freezing is necessary to obtain reliable bounds for admission control or capacity analysis in an infrastructure BSS. As a final note, although Ziouva's backoff freezing underestimates the collision probability for an IBSS with homogeneous STAs, it remains a good approximation for an infrastructure BSS where the AP has a relatively higher load than its associated STAs.

6.5 Chapter Summary

This chapter has proposed a simple, elegant unified analytical model to analyze the IEEE 802.11 DCF infrastructure BSS VoWLAN. Through this analysis, the key performance metrics of MAC delay, PLR, and throughput efficiency have been accurately obtained to enable efficient and effective admission control. Furthermore, they have provided insights into the capacity analysis of VoWLAN where its saturation point can be predicted by capturing the transition from the non-saturation to saturation mode of operation.

This chapter has also addressed the impact of backoff freezing on both IBSS and infrastructure BSS. Particularly, it has revealed that backoff freezing should be properly modeled in order to derive accurate performance metrics and consequently tight bounds for admission control or capacity analysis. Moreover, extensive analyses and simulations have shown that ignoring backoff freezing in an infrastructure BSS will result in overly conservative bounds. In fact, both the improper modeling and ignorant of backoff freezing will manifest as overly conservative bounds, particularly, in heavy load situations.

This implies that residual capacity will be created, which will in turn lead to low network utilization when such models are deployed as admission or load control.

The next chapter gives an overview of the recent IEEE 1900.4 standard. Particularly, the similarities between the TONA handover architecture proposed in Section 3.2 and the IEEE 1900.4 system architecture are discussed. In addition, how the LAS framework in Section 5.5.2, built on the concept of RQB, can be directly applied within the IEEE 1900.4 functional architecture is exemplified through the LAP. Subsequently, a comparative performance analysis of three dynamic load distribution algorithms suitable for multi-AP WLAN or multi-RAT environment is presented. The unified analytical model developed in this chapter will be deployed in the admission controller of an infrastructure BSS VoWLAN as the model-based PQB algorithm. This serves as a benchmark where the performance of the RQB and PLB algorithms will be evaluated.

CHAPTER 7

TOWARD REALIZATION OF THE IEEE 1900.4 STANDARD

From the exposition of the previous chapters, it is clear that vested benefits of future wireless networks can be attained by the cross-fertilization of cooperative and cognitive principles within an advanced RRM architecture. Thus, the motivation of exploiting both cooperation and cognitive functionality, advocated in this thesis, to improve network and end-user performances through better utilization of radio resources in the presence of heterogeneity and convergence associated with future wireless networks will be further explored in this chapter.

The recently approved IEEE 1900.4 standard [26] specifies a policy-based RRM framework in which the decision making process is distributed between network-terminal entities. By exploiting the cooperative exchange of context information between network-terminal entities, the standard aims to facilitate the optimization of radio resource usage to improve the overall composite capacity and QoS of heterogeneous wireless access networks in a multi-RAT landscape. Prior to the standardization of the IEEE 1900.4 standard, the QLO framework presented in Section 5.2.2, which is based on the novel concept of RQB, has been developed to maintain a QoS-balanced system by redistributing network load across a single rate multi-AP WLAN in a self-adjusting manner. The key idea of broadcasting network context information to STAs, which subsequently select an optimal AP to fulfill their QoS requirements dynamically, through the TONA handover architecture exposited in Section 3.2 is similar to the IEEE 1900.4 standard. Furthermore, the LAS framework described in Section 5.5.2, which is an extension of the QLO framework to mitigate rate anomaly in multirate WLAN-based cognitive networks, is directly applicable to the IEEE 1900.4 standard.

In the first half of this chapter, the intricate relationships between the TONA handover architecture and IEEE 1900.4 standard will be discussed. In addition, the relevance of the LAS framework to the IEEE 1900.4 standard is exemplified through the LAP which is implemented based on the distributed radio resource usage optimization use case, stated in Annex A, §A.3 of the IEEE 1900.4 standard, for a multirate multi-AP WLAN. To the best of the author's knowledge, this is one of the first studies to provide insights on the benefits of the recent IEEE 1900.4 RRM framework [26], based on the distributed radio resource usage optimization use case, as compared to the network-distributed RRM framework described in [178] and implemented in [101].

To this end, it is clear that load distribution algorithms such as the iLB scheme, QLO framework, and LAS framework are effective for ensuring uniform traffic distribution in a multi-AP WLAN so as to maximize trunking gain by reducing call blocking probability and maintaining lower average delay in the network, as well as minimizing unnecessary handovers. It is worth to note that these advantages from the suite of RQB algorithms are consistent to the study reported in [179]. Although various load distribution algorithms for WLAN have been investigated in literature, there is a lack of performance comparison between the different algorithms. The remainder of this chapter concentrates on the comparative performance analysis of three main dynamic load distribution algorithms, viz., PLB, PQB, and RQB for a single rate multi-AP WLAN under diverse network conditions. Among these algorithms, the PLB and PQB belong to a class of predictive algorithm while the RQB belongs to a class of reactive algorithm. To the best of the author's knowledge, this is one of the first studies to evaluate the comparative performance between predictive and reactive dynamic load distribution algorithms where the latter is advocated in this thesis.

This chapter is outlined as follows. Section 7.1 gives an overview of the IEEE 1900.4 standard. Section 7.2 relates the TONA handover architecture to the IEEE 1900.4 system architecture, and the LAS framework to the IEEE 1900.4 functional architecture through the LAP. Section 7.3 presents the design rationale and performance evaluation of the LAP. Section 7.4 establishes the motivations and classification of load distribution algorithms. Section 7.5 gives a comparison of three different dynamic load distribution algorithms under study and an exposition on their implementation aspects. Section 7.6 discusses the results of the comparative performance evaluation. Finally, conclusions are drawn in Section 7.7.

7.1 IEEE 1900.4 Standard: An Overview

The IEEE 1900.4 standard specifies a policy-based management framework which supports distributed decision making between network-terminal entities to improve the overall composite capacity and QoS of a composite wireless network (CWN). According to the standard, CWN refers to a coalition of RANs, each connected through a packet (IP) based core network with the deployment of the IEEE 1900.4 entities. To realize this, the IEEE 1900.4 standard defines three use cases, viz., dynamic spectrum assignment, dynamic spectrum sharing, and distributed radio resource usage optimization.

Figure 7.1 illustrates the dynamic spectrum assignment use case for a single operator scenario in which several frequency bands are dynamically assigned to RANs to optimize spectrum usage. Note that in this case, the operator spectrum manager (OSM) entity is required for generating spectrum assignment policies to guide the network reconfiguration manager (NRM) in making dynamic spectrum assignment decisions. As a consequence, both RANs and terminals are reconfigurable according to the dynamic spectrum assignment decisions with the help of both RAN reconfiguration controller (RRC) and terminal reconfiguration controller (TRC), respectively. The dynamic spectrum assignment use case for multiple operator scenarios can be found in Annex A, §A.1.2 and A.1.3 of [26].

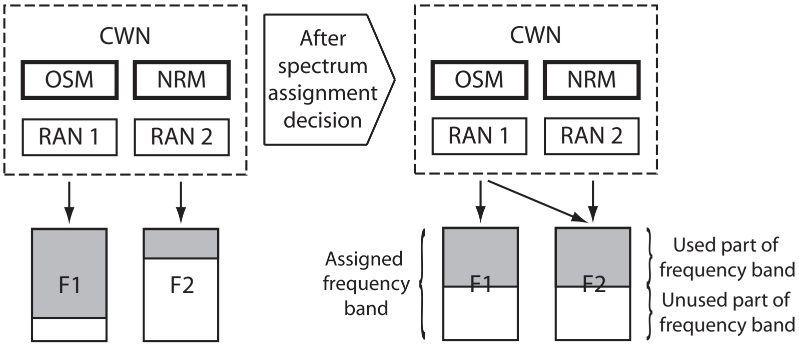

CWN
OSM
NRM
RAN 1
RAN 2
F1
F2
After spectrum assignment decision
CWN
OSM
NRM
RAN 1
RAN 2
Assigned frequency band
F1
F2
Used part of frequency band
Unused part of frequency band

Figure 7.1: Dynamic spectrum assignment use case.

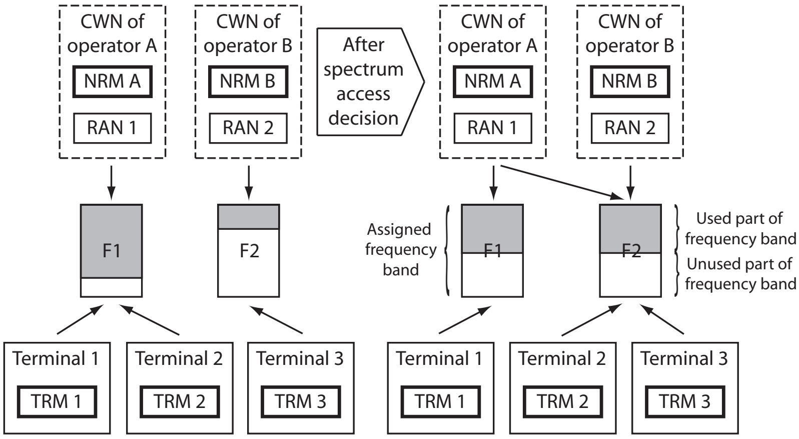


Figure 7.2: Dynamic spectrum sharing use case.

Figure 7.2 depicts the dynamic spectrum sharing use case in which frequency bands are fixed and the OSM entity is not required. However, the frequency bands are available for joint use by operators A and B according to regulatory rules. In other words, the fixed frequency bands can be dynamically shared and/or used by both RANs and terminals, powered by the NRMs, terminal reconfiguration managers (TRMs), and collaborations between them. Similarly, both RANs and terminals are reconfigurable according to dynamic spectrum sharing decisions with the help of both RRC and TRC, respectively.

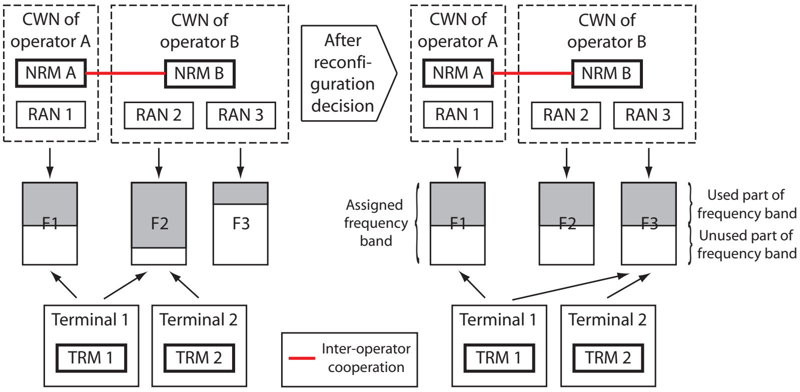


Figure 7.3: Distributed radio resource usage optimization use case.

Figure 7.3 shows the distributed radio resource usage optimization use case for legacy RANs in which frequency bands are fixed and RANs are not reconfigurable. It supports terminals with or without multi-homing capability and allows terminal reconfiguration to be performed in a distributed manner with the help of the TRC. Similarly, distributed radio resource usage optimization is enabled by the NRMs, TRMs, and collaborations between them. Note that in this case, both the OSM and RRC entities are not required. Furthermore, the network reconfiguration decision and control, as well as spectrum assignment evaluation function blocks within the NRM can also be omitted, which will be further discussed in Section 7.2. This thesis focuses on the last use case and shows how the TONA handover architecture and LAP can be easily extended to the other two use cases. In particular, this thesis advocates a generalized CCRRM architecture with the end-to-end goal of promoting a QoS-balanced system to maximize overall composite capacity through the collaborative use of radio resources, realized by the novel concept of RQB. The basic idea is to optimize and exhaust the current network capacity before recourse to acquire additional capacity using the dynamic spectrum assignment and dynamic spectrum sharing use cases. The rationale behind this idea can be easily appreciated by recognizing that any forms of dynamic spectrum access, in general, are expensive, e.g., spectrum leasing and often unpredictable, e.g., spectrum relinquishing. As mentioned in Section 3.2.1, inter-operator cooperation (in red) of Figure 7.3 is not the focus in this thesis. Hence, by inter-network cooperation, it is implicitly assumed that such cooperation exists so that all relevant context information is available for exchange between the access networks of different administrative domains.

7.1.1 System Architecture

To implement the defined use cases, the IEEE 1900.4 standard defines the system architecture as illustrated in Figure 7.4 based on three fundamental system requirements, viz., context awareness, decision making, and reconfiguration. The system architecture comprises of seven entities and six interfaces. Four entities are defined on the network side, viz., OSM, NRM, RAN measurement collector (RMC), and RRC. The OSM is used to reflect the operators' control over the spectrum assignment policies for the NRM. The NRM manages CWN and terminals by generating radio resource selection policies to guide terminals through the optimized radio resource allocations. The RMC acquires RAN context information for the NRM. The RRC, which acts upon the NRM requests, is responsible for the reconfigurations of RANs. Note that according to the standard, the NRM, RMC, and RRC may also be implemented in a distributed manner.

Three entities are defined on the terminal side, viz., TRM, terminal measurement collector (TMC), and TRC. The TRM manages the terminal by making the final decision regarding radio resource allocations within the bounds of guiding policies defined by the NRM, user preferences, and the available context information. The TMC acquires terminal context information for the TRM. The TRC, which acts upon the TRM requests, is responsible for the reconfigurations of terminals. Of all the seven defined entities, both NRM and TRM are the key decision making entities where the exchanges of RAN context information and terminal context information, including the dissemination of network policies are carried out by the radio enabler (RE). The six key interfaces defined in the IEEE 1900.4 standard, viz., interface between NRM and TRM, TRM and TRC, TRM and TMC, NRM and RRC, NRM and RMC, and lastly NRM and OSM are listed in Table 7.1. In this

Table 7.1: IEEE 1900.4 key interfaces between entities.

Between NRM and TRM	**Between TRM and TRC**	**Between TRM and TMC**
From NRM to TRM	From TRM to TRC	From TRM to TMC
● Radio resource selection policies	● Terminal reconfiguration requests	● Terminal context information requests
● RAN context information		
● Terminal context information		
From TRM to NRM	From TRC to TRM	From TMC to TRM
● Terminal context information	● Terminal reconfiguration responses	● Terminal context information
Between NRM and RRC	**Between NRM and RMC**	**Between NRM and OSM**
From NRM to RRC	From NRM to RMC	From NRM to OSM
● RAN reconfiguration requests	● RAN context information requests	● Spectrum assignment policies
From RRC to NRM	From RMC to NRM	From OSM to NRM
● RAN reconfiguration responses	● RAN context information	● Information on spectrum assignment decisions

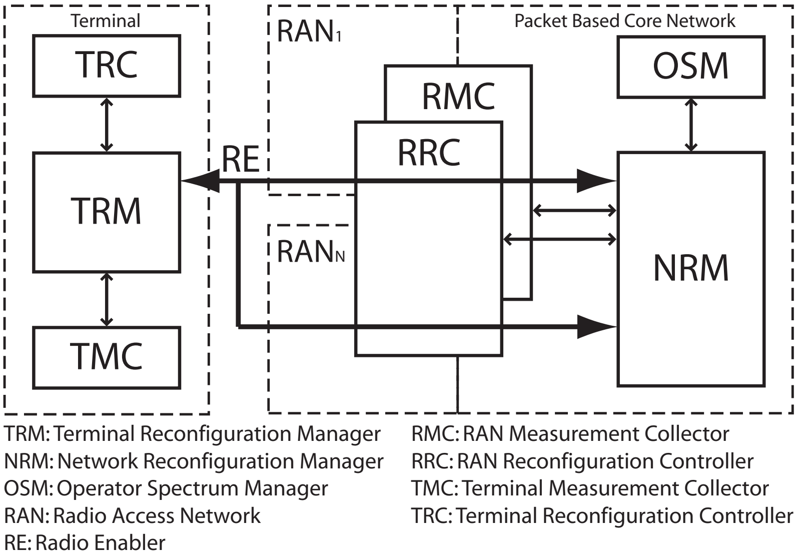

Terminal
TRC
TRM
TMC
RE
RAN1
RANN
RMC
RRC
Packet Based Core Network
OSM
NRM
TRM: Terminal Reconfiguration Manager
NRM: Network Reconfiguration Manager
OSM: Operator Spectrum Manager
RAN: Radio Access Network
RE: Radio Enabler
RMC: RAN Measurement Collector
RRC: RAN Reconfiguration Controller
TMC: Terminal Measurement Collector
TRC: Terminal Reconfiguration Controller

Figure 7.4: IEEE 1900.4 system architecture.

thesis, the focus is on the implementation aspects of the LAP associated with both NRM and TRM. Readers are referred to [180] for an excellent coverage on the development aspects of the RE.

7.1.2 Functional Architecture

The IEEE 1900.4 standard also defines the functional architecture depicted in Figure 7.5 which focuses on the key functions within the NRM and TRM as they are the key decision making entities. Six functional blocks are defined in the NRM, viz., network reconfiguration decision and control function, spectrum assignment evaluation function, policy derivation function, policy efficiency evaluation function, information extraction, collection, and storage function, and RAN selection function. The network reconfiguration decision and control function makes decisions on RAN reconfiguration compliant with the spectrum assignment policies received from the OSM, and subsequently it sends the corresponding reconfiguration requests to the RRC. In addition, it sends the outcome of the reconfiguration decisions back to the OSM. The spectrum assignment evaluation function evaluates the efficiency of spectrum usage under the prevailing spectrum assignment, and

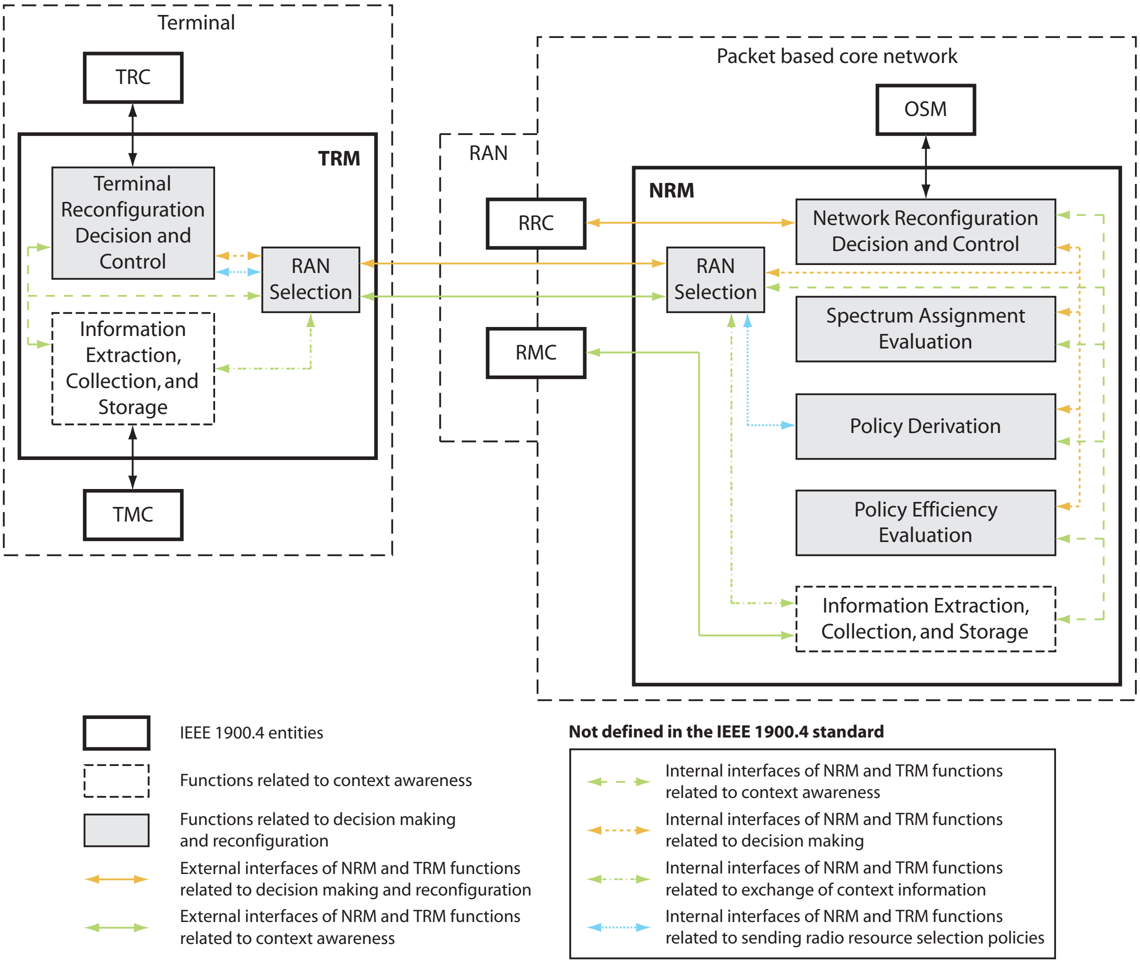

Terminal
TRC
TRM
Terminal Reconfiguration Decision and Control
RAN Selection
Information Extraction, Collection, and Storage
TMC
RAN
RRC
RMC
Packet based core network
OSM
NRM
Network Reconfiguration Decision and Control
RAN Selection
Spectrum Assignment Evaluation
Policy Derivation
Policy Efficiency Evaluation
Information Extraction, Collection, and Storage
IEEE 1900.4 entities
Functions related to context awareness
Functions related to decision making and reconfiguration
External interfaces of NRM and TRM functions related to decision making and reconfiguration
External interfaces of NRM and TRM functions related to context awareness
Not defined in the IEEE 1900.4 standard
Internal interfaces of NRM and TRM functions related to context awareness
Internal interfaces of NRM and TRM functions related to decision making
Internal interfaces of NRM and TRM functions related to exchange of context information
Internal interfaces of NRM and TRM functions related to sending radio resource selection policies

Figure 7.5: IEEE 1900.4 functional architecture.

the outcome of this evaluation may be used by the network reconfiguration decision and control function to make future decisions on RAN reconfiguration. The policy derivation function generates radio resource selection policies to guide the TRMs in terminals' reconfiguration decisions. The radio resource selection policies are derived according to the context information from the information extraction, collection, and storage function. The policy efficiency evaluation function evaluates the efficiency of prevailing radio resource selection policies, and the outcome of this evaluation may be used by the policy derivation function to generate future radio resource selection policies. The information extraction, collection, and storage function receives, processes, and stores both RAN context information and terminal context information. The RAN context information is received from the RMC while the terminal context information is received from the TRM periodically, based on request, or based on event. The information extraction, collection, and storage function provides information to functions within the NRM, forwards the RAN context information to the TRM, and may also forward the terminal context information that is related to other terminals to the TRM. The RAN selection function selects RANs for exchanging radio resource selection policies and context information between the NRM and TRM through the RE. The exchanges are carried out in a localized manner, e.g., based on geo-locations and topology information to ensure timely delivery of radio resource selection policies and context information, as well as minimize signaling load. The network reconfiguration decision and control function, spectrum assignment evaluation function, policy derivation function, policy efficiency evaluation function, and RAN selection function cooperate during their operation by leveraging on information from the information extraction, collection, and storage function.

Three functional blocks are defined in the TRM, viz., terminal reconfiguration decision and control function, information extraction, collection, and storage function, and RAN selection function. The terminal reconfiguration decision and control function makes decisions on terminal reconfiguration based on the radio resource selection policies received from the NRM, and subsequently it sends the corresponding reconfiguration requests to the TRC. Similarly, the information extraction, collection, and storage function receives, processes, and stores both terminal context information and RAN context information. The terminal context information is received from the TMC while the RAN context information is received from the NRM periodically, based on request, or based on event. Additionally, the terminal context information regarding other terminals may be received from the NRM. The information extraction, collection, and storage function provides information to functions within the TRM and forwards the terminal context information to the NRM. The RAN selection function selects RANs for exchanging radio resource selection policies and context information between the TRM and NRM through the RE. The terminal reconfiguration decision and control function and RAN selection function cooperate during their operation by leveraging on information from the information extraction, collection, and storage function.

7.2 TONA Handover Architecture and Load Adaptation Policy

The main idea of the IEEE 1900.4 is enabling terminals to participate in the decision making process autonomously while adhering to some policies and constraints imposed by the network. The LAP first proposed in [181] is based on such distributed decision making process between network-terminal entities by leveraging on the TONA handover architecture to exploit the cooperative exchange of context information as depicted in Figure 7.6. As exposited in Section 3.2.2, the TONA handover architecture supports: (i) network-assisted discovery where the source AP broadcasts the network context information of neighboring APs together with its own and recommended LAP; and (ii) terminal-oriented decision where terminals make the *final* RRM decision in selecting an optimal AP to fulfill their user preferences and QoS requirements while operating within the bounds of the recommended LAP. In effect, the TONA handover architecture corresponds to the IEEE 1900.4 system architecture in the following aspects.

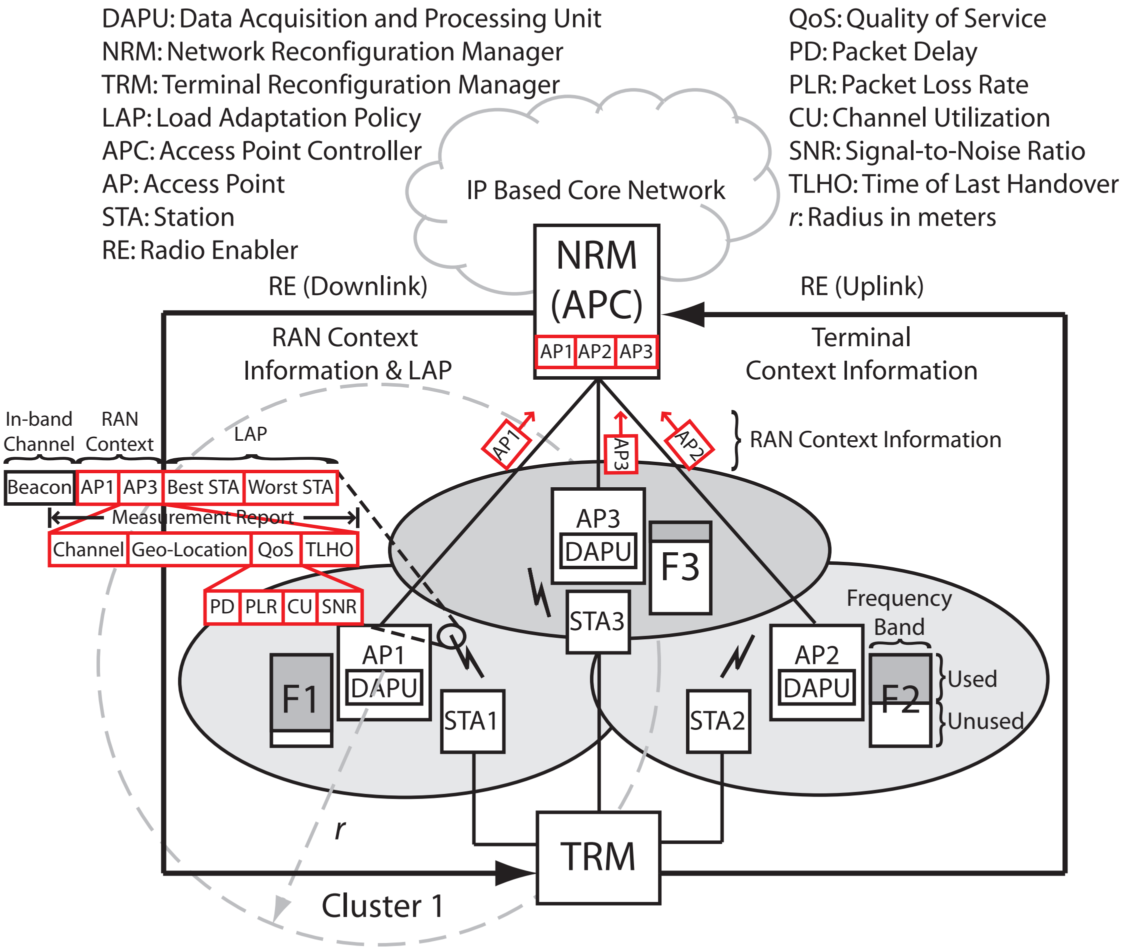


Figure 7.6: Relationships between the TONA handover architecture and the IEEE 1900.4 system architecture.

First, the measurement report, which consists of the QoS context information, time of last handover event, channel number, and geo-location of the AP, is analogous to the RAN context information. Second, the QoS context information comprises of PD, PLR, CU, and SNR. These are obtained from the DAPU described in Section 3.4.1 (cf. Figure 3.6) which is analogous to both RMC and TMC. Third, the measurement report is periodically transmitted to the APC which is analogous to the NRM. Fourth, the APC would collect these measurement reports from every AP in its subnet and facilitate the exchange of measurement reports between different subnets through the IP-based core network should multiple NRMs exist as stated in the standard. Fifth, the consolidated measurement reports of the source and neighboring APs would be disseminated from the APC using cluster-based broadcast based on the geo-location of both AP and terminals (or TRMs) as suggested in the standard. The cluster is defined as a group of 'reachable' APs and terminals bounded by the cluster radius r w.r.t. the geo-location of the source AP. The motivation is to dispense terminal from monitoring the network conditions of distant APs which are 'unreachable'. This is in addition to the standard which requires only the NRM to send radio resource selection policies and context information to selected terminals within the geo-localized cluster. Sixth, the LAP is analogous to the radio resource selection policies provided by the NRM to the TRMs. Last, the RE would be mapped to the beacon (in-band channel) of the source AP for broadcast as one of the two options specified by the standard.

Given that the LAP is an avatar of the LAS framework implemented within the IEEE 1900.4 architecture, the design philosophy is also based on the fundamental principle of RQB (cf. Section 4.2). In essence, the LAP is responsible for synergetic interactions within the MAC layer to optimize load distribution opportunistically, and between the PHY and MAC layer to exploit the benefits of both link adaptation and load adaptation on-demand. Through the TONA handover architecture, the QoS context information of each AP is cooperatively exchanged between network-terminal entities to facilitate the joint optimization of radio resource usage. The network-QoS entity, which consists of service prioritization and admission control to deal with different user service profiles, resides in each AP (or RAN). On the other hand, connection-QoS entity, which consists of network selection and handover control to deal with dynamic network conditions, resides in each terminal. Note that dynamic network conditions could be associated with network congestions, channel impairments, and dynamic spectrum assignment/sharing. The details of the LAP, which are the same as the LAS framework, can be found in Section 5.5.2.

The motivation here is to exemplify the algorithmic implementations of the LAP in the NRM and the load adaptation decision in the TRM of the IEEE 1900.4 functional architecture as illustrated in Figure 7.7 and Figure 7.8, respectively. It is important to note that these algorithmic implementations, inherited from the LAS framework, resemble and in fact are compliant to the IEEE 1900.4 functional architecture discussed in Section 7.1.2. To be more specific, in the NRM, the HCUFO policy is analogous to the policy derivation function in which candidate STA is being nominated to perform either planned or unplanned handover. In addition, RQB is analogous to the policy efficiency evaluation function in which QoS balance has been shown and advocated in the previous chapters as the criterion to quantify the state of balance in a CWN when deploying load distribution algorithms. Furthermore, bootstrap approximation in conjunction with Bayesian learning, which is employed in the NRM for both PD and PLR to obtain the required network quality probabilities for network selection (cf. Section 3.4) in the TRM, are analogous to the NRM's information extraction, collection, and storage function. In the TRM, the load adaptation decision is analogous to the terminal reconfiguration decision and control function. Moreover, the terminal context information such as user preferences and QoS requirements is analogous to the TRM's information extraction, collection, and storage function.

To this end, a detailed exposition on the implementation aspects of the distributed radio resource usage optimization use case based on the LAP has been given. Without loss of generality, the TONA handover architecture and LAP can be easily extended to the

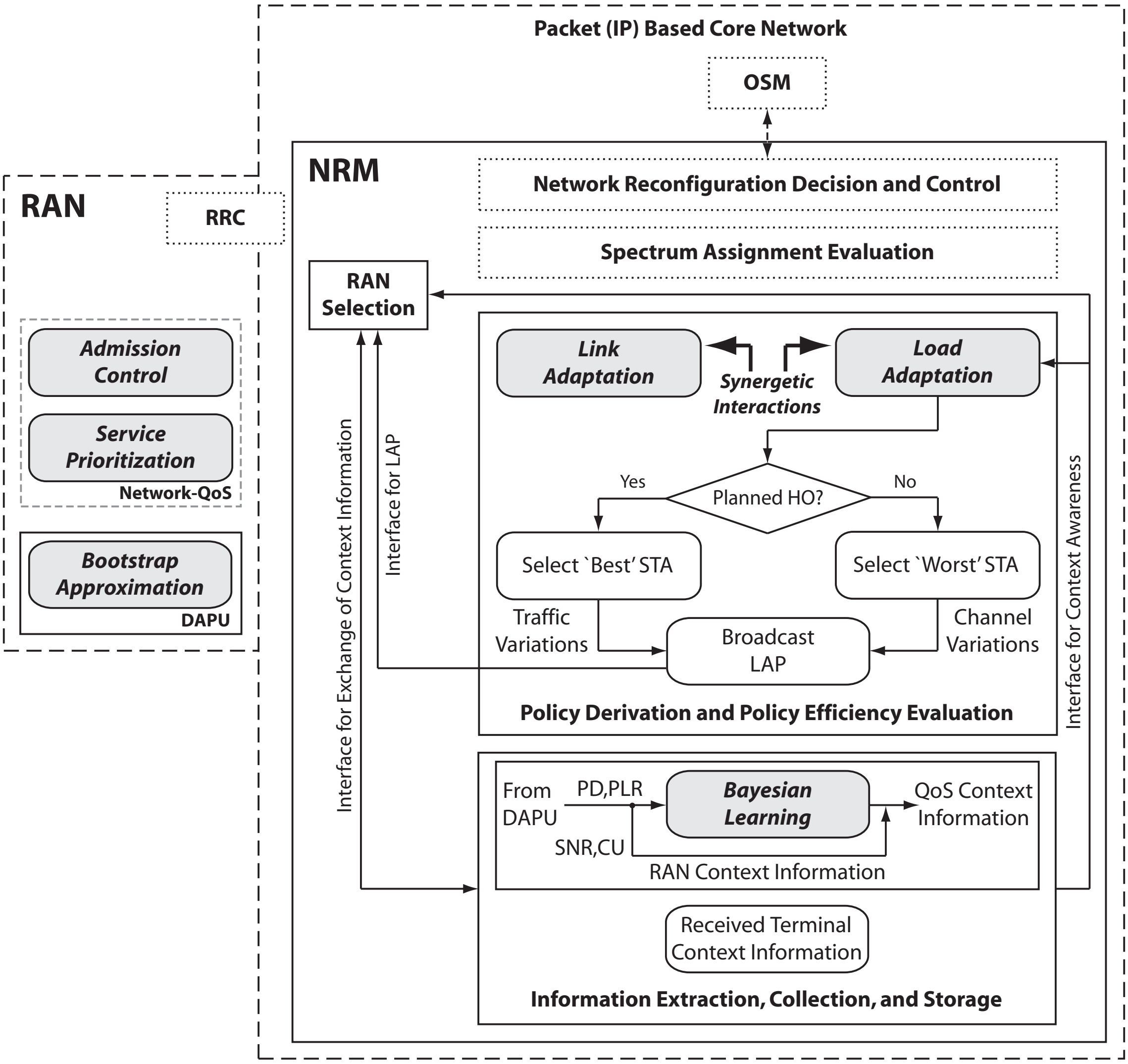


Figure 7.7: LAP implementation in the NRM of the IEEE 1900.4 functional architecture.

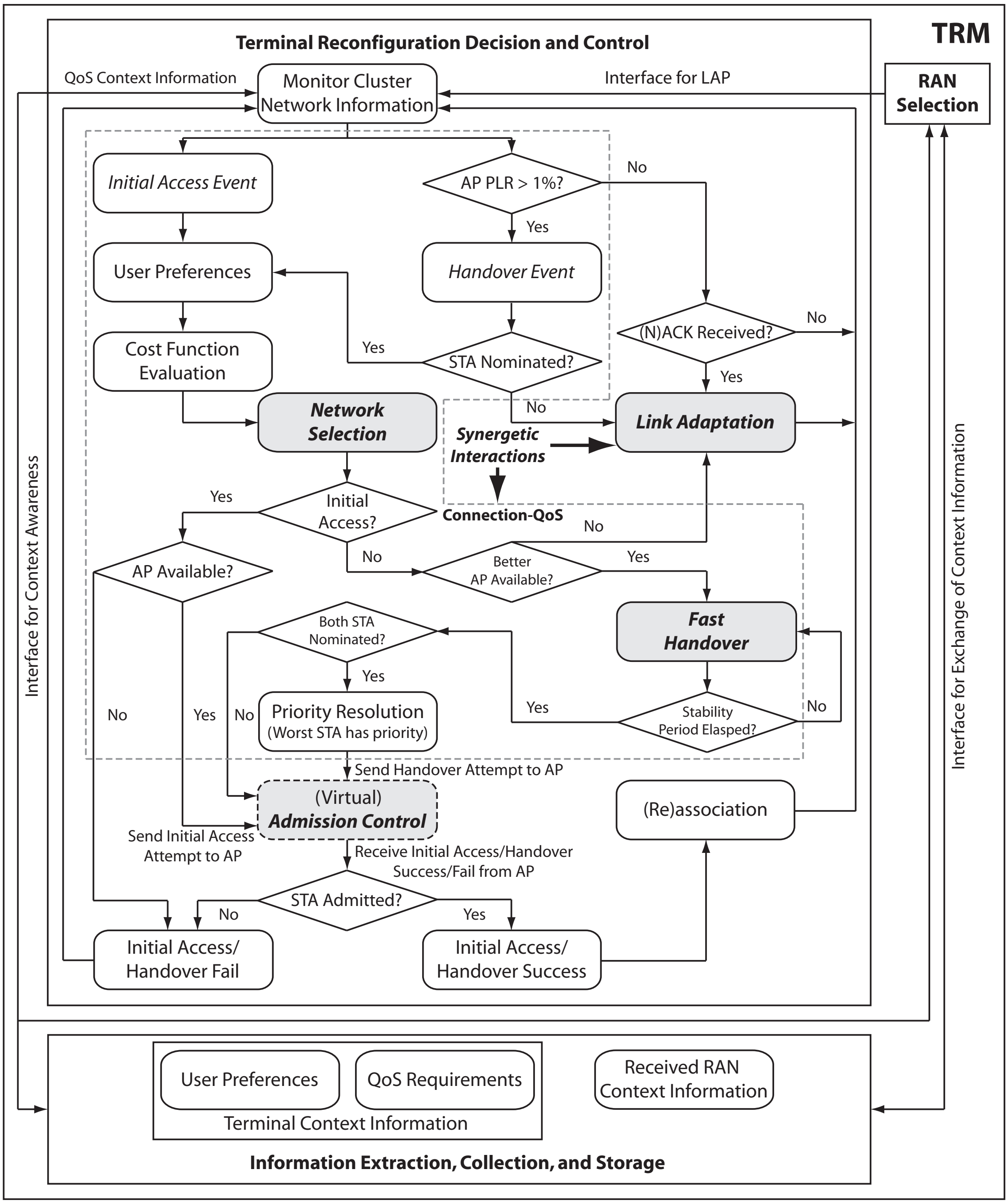

TRM
Terminal Reconfiguration Decision and Control
QoS Context Information
Monitor Cluster Network Information
Interface for LAP
RAN Selection
Initial Access Event
AP PLR > 1%?
No
Yes
User Preferences
Handover Event
(N)ACK Received?
No
Yes
Cost Function Evaluation
STA Nominated?
Yes
No
Network Selection
Link Adaptation
Synergetic Interactions
Connection-QoS
Initial Access?
Yes
No
AP Available?
Better AP Available?
No
Yes
Fast Handover
Both STA Nominated?
Yes
Priority Resolution (Worst STA has priority)
Stability Period Elasped?
Yes
No
No
Yes
No
Send Handover Attempt to AP
(Virtual) Admission Control
(Re)association
Send Initial Access Attempt to AP
Receive Initial Access/Handover Success/Fail from AP
STA Admitted?
No
Yes
Initial Access/ Handover Fail
Initial Access/ Handover Success
Interface for Context Awareness
Interface for Exchange of Context Information
User Preferences
QoS Requirements
Terminal Context Information
Received RAN Context Information
Information Extraction, Collection, and Storage

Figure 7.8: Load adaptation decision implementation in the TRM of the IEEE 1900.4 functional architecture.

dynamic spectrum assignment and dynamic spectrum sharing use cases. To realize this, two additional IEEE 1900.4 entities, viz., OSM and RRC, as well as two additional functional blocks (in dotted lines) inside the NRM, viz., network reconfiguration decision and control function and spectrum assignment evaluation function, all on the network side are required. Nevertheless, the effects of both dynamic spectrum assignment and dynamic spectrum sharing will be simulated for a preliminary study in the next section.

7.3 Performance Evaluation of the LAP

The design rationale of the LAP is to overcome: (i) the inherent channel impairments apparent in legacy WLAN deployments at hotspots and indoor environments in which structures and obstacles cause frequent NLOS transmissions; and (ii) the effects of dynamic spectrum assignment/sharing capabilities supported by reconfigurable RANs and terminals, corresponding to two of the defined use cases in the IEEE 1900.4 standard. The dynamic spectrum assignment/sharing results in opportunistic access to diverse channels of the available frequency spectrum which will have largely different propagation characteristics. Moreover, although dynamic spectrum assignment/sharing exploits the spectrum holes of primary users opportunistically, the secondary users are mandated to utilize the spectrum only when their transmissions do not interfere with that of the primary users. This hard requirement means that the availability of spectrum and the corresponding system capacity would essentially be time-varying, depending on the load of the primary users. In a nutshell, these issues imply that a suitable LAP must be in place to manage fruitful utilization of heterogeneous channels. Particularly, the LAP aims to answer two important questions of: (i) what happens when additional spectrum becomes available; and (ii) more importantly, what happens when such additional spectrum ceases to exist. These answers will enable one to maximize system capacity during condition (i) and prevent any serious QoS degradation during condition (ii).

Figure 7.9 illustrates the existence of a desired operating region which can satisfy the QoS requirements of end-users. Note that this desired operating region is lower bounded by the horizontal axis, vertical axis, and the arc of the QSF unit quadrant. Ideally, according to the definition in (A-7a) of Appendix A-3.1, QSF should be greater than 1 in order to meet the QoS requirements of end-users. This leads to the question of how to operate networks in the desired region. As explained in Section 5.5.2, the LAP relies on the concept of RQB to distribute network load based on opportunistic yet altruistic exploitation. On the other hand, the HCUFO policy will alleviate the affected network rapidly by transferring the most aggressive traffic source to another network of a better quality as the affected network will recover once its load reduces and/or wireless channel conditions improve according to Figure 7.9. Hence, when the affected network sustains prolonged degradation of wireless channel conditions or capacity outages due to rate anomaly, the 'best' possible option is to transfer most of its load to an alternative network over the least time. Such a policy has two significant advantages. First, it reduces the number of handover events since transferring the most aggressive traffic source implies that it requires fewer of such transfers in order to reach the recovery state as compared to transferring the less aggressive traffic source. Second, it results in the long-term uniform distribution of aggressive traffic sources which is attractive from a load distribution perspective.

The OPNET™ Modeler® 14.0 with wireless module is used in this simulation study. The simulation scenario is a typical hotspot which consists of a homogeneous multirate multi-AP WLAN-based cognitive network with three IEEE 802.11g APs, each operating at an initial data rate of 54 Mbps as in the LAS framework. An error-prone channel is considered by including shadow fading, multipath, and variable path loss exponent according to Appendix A-2.2 to capture different propagation characteristics as a result of NLOS transmissions and/or dynamic spectrum assignment/sharing. A balanced load of three voice, three video, and three FTP STAs in each AP is introduced as the motivation is to investigate the consequence of load distribution under varying capacity and wireless channel conditions. Specifically, this study simulates: (i) capacity variations by intro-

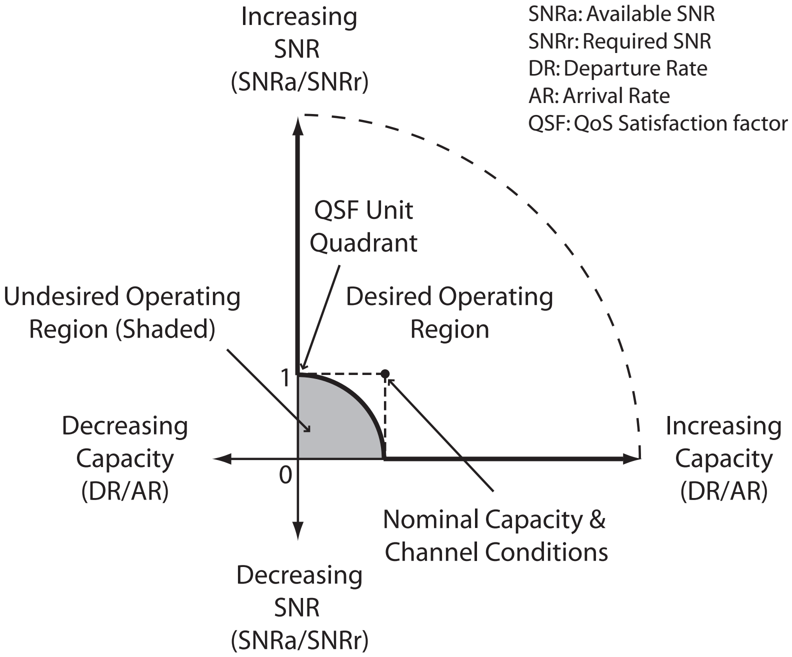


Figure 7.9: Design rationale of the LAP.

ducing the effects of dynamic spectrum assignment/sharing in each AP; and (ii) wireless channel variations by introducing NLOS transmissions and the effects of dynamic spectrum assignment/sharing in each BSS. The capacity and wireless channel variations are simulated according to Table 7.2. For states 1 – 7, it is assumed that an AP with a reduced capacity of 6 Mbps has a high average SNR of 35 dB due to its more robust modulation and coding scheme while an AP with a higher capacity of 54 Mbps may have either a high average SNR of 35 dB or low average SNR of 15 dB depending on propagation characteristics. Shadow fading and multipath are included for the entire simulation duration, and multimedia traffic sources are simulated according to Table 5.1. The QoS performance of the LAP is examined based on the QSF, TFI, and QBI defined in Appendix A-3.1. Further details on the general simulation models can be found in Appendix A-2.

The effectiveness of the IEEE 1900.4 RRM, based on the LAP, is evaluated in terms of the average aggregate throughput of system, QBI of STAs' QSF, TFI of STAs, and QSF of STAs, as well as the total number of handover events. A comparison between the IEEE

Table 7.2: Capacity and wireless channel variations.

State	Data Rate (Mbps)			Avg. SNR (dB)		
	AP1	AP2	AP3	AP1	AP2	AP3
1	54	54	54	35	35	15
2	54	54	6	35	15	35
3	54	6	54	15	35	15
4	54	6	6	15	35	35
5	6	54	54	35	15	15
6	6	54	6	35	15	35
7	6	6	54	35	35	35
8	6	6	6	15	35	35

1900.4 RRM and the network-distributed RRM is also presented. The latter is based on a typical LBM implementation in [101] which utilizes absolute network load, particularly, the CU as the only load metric (cf. Figure 5.18). Ideally, according to the definitions of the three KPIs in (A-7) – (A-9), the QSF should be greater than 1, the TFI should be close to 0, and the QBI should be close to 1 so as to offer QoS guarantee, throughput fairness, and QoS fairness, respectively. The results are analyzed in eight states, which correspond to the simulated scenario, starting from 100 s (0 – 100 s is the warm-up period).

In general, it is observed that the average aggregate throughput, QBI, and TFI of LAP outperform LBM by 15%, 23%, and 48%, respectively as shown in Figures 7.10 – 7.12 for the simulated scenario. However, LBM has higher average QSF than LAP in states one, two, five, and six as shown in Figure 7.13. This counterintuitive result is a direct consequence of the design philosophy of RQB to preclude unnecessary handovers when the QoS requirements of STAs can be achieved. In state one, LBM does not trigger any handovers as the CU of all three APs is balanced. However, LAP is aware of the low SNR in AP 3 and triggers five handovers to redistribute load from AP 3 to both AP 1 and AP 2 which have better SNR. Although LBM has higher average QSF than LAP, the aggregate throughput of LAP outperforms LBM by 22%, the QBI of LAP outperforms LBM by 36%, and the TFI of LAP outperforms LBM by 89%.

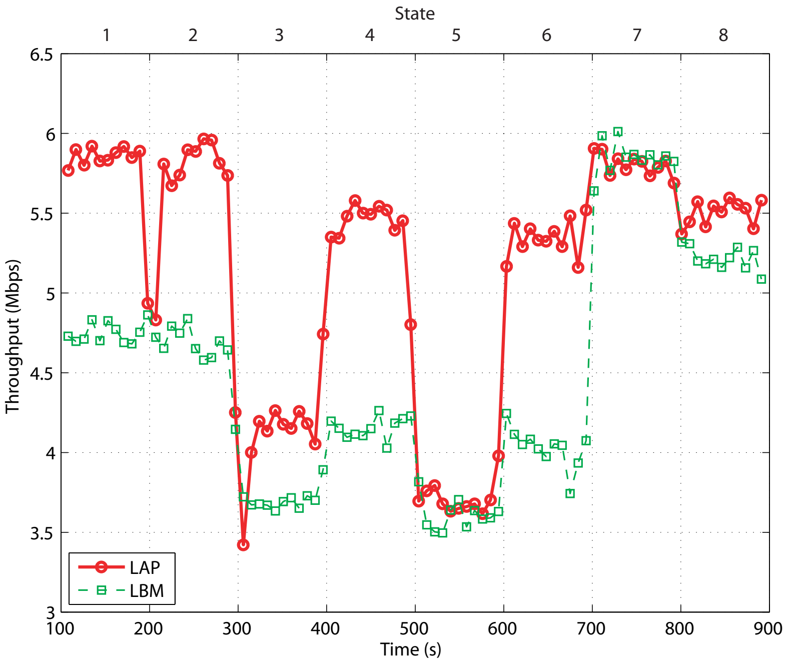


Figure 7.10: Average aggregate throughput of system.

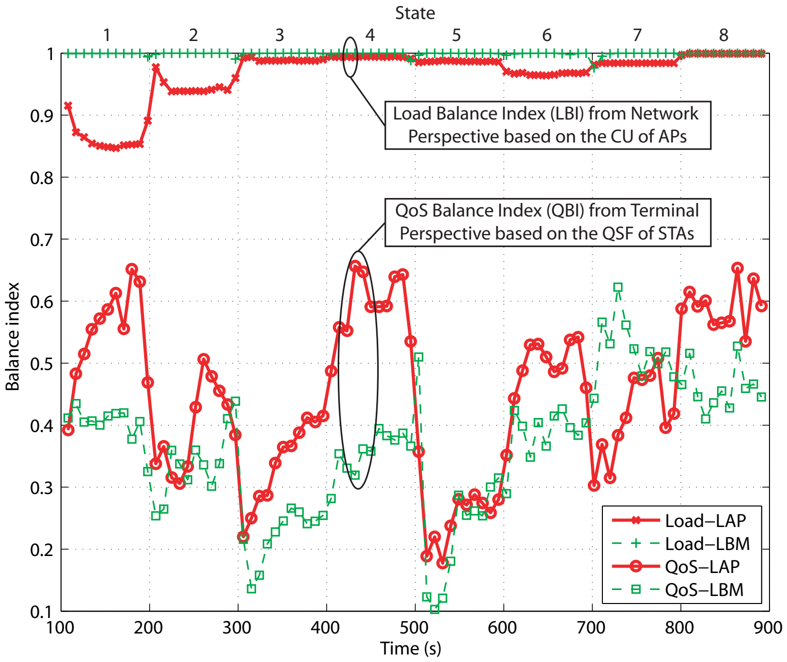


Figure 7.11: Average QBI of STAs' QSF and LBI of APs' CU.

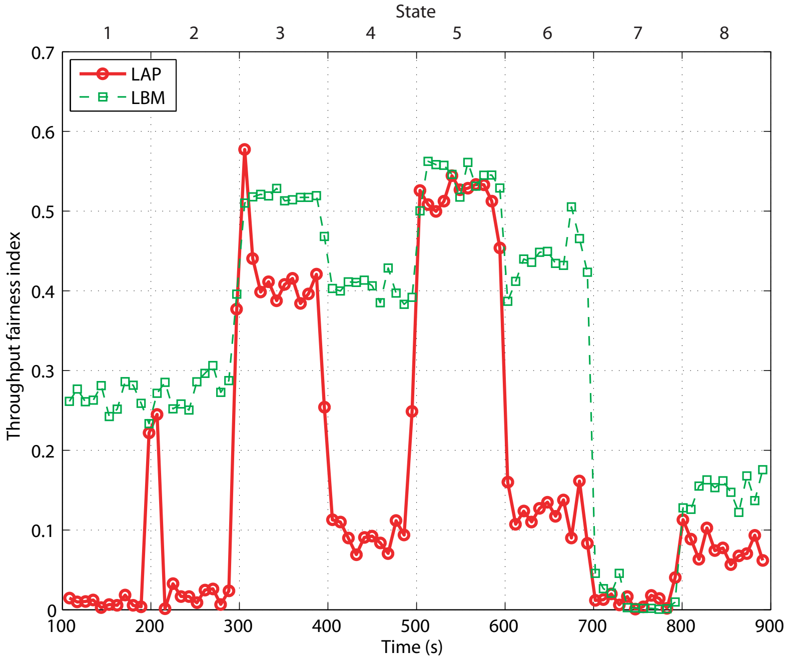


Figure 7.12: Average TFI of STAs.

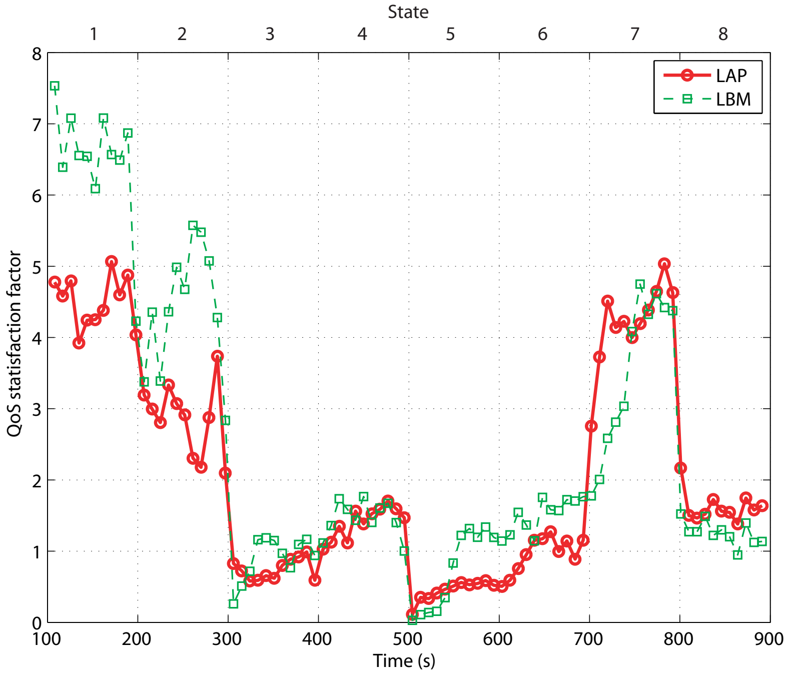


Figure 7.13: Average QSF of STAs.

In state two, LBM triggers five handovers from AP 3 to both AP 1 and AP 2 as the capacity of AP 3 is reduced. However, it is important to note that load distribution based on the CU as the only load metric is catastrophic as the load, in this case, is ‘wrongly’ transferred to AP 2 with low SNR. This has a strong negative impact on the aggregate throughput, QBI, and TFI. On the other hand, LAP triggers four handovers to transfer the load from AP 2 to AP 3 instead since it is aware that AP 3 has higher SNR than AP 2, and it can still accept some additional load after some of its original load has been redistributed in state one. Although LBM has higher average QSF than LAP, the aggregate throughput of LAP outperforms LBM by 21%, the QBI of LAP outperforms LBM by 17%, and the TFI of LAP outperforms LBM by 75%.

In states three and five, both LAP and LBM have an average QSF of less than 1. The compromise of QSF is unavoidable as these two states simulate the extreme case where the capacity of one AP is reduced and the alternative APs have low SNR. LBM, agnostic of the wireless channel conditions, continues to transfer load to APs with low SNR by triggering eleven handovers in state three and twelve handovers in state five. On the contrary, LAP does not transfer any load since it is aware of the low SNR in the alternative APs which inhibits any exploitation of heterogeneity. Under such conditions, it is worth to highlight that the aggregate throughput of LAP still outperforms LBM by 12%, the QBI of LAP also outperforms LBM by 52%, and likewise the TFI of LAP outperforms LBM by 20% in state three, whereas similar performances between LAP and LBM are observed in state five.

States four and six correspond to the scenario where the capacity of two APs is reduced and the alternative AP suffers low SNR. LBM triggers eight handovers from AP 2 and AP 3 to AP 1 with low SNR in state four, and it triggers eleven handovers from AP 1 and AP 3 to AP 2 with low SNR in state six. On the other hand, LAP does not trigger any handovers in state four while one handover is triggered from AP 2 to AP 1 in state six. Although LBM has higher average QSF than LAP in state six, the aggregate throughput of LAP outperforms LBM by 33%, the QBI of LAP outperforms LBM by 26%, and the

TFI of LAP outperforms LBM by 72%. As for state four, the aggregate throughput of LAP outperforms LBM by 30%, the QBI of LAP outperforms LBM by 66%, and the TFI of LAP outperforms LBM by 74%.

State seven corresponds to the scenario where the capacity of two APs is reduced and all APs have high SNR. LBM triggers five handovers from AP 1 and AP 2 to AP 3, but this time AP 3 has high SNR. On the other hand, LAP does not trigger any handovers since QoS requirements of STAs associated with AP 1 and AP 2 can still be met. This is the very reason why the QBI of LBM is higher than LAP as it tries to perform load balancing proactively, whereas LAP takes the reactive approach to trigger handover *only* when the QoS requirements of STAs cannot be supported. It is worth noting that, in this case, the aggregate throughput, TFI, and QSF of both LAP and LBM are similar. Moreover, it is important to emphasize that the QoS requirements of STAs are not compromised as the average QSF of LAP is still greater than 1. Finally, state eight represents the scenario where the capacity of all APs is reduced and one of the APs has low SNR. Under such conditions, the QSF of LAP outperforms LBM by 28%, the aggregate throughput of LAP outperforms LBM by 5%, the QBI of LAP outperforms LBM by 29%, and the TFI of LAP outperforms LBM by 47%.

7.3.1 Discussions

Although LBM is balanced from the network perspective in terms of the CU as shown in Figure 7.11, it fails to provide QoS fairness in six out of the eight simulated states, particularly, when the target AP has a low SNR. In essence, low QBI of LBM implies that there is a huge disparity in the QoS or throughput between STAs of the same service class. In fact, Figure 7.12 confirms that LBM also fails to offer throughput fairness in six out of the eight simulated states. Evidently, the QoS-agnostic LBM is incapacitated by wireless channel variations which could arise due to NLOS transmissions in legacy WLAN and/or dynamic spectrum assignment/sharing supported by reconfigurable RANs

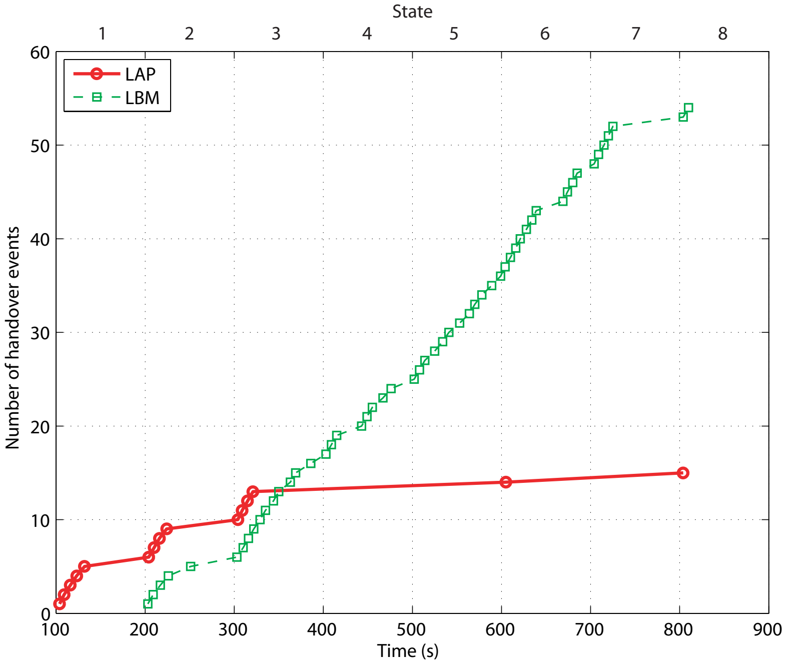


Figure 7.14: Total number of handover events.

and terminals. This has serious implications on the three defined use cases in the IEEE 1900.4 standard, which focus on load balancing of available spectrum. It is important to stress that these simulation results underpin that any feasible spectrum load balancing shall be QoS-aware. This can be realized by the cooperative exchange of QoS context information within the CLM. Such cooperation will facilitate the joint optimization of network-terminal distributed decision making to orchestrate informed VHO and dynamic load distribution, supported by the generalized CCRRM architecture. As a final note, it is imperative to highlight that there are a total of fifty-four handovers triggered by LBM but only fifteen handovers triggered by LAP as shown in Figure 7.14. This translates to a massive 72% reduction in handover while achieving high overall composite capacity and maintaining a QoS-balanced system.

Therefore, this thesis advocates QoS balance as the criterion to quantify the state of balance in a CWN from the terminal or end-user perspective based on perceived QoS, i.e., the QSF of STAs instead of from the network perspective based on absolute network load, i.e., the CU of APs. Moreover, RQB provides statistical QoS guarantee for all states except three and five which are unavoidable. Although not explicitly mentioned, the performance gains of LAP in this multirate WLAN scenario is also attributed to the mitigation of the rate anomaly problem by RQB as in the LAS framework.

7.4 Why Load Distribution?

Future wireless networks will enjoy ubiquitous connectivity by taking advantage of the IP core convergence which is seen as the '*epoxy resin*' of heterogeneous access networks ecosystem. It is expected that the prevalence of WLAN and the advent of the IEEE 802.11n standard [87] will continue to offer compelling opportunities. Therefore, WLAN will be considered as one of the de-facto wireless access networks. However, it is known that wireless network conditions, in general, are diverse owing to both traffic and wireless channel variations. Moreover, the delivery of QoS-demanding applications such as in VoWLAN is very challenging as shown in Chapter 4, particularly, in the context of future IP-based wireless networking scenario where hotspots of multi-AP are physically co-located. This raises the importance of exploiting heterogeneity across a multi-AP WLAN, which requires an advanced network control mechanism to effectuate uniform load distribution, so that the QoS of end-users and overall composite capacity can be improved.

First and foremost, it is important to recognize that differing objectives between network operators and end-users exist. In general, network operators are motivated to maximize their revenue by maintaining high network utilization while end-users demand good perceived QoS. Hotspots are typically deployed to cope with heightened traffic demands. However, the overall composite capacity will not scale with the increasing number of APs when STAs select an AP based only on RSSI, without any QoS considerations such as load control or an appropriate network control mechanism such as admission control. This problem is further complicated by the typical non-uniform load distribution across the APs in public hotspots such as convention centers and airports where end-users tend

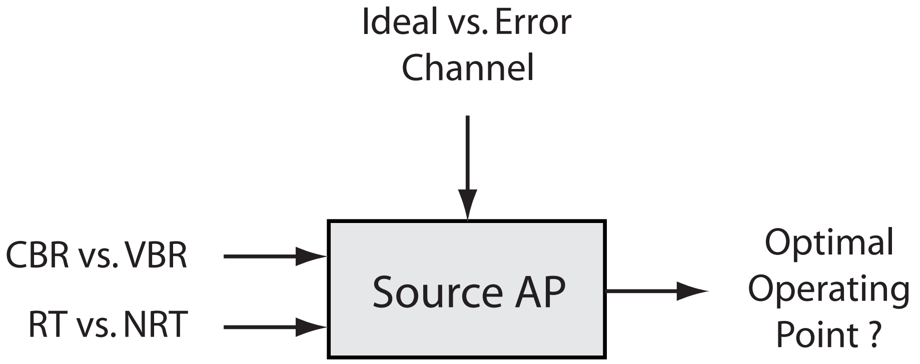


Figure 7.15: Non-unique optimal operating point: Saturation point of WLAN depends on the class of services, the type of traffic sources, the proportion of traffic mixes, and prevailing wireless channel conditions.

to correlate temporally and spatially. Consequently, these cause sporadic congestions in the AP with the strongest signal strength.

Next, the problem of finding an optimal operating point of WLAN as depicted in Figure 7.15 is not trivial due to different traffic mixes and channel impairments arising from frequent NLOS transmissions caused by structures and obstacles, which are commonplace in hotspot and indoor WLANs. These uncertainties will result in *non-unique* saturation points in which the QoS will inevitably deteriorate when network is driven beyond these points. Therefore, load and/or admission control must be incorporated in such multi-AP hotspots so that heterogeneity can be exploited to harness overall composite capacity and QoS improvements. The context of heterogeneity here refers to the dynamic network conditions in the AP due to both traffic and wireless channel variations. The former could depend on the class of services, e.g., RT and NRT, the type of traffic sources, e.g., CBR and VBR, and the proportion of traffic mixes whilst the latter could depend on different propagation and fading environments.

Load distribution will become more imperative in future wireless networks, which may comprise of highly heterogeneous technologies such as WLAN, WiMAX, and LTE systems, since the key motivation of such integration is to exploit all possible heterogeneity within complementary access networks to orchestrate better utilization of radio resources and provide better end-user experiences. Hence, the idea of effectuating uniform load distribution will continue to serve as a fundamental solution due to its advantages of maximizing trunking efficiency by reducing call blocking probability and maintaining lower average access delay, as well as minimizing unnecessary handovers. Without loss of generality, the optimal operating point model of Figure 7.15 can also be used to derive capacity bounds for various access networks and facilitate load distribution, albeit, it will pose a more significant challenge in such a multi-RAT environment. Therefore, one of the key motivations in the remainder of this chapter is to conduct a comparative performance analysis of different dynamic load distribution algorithms, highlighting the core advantages of the measurement-based approach adopted in the generalized CCRRM architecture.

7.4.1 Classification of Load Distribution Algorithms

The key issues in designing any load or admission control algorithms are: (i) identifying a suitable load metric to accurately estimate the available network capacity; and (ii) adopting a suitable decision trigger to effectively exploit and maximize the overall composite capacity. Traditionally, load control is concerned with load distribution to improve network QoS performance by transferring STAs from heavily to lightly loaded networks. This allows STAs to take advantage of the spare network capacity which would otherwise be left unused. However, it is also important to consider the state of wireless channel, which places fundamental limits on the network QoS performance, when distributing load across wireless networks as explained in Chapter 5. Admission control is also critical for the provisioning of QoS by regulating input traffic and preventing the overloading of network. It works by conducting an assessment to check whether a new flow could be admitted without compromising the QoS requirements of existing flows. Hence, admission control policy dictates the provisioning of either guaranteed or predictive QoS [176]. In fact, admission control and load control are often not dissociable. The main reason is that both rely on the knowledge of the load metric in order to make their decisions.

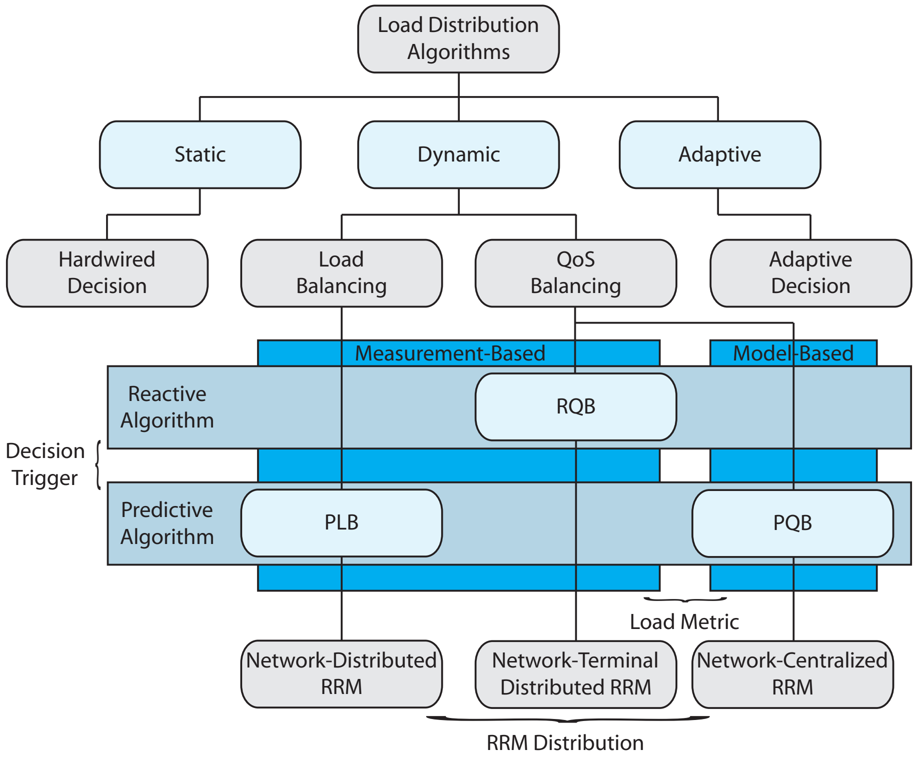


Figure 7.16: Classification of load distribution algorithms.

Henceforth, both load and admission controls are treated interchangeably in the context of the following discussions.

In essence, load distribution algorithms can be broadly classified as static, dynamic, and adaptive as illustrated in Figure 7.16. The main difference between static and dynamic load distribution algorithms is that the latter utilizes additional system state information, which enables the exploitation of short-term fluctuations, to improve the quality of its decisions. In contrast, the decision of static load distribution algorithm is typically hardwired, e.g., in a round-robin manner. On the other hand, adaptive load distribution algorithm is an extension of dynamic load distribution algorithm with the capability to adapt its parameters or policies dynamically in response to varying system states.

This thesis is concerned with dynamic load distribution algorithms which can be categorized as load balancing or QoS balancing algorithm. Both algorithms have the same primary function of avoiding under-utilized networks when distributing load. The subtle difference is the former attempts to equalize load while the latter attempts to equalize QoS across networks in order to improve perceived QoS for all flows. Based on the acquisition of load metric, QoS balancing algorithm can be further classified into model-based or measurement-based approach, whereas load balancing algorithm belongs only to the class of measurement-based approach.

In the model-based approach, the load metric is obtained by analyzing the WLAN DCF using a two-dimensional Markov chain model either with or without the aid of theoretical queueing models. E.g., Zhai *et al.* [133] integrate Bianchi's model [131] with the $M/M/1/K$ and $M/G/1/K$ queueing models to give non-saturation throughput, delay, and loss bounds. Malone *et al.* [145] extend Bianchi's model to non-saturation conditions by incorporating post-backoff states under bufferless network assumption (cf. Chapter 6 for a comprehensive review of pertinent analytical models). In the measurement-based approach, the load metric is obtained by either direct measurements or estimations from the system itself. E.g., Velayos *et al.* utilize the throughput of AP to reflect the load of a network. Ong and Khan [102] employ the PD of AP to capture both network and wireless channel variations, which are indicative of the network load. Above all, the CU estimation (cf. Section 5.2.2) first proposed by Garg and Kappes [96] gives the best representation of the effective network load. In addition, QoS balancing algorithm can be categorized as reactive or predictive algorithm according to its decision trigger, whereas load balancing algorithm belongs only to the class of predictive algorithm. Reactive algorithm is defined as the load distribution process in which its decision trigger is based on the *observation* of some KPIs, whereas predictive algorithm is defined as the load distribution process in which its decision trigger is based on the *prediction* of future dynamics in some KPIs, both against a set of pre-defined thresholds.

It is worth mentioning that the types of load metric and decision trigger jointly determine the behavior of admission control. In particular, soft admission control (cf. Section 4.2.1) is defined as one which operates on a soft limit where its load metric is obtained by link layer measurements, and its corresponding decision trigger is based on the observation of some KPIs. On the contrary, an admission control is said to operate on a hard limit when its decision trigger is based on the prediction of future dynamics in some KPIs, irrespective of how it acquires its load metric. Hence, only the RQB algorithm satisfies the definition of soft admission control.

The level of centralization also plays a crucial role in dynamic load distribution algorithms. Balachandran *et al.* [100] present an adaptive load balancing solution where a centralized admission control server contains the load information of all APs. Velayos *et al.* [101] propose a decentralized load balancing scheme where the APs are classified based on their throughput in one of the three states, viz., underloaded, overloaded, or balanced. It is known that both centralized and decentralized architectures have their pros and cons [178]. Hence, the recently approved IEEE 1900.4 standard advocates a network-terminal distributed RRM framework, which can be seen as a compromise between the centralized and decentralized ones, such as the TONA handover architecture presented in Section 7.2 (cf. Figure 7.6).

7.5 Dynamic Load Distribution Algorithms

The comparison of the three dynamic load distribution algorithms first presented in [182] and [183] is summarized in Table 7.3. It is important to note that these dynamic load distribution algorithms have different RRM distributions [178]. This implies that their RRM functions which typically consist of network selection, load control, and admission control, together with their corresponding RRM decisions have different levels of centralization. Accordingly, the network-centralized RRM framework refers to RRM decisions which are made in a central APC. The network-distributed RRM framework refers to RRM decisions which are distributed between APs. Lastly, the network-terminal distributed RRM framework refers to RRM decisions which are distributed between APs and STAs via the APCs. Since these algorithms span across different levels of centralization,

Table 7.3: Comparison between the dynamic load distribution algorithms.

Attributes	Model-Based	Measurement-Based	
Algorithm Type	QoS balancing (PQB)	Load balancing (PLB)	QoS balancing (RQB)
Traffic Profiling	Mean arrival rates, collision probability, queue characteristics	Estimated peak and/or mean CU	Measured PD, estimated mean CU
Load Metric	PD, PLR (cf. Section 6.2)	CU [96]	PD [102], CU [96]
Decision Trigger	Predictive	Predictive	Reactive
Admission Control	Hard Limit	Hard Limit	Soft Limit
RRM Distribution	Network-centralized	Network-distributed	Network-terminal distributed
Information Exchange	Between APC-APs	Between APs	Between APCs-APs-STAs
Utilization	Medium	Low	High
Handover Events and Complexity	High	Low	Medium
QoS Provision	Predictive QoS		
Stability Period	10 QoS Broadcast Intervals		
Candidate Selection	QoS Satisfaction Factor ($QSF < 1$)		
Network Selection	Greedy Approach		

their performance is investigated based on the IP-based TONA handover architecture (cf. Section 3.2 and Section 7.2) which can be configured to support different RRM distributions. In what follows, an overview of the three dynamic load distribution algorithms, which aim to redistribute load across a multi-AP WLAN by exploiting the heterogeneity of dynamic network conditions to trigger VHO, is given.

7.5.1 Predictive QoS Balancing Algorithm

In the PQB algorithm, the load metric is based on the QoS metrics of PD and PLR, which are derived by combining two analytical models as illustrated in Figure 7.17. Accordingly, the Markov chain model is used to analyze the WLAN DCF operation and the $M/M/1/K$ queueing model is used to analyze the WLAN QoS performance under varying traffic and wireless channel conditions. Here, Zhai's model [133] is modified to reflect the asymmetric load situation of an infrastructure BSS VoWLAN configured in a wireline-to-wireless topology. The VoWLAN consists of one AP, $N-1$ WLAN STAs, and $N-1$ ethernet STAs which are connected through a wireline backbone. When considering 2-way voice conversations between WLAN and ethernet STAs, the traffic load flowing through the AP is $N-1$ times that of a WLAN STA since the AP transmits half of the voice traffic to WLAN STAs. In addition, traffic variability between WLAN

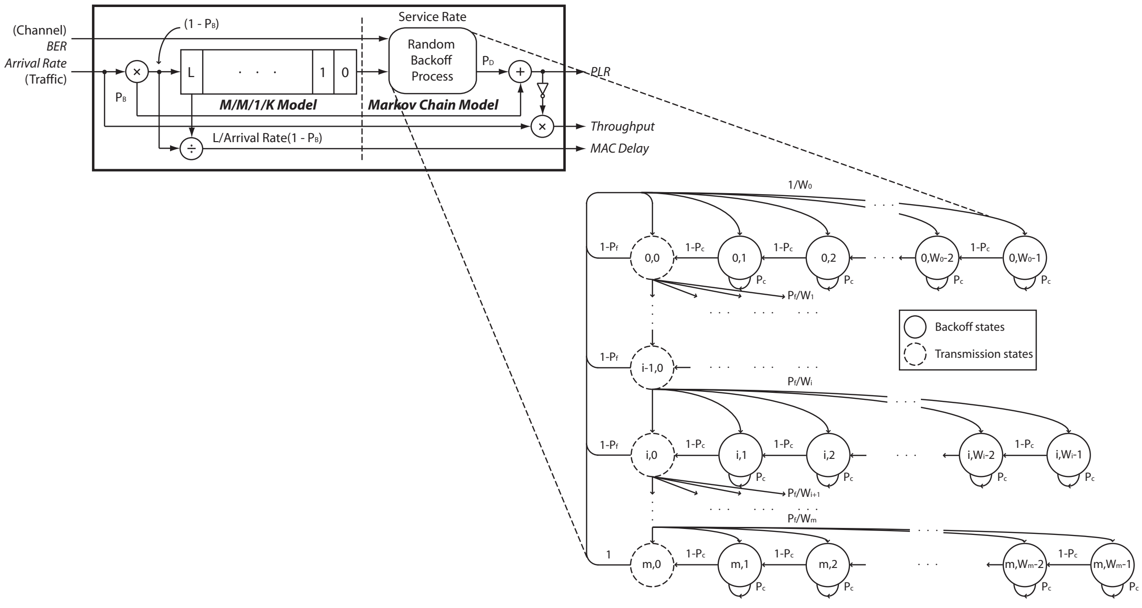


Figure 7.17: Implementation of the PQB algorithm.

STAs is introduced by considering heterogeneous voice codecs of different packetization intervals and packet lengths. Furthermore, wireless channel variability between BSSs is considered by factoring in transmission failures in both MAC data frame and ACK frame. An AWGN wireless channel is assumed where each bit has the same bit error probability and bit errors are i.i.d. over the entire frame. The effects of distance are ignored, and hence it is assumed that all STAs have the same BER and FER as in Ni's model [134]. The freezing of backoff counter during the times when medium is busy is also modeled according to Ziouva's model [135].

Collectively, the unified analytical model (cf. Section 6.2 for details) accounts for: (i) asymmetric traffic load between the AP and its associated STAs of an infrastructure BSS VoWLAN; (ii) diverse traffic flows between STAs; (iii) transition from the non-saturation to saturation mode (and vice-versa); and (iv) diverse wireless channel conditions between BSSs of a multi-AP hotspot scenario. The PQB algorithm is implemented as the network-centralized RRM framework where the load metric is used as the upper bounds of admissible traffic load, which include the new flow and any existing flows of an AP, in a centralized admission control to provision predictive QoS. It is worth to remark that these bounds are more proper as compared to that used in the PLB algorithm since the collision probability and queue characteristics of the AP are considered. However, the PQB algorithm will generally result in higher complexity.

7.5.2 Predictive Load Balancing Algorithm

In the PLB algorithm, the load metric is based on the CU which estimates the fraction of channel occupation time per observation interval. The CU is widely used as the load metric for both load and admission control algorithms due to its simplicity and accuracy. Accordingly, the CU of each flow and the corresponding network capacity can be estimated as

$$CU_{total}^{n} = \sum_{k \in Flows} CU_{k}^{n}, \quad n \in APs, \tag{7.1a}$$

$$CU_{j}^{n} + CU_{total}^{n} < CU_{max}, \tag{7.1b}$$

where $0 \leq CU_{total}^{n} \leq 1$ is the total CU of nth AP, CU_{j}^{n} is the CU of jth flow (cf. Section 5.2.2 for details on computing the CU of each flow), and CU_{max} is the admission threshold. A new RT flow can be accepted without affecting the QoS of existing flows if (7.1b) is true. For error-prone channel, it is necessary to consider the average FER and account for the factor of $(1 - FER)$, when computing the CU of each flow, i.e., $CU_{j}^{n}/(1 - FER)$, since the entire transmission will fail.

The PLB algorithm is implemented as the network-distributed RRM framework which is reported in [101] and depicted in Figure 7.18. Here, the APs are classified in one of the three states, viz., underloaded which will accept any requests, balanced which will accept only new connections, or overloaded which will not accept any requests but nominate a candidate STA for handover to an underloaded AP instead. The key characteristic of the PLB algorithm is that it attempts to equalize load across APs proactively. Guaranteed QoS can be provisioned when both peak and mean CU are used as the upper bounds of admissi-

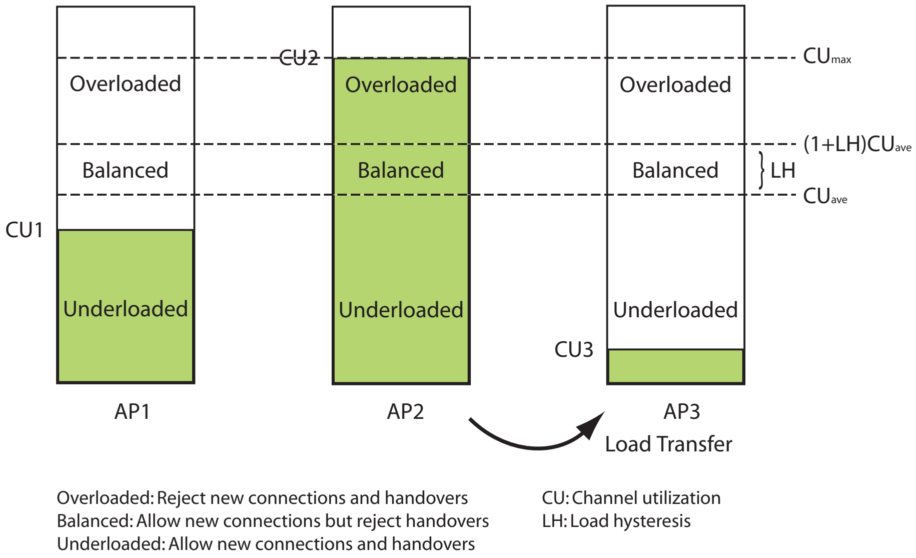


Figure 7.18: Implementation of the PLB algorithm.

ble traffic load. The network utilization is usually acceptable when flows are smooth with CBR sources. However, when flows are bursty with VBR sources, such guaranteed QoS inevitably results in low network utilization. Higher network utilization can be achieved by relaxing the bounds to use only the mean CU, but this implies that only predictive QoS can be provisioned. Furthermore, the admission threshold CU_{max} for RT flows is typically restricted to 80 – 90%. It is often argued that this buffer caters for the variability of VBR sources and ensures that NRT flows can be accommodated within the buffered capacity. However, in reality, finding an optimal admission threshold in not trivial since the saturation point of WLAN is non-unique as explained in Section 7.4 (cf. Figure 7.15). In other words, there will be a different impact on the network load *even* for the same average data rate. Hence, a better approach might be to remove the admission threshold and rely on the measurements of existing flows to regulate input flows. However, such measurements should be conservative by using the historical knowledge of fluctuations in network conditions.

7.5.3 Reactive QoS Balancing Algorithm

In the RQB algorithm, the load metric is based on the measured PD and mean CU, which are utilized as the upper bounds of admissible traffic load as shown in Figure 7.19. The RQB algorithm leverages on the link layer measurements of PD as QoS metric to characterize the perceived quality of each AP. The key advantages of adopting link layer measurements are that: (i) it could be used to quantify traffic variations explicitly and wireless channel variations implicitly since QoS metric, in general, varies accordingly to wireless channel conditions; and (ii) it mitigates the difficulty of estimating the actual bandwidth occupancy for each flow, particularly, in the presence of dynamic traffic patterns and wireless channel conditions when employed as the load metric for soft admission control. As explained in Section 7.4.1, soft admission control differs from the traditional hard admission control, which is typically used for homogeneous voice traffic, where the number of admissible connections can be easily pre-determined. These bounds are more relaxed as compared to the previous two algorithms, and thus are referred as soft limits.

Here, the mean CU is used *without* imposing an admission threshold to RT flows by setting $CU_{max} = 1.0$. This essentially removes the hard limit and encourages higher network utilization. E.g., Figure 7.19 illustrates that load transfers to AP 1 and AP 3 could be possible to exploit the buffered capacity on the conditions that the admission threshold for RT flows is removed and better QoS in AP 1 and AP 3 are perceived. However, additional PD measurements need to be incorporated to account for the past variations of network conditions. Accordingly, the measurements directly optimize the expected PD, making it adaptive to dynamic network conditions. This improves the flexibility of the admission control but at the expense of occasional violations, which limit it to provision predictive QoS, and moderate complexity. The network utilization gain would become more significant when there is a high degree of statistical multiplexing, e.g., in broadband WLANs.

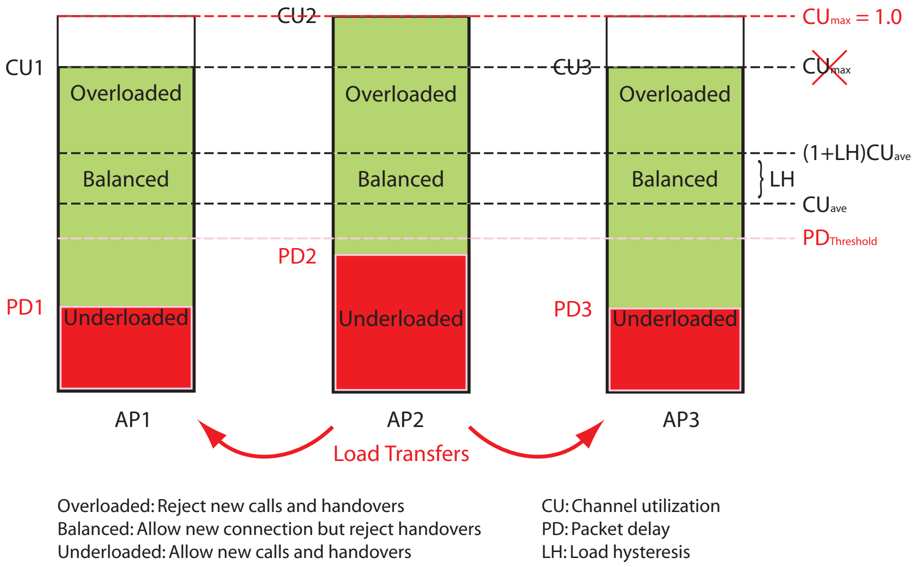


Figure 7.19: Implementation of the RQB algorithm.

The RQB algorithm is implemented as the network-terminal distributed RRM based on the IEEE 1900.4 standard as discussed in Section 7.2. In this study, however, the RQB algorithm invokes only bi-domain cooperation which essentially reduces to the iLB scheme found in Section 4.2. Additionally, the RQB algorithm will provide an important property for CWNs:

Baseline QoS. The long-term QoS performance of a CWN when subjected to load distribution using the measurement-based approach is similar to the QoS performance that the CWN could achieve when subjected to load distribution using the model-based approach.

In other words, the QoS performance of the CWN employing the RQB algorithm is similar to what could be achieved if the PQB algorithm is deployed. It is important to note that this baseline QoS property is unique to the RQB algorithm and does not apply to the PLB algorithm, albeit, it also belongs to the class of measurement-based approach. This baseline QoS property will be validated in Section 7.6.3.

7.5.4 Candidate Selection and Network Selection

To facilitate candidate selection, the QoS performance of STA is quantified as a function of two QoS metrics in which the QSF is defined in (A-7a) of Appendix A-3.1. Accordingly, $QSF < 1$ when the QoS requirements of STAs cannot be met and this condition is used by STAs in all the three dynamic load distribution algorithms to trigger VHO.

On the other hand, the network selection of all the three dynamic load distribution algorithms is based on the greedy approach. The reason being obtaining an optimal allocation of STAs to the available APs such that the allocation maximizes the overall composite capacity is a combinatorial problem which is NP-hard. For PQB (PLB) algorithm, the AP which maximizes the difference between the estimated bounds and predefined QoS metric (load metric) thresholds is selected. For RQB algorithm, network selection is implemented according to Section 3.4 where an AP with the highest network quality probability, which is based on PD measurements, is selected. A Bayesian learning process is used to capture the historical variations of network conditions conservatively, making it reliable for use in soft admission control.

7.6 Comparative Performance Evaluation

In order to compare the performance of the three different dynamic load distribution algorithms and ascertain the effectiveness of the RQB algorithm, which is based on the network-terminal distributed RRM framework advocated by the recent IEEE 1900.4 standard, two simulation scenarios are examined. The simulation models are developed by using OPNET™ Modeler® 14.5 with Wireless Module. Modifications to the existing DCF models are performed to incorporate the three different dynamic load distribution algorithms as described in Section 7.5 which are the focus of this study. A typical hotspot which consists of a multi-AP VoWLAN with three IEEE 802.11b APs operating at the data rate of 1 Mbps is simulated according to Figure 7.20. The VoIP traffic is generated using the heterogeneous voice codecs of different packetization intervals and packet

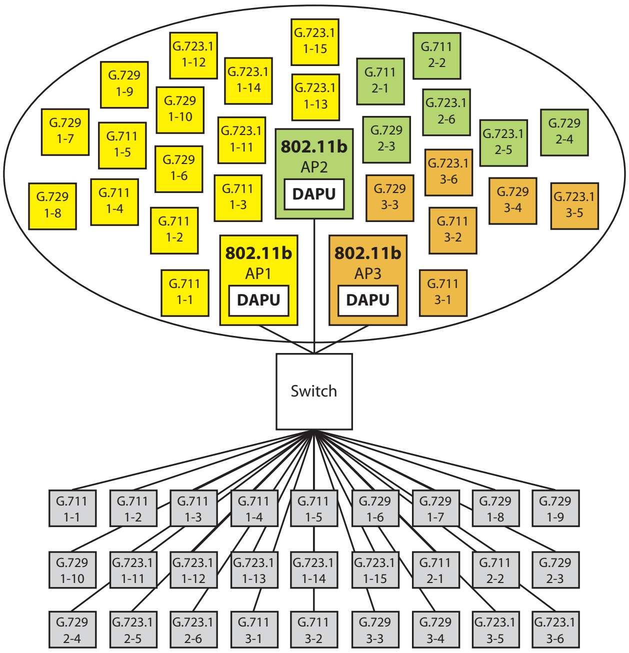


Figure 7.20: Simulation model of a homogeneous multi-AP WLAN with the IEEE 802.11b APs under diverse network traffic and wireless channel conditions.

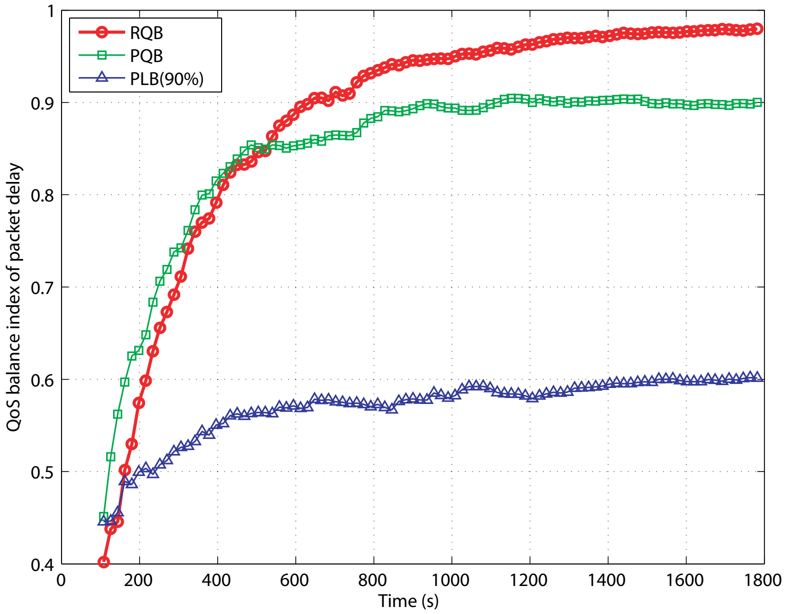
lengths as shown in Table 7.4, and the VBR source is simulated using ON-OFF model. These traffic generation parameters are consistent to the VoWLAN study in Section 4.3.1.

In this simulation, an unbalanced load of five G.711, five G.729, and five G.723.1 STAs in BSS 1 while two G.711, two G.729, and two G.723.1 STAs in each of BSS 2 and BSS 3 is initially introduced. An ideal channel is considered in the first scenario while an error-prone AWGN wireless channel is simulated in the second scenario where the BER of wireless channels in BSS 1, BSS 2, and BSS 3 are 10^{-9}, 10^{-5}, and 10^{-6}, respectively. The motivation is to examine the worst case scenario when the total offered load approaches the overall composite capacity of the three BSSs under diverse network traffic and wireless channel conditions. Furthermore, although ideal channel, i.e., BER = 0 is rarely achievable in real world scenarios, it serves to illustrate the performance of these dynamic load distribution algorithms without the influence of channel errors. Additional details on the general simulation models are available from Appendix A-2.

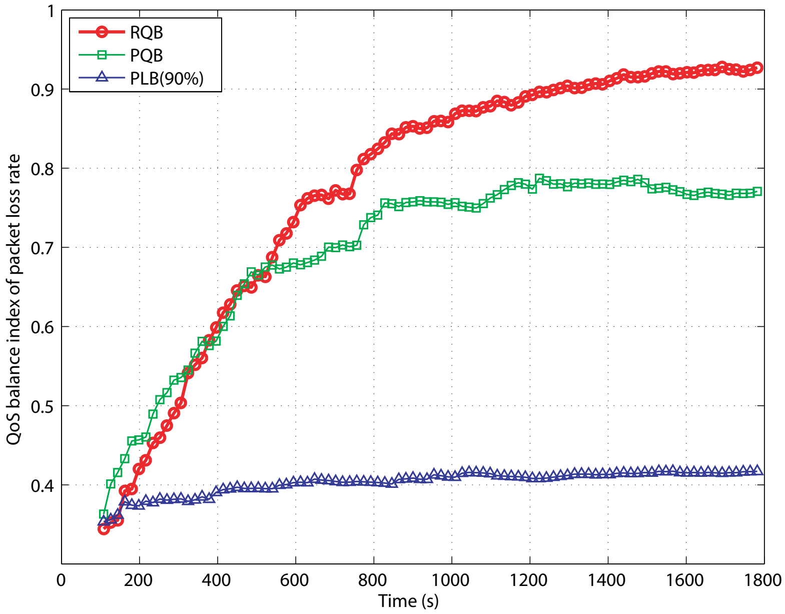
Table 7.4: Traffic generation parameters.

Traffic Type	Packet Size (Bytes)	Inter-arrival (ms)	Avg. Data Rate (Kbps)
G.711 (VBR)	80	10	64
G.729 (VBR)	20	20	8
G.723.1 (VBR)	24	30	6.4

For the QoS performance evaluation, the QBI defined in (A-9) of Appendix A-3.1 is adopted to quantify the effect of different dynamic load distribution algorithms on the QoS fairness among APs. The considered QoS metrics are PD and PLR, which are typically used to characterize the quality of VoIP traffic. On the other hand, the η_{cc} defined in (A-10) of Appendix A-3.1 is employed to evaluate the overall composite capacity achievable by different dynamic load distribution algorithms. For the case of ideal channel conditions, the PLB algorithm is evaluated with an admission threshold of $CU_{max} = 0.9$ denoted as PLB(90%). For the case of error-prone channel conditions, the PLB algorithm is evaluated with the admission threshold of $CU_{max} = 0.8$ denoted as PLB(80%), in addition to the PLB(90%). The PQB algorithm is evaluated with a PD threshold of 60 ms and PLR threshold of 1%. The RQB algorithm is evaluated with $CU_{max} = 1.0$, i.e., no admission threshold for RT flows and also a PD threshold of 60 ms. Where appropriate, these three dynamic load distribution algorithms are also compared to the initial case of an unbalanced load with no load distribution denoted as NLD, and the case of a balanced load in which all three APs have the same number of associated STAs denoted as BAL. The key motivation is to compare the QoS fairness between APs, overall composite capacity, number of handover events, and end-user throughput when different dynamic load distribution algorithms are deployed. Additionally, an interesting relationship, which is the salient trait of QoS balancing algorithm, between the aggregate QSF of STAs defined in (A-7a) and aggregate throughput of STAs, as well as the QoS fairness between APs defined in (A-9) is unveiled.

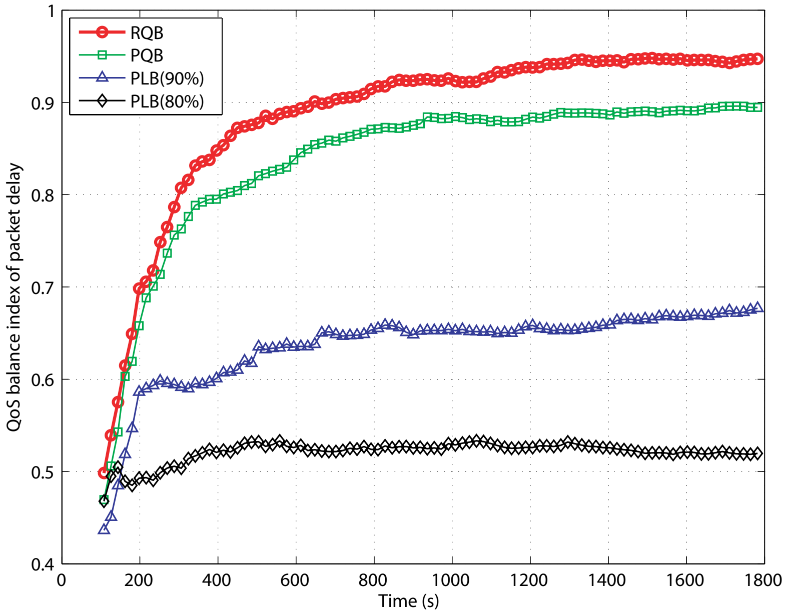


(a) Average QBI of PD.

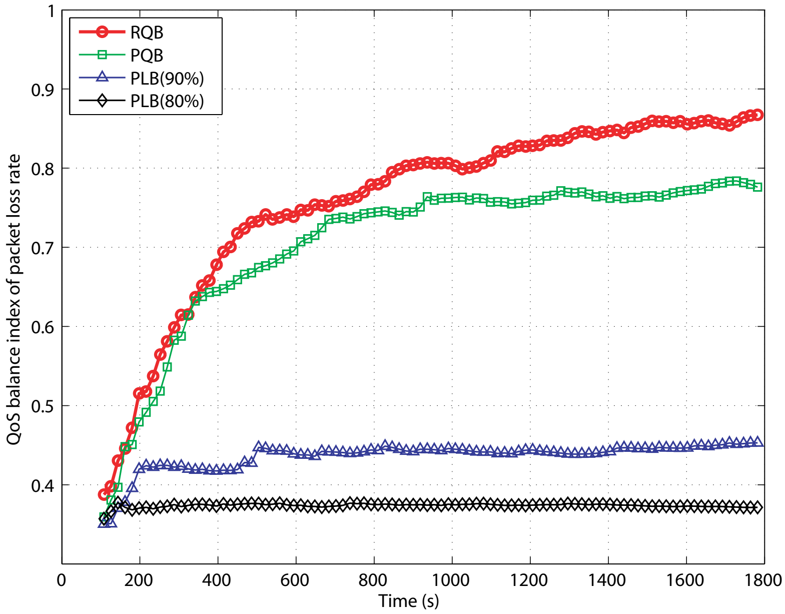


(b) Average QBI of PLR.

Figure 7.21: Average QBI of PD and PLR between APs under ideal channel conditions.



(a) Average QBI of PD.



(b) Average QBI of PLR.

Figure 7.22: Average QBI of PD and PLR between APs under error-prone channel conditions where the BER of wireless channels in BSS 1, BSS 2, and BSS 3 are 10^{-9}, 10^{-5}, and 10^{-6}, respectively.

7.6.1 QoS and Composite Capacity Performances

The presented results are analyzed starting from 100 s (0 – 100 s is the warm-up period). Ideally, according to the definition of (A-9) – (A-10), QBI and η_{cc} should be close to 1 so as to offer QoS fairness and maximize overall composite capacity, respectively. First, Figure 7.21(a) and Figure 7.21(b) illustrate that RQB outperforms both PQB and PLB by 4% (11%) and 54% (95%) in terms of the QBI of PD (PLR) between APs, respectively under ideal channel conditions. On the other hand, Figure 7.22(a) shows that RQB outperforms PQB by 5%, PLB(90%) by 39%, and PLB(80%) by 67% in terms of the QBI of PD between APs under error-prone channel conditions. Similarly, Figure 7.22(b) depicts that RQB outperforms PQB by 7%, PLB(90%) by 72%, and PLB(80%) by 100% in terms of the QBI of PLR between APs under error-prone channel conditions. Two important observations are made on the QoS performance of the different dynamic load distribution algorithms. First, it is evident that all the three algorithms are able to achieve higher QBI of both PD and PLR under ideal channel conditions, which is expected, and they exhibit similar trends when compared to the case under error-prone channel conditions. Second, it is clear that the state of balance, i.e., the QoS fairness between APs is dependent on the type of dynamic load distribution algorithms, which would now be discussed. Since all the three algorithms have similar trends under both ideal and error-prone wireless channel conditions, the following discussions will be based only on error-prone channel conditions.

Next, Figure. 7.23 illustrates that RQB yields an overall composite capacity improvement over PQB by 0.5%, PLB(90%) by 6%, PLB(80%) by 14%, and NLD by 26%. In fact, NLD and BAL essentially form the capacity lower and upper bounds, respectively. Specifically, NLD represents the initial case without load distribution where BSS 1 is overloaded. In constrast, BAL reflects the case of the best possible load distribution, which results in three G.711, three G.729, and three G.723.1 STAs in each BSS, with the simulated scenario. It is evident that both QoS balancing algorithms converge to the capacity upper bound of 86% with an average η_{cc} of 84%. On the other hand, the

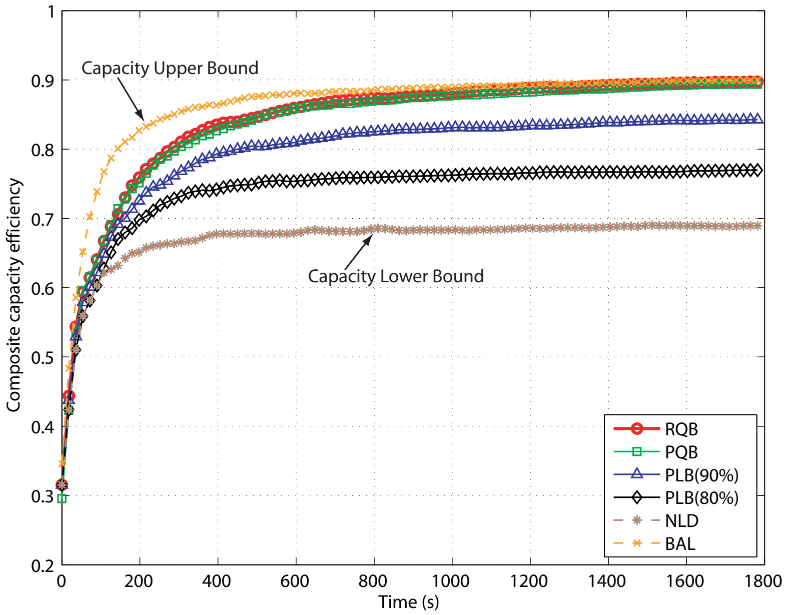
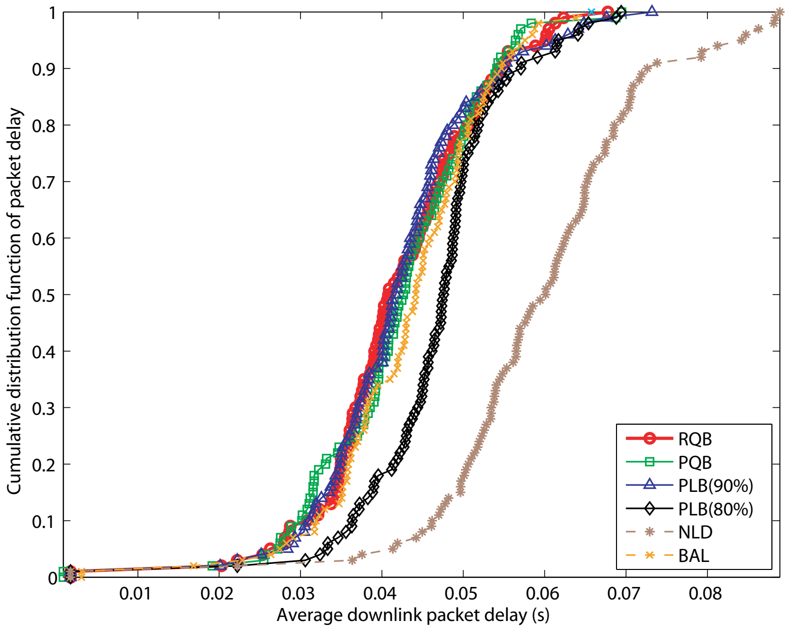


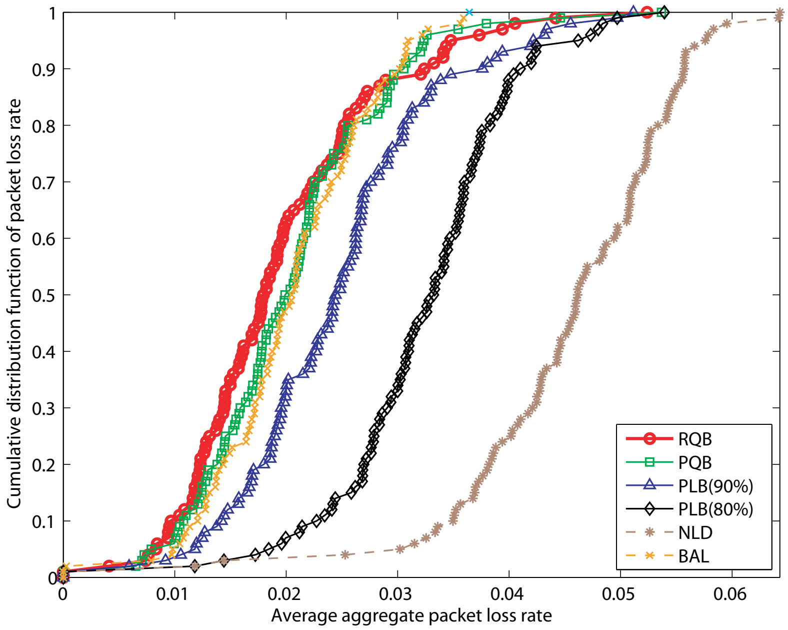
Figure 7.23: Composite capacity efficiency of different dynamic load distribution algorithms in a multi-AP WLAN under error-prone channel conditions where the BER of wireless channels in BSS 1, BSS 2, and BSS 3 are 10^{-9}, 10^{-5}, and 10^{-6}, respectively.

load balancing algorithm can achieve only an average η_{cc} of 79% and 73% for admission thresholds of $CU_{max} = 0.9$ and $CU_{max} = 0.8$, respectively.

On the whole, both QoS balancing algorithms achieve better QoS fairness as compared to the load balancing algorithm. The QoS balancing algorithms exhibit better performance for two main reasons. First, the load metrics of both PQB and RQB contain at least one of the QoS metrics under study. This directly optimizes the expected PD and PLR while the load metric of PLB is indirectly related to the investigated QoS metrics. Second, the load metric of PLB is based on the mean CU where the admission threshold is set to 80% (90%) of an AP's maximum capacity. Since only BSS 1 is overloaded in the simulated scenario, the admission threshold preemptively creates an aggregate buffer capacity of 40% (20%) in BSS 2 and BSS 3. This places a hard limit which prevents the opportunistic exploitation of possible spare capacity. Although this strategy attempts to protect existing flows, it inevitably results in higher blocking probability for incoming handover



(a) CDF of average DL PD.



(b) CDF of average aggregate PLR.

Figure 7.24: CDF of average DL PD and aggregate PLR in a multi-AP WLAN under error-prone channel conditions where the BER of wireless channels in BSS 1, BSS 2, and BSS 3 are 10^{-9}, 10^{-5}, and 10^{-6}, respectively.

attempts. Hence, BSS 1 suffers sustained overloading which degrades the QoS fairness between APs and overall composite capacity. This impact will be magnified with decreasing admission threshold, which acts to create more buffered capacity, and this is evident from Figure 7.22 and Figure 7.23, respectively. Moreover, choosing an optimal admission threshold is not trivial since the saturation point of WLAN is non-unique as explained in Section 7.4 (cf. Figure 7.15). Therefore, it is very difficult to obtain an accurate characterization of RT flows as a priori knowledge in the presence of such dynamic network conditions. It is worth noting that PQB also utilizes hard limit but admission threshold is not required. Hence, QoS fairness of PQB comes in between RQB, as well as PLB(80%) and PLB(90%).

On the other hand, RQB also employs the mean CU as one of its load metric but relaxes the bounds by eliminating the admission threshold. Instead, it operates on a soft admission control using PD measurements. The salient advantage of measurement-based soft admission control is that it relies on the historic variations of network conditions captured through measurements to mitigate the difficulty in characterizing the bandwidth occupancy of RT flows. Hence, a higher network utilization can be achieved by allowing the exploitation of spare capacity opportunistically. This is evident in the case of RQB over PLB(80%) and PLB(90%) as shown in Figure 7.23 where both are designed to provision predictive QoS. Although there would be sporadic violations of PD as shown in Figure 7.24(a), this would be outweighed by the remarkable QoS fairness and overall composite capacity improvements as shown in Figure 7.22 and Figure 7.23, respectively. Note that these improvements are the direct consequences of the PLR improvements as shown in Figure 7.24(b).

7.6.2 Handover Performance

In terms of the handover performance as shown in Figure 7.25, PLB(80%) has the least number of handover events as compared to PLB(90%), RQB, and PQB under error-prone channel conditions. On the other hand, Figure 7.26 illustrates that, under ideal channel conditions, PLB(90%) has the least number of handover events as compared to both RQB and PQB where PQB has an additional handover event as compared to RQB. Note that both NLD and BAL do not generate any handover events, and thus are not depicted. In both cases, when comparing between the two QoS balancing algorithms, PQB has the highest number of handover events while RQB has moderate number of handover events which comes in between PLB and PQB. It is also worth mentioning that all the three algorithms generate fewer number of handover events under ideal channel conditions. This is due to the fact that the candidate selection in all the three algorithms is triggered by the QSF of STA as explained in Section 7.5.4. Hence, all the three algorithms will converge to the state of balance sooner in the absence of channel errors since the QoS requirements of STAs will be affected only by network congestions. In general, both QoS balancing algorithms tend to accrue more handover events as compared to the load balancing algorithm since their load metrics do not impose any admission thresholds to create buffered capacity preemptively. However, the QoS balancing algorithms provide better overall QoS performance in terms of PD and PLR as compared to the load balancing algorithm since their load metrics contain at least one of the QoS metrics under study.

7.6.3 Throughput Performance

From Figure 7.27, it is interesting to observe that both QoS balancing algorithms have lower aggregate QSF but higher aggregate throughput from the STA perspective as compared to the load balancing algorithm. More specifically, the aggregate throughput increases with decreasing QSF. Similarly, from Figure 7.22(a) and Figure 7.22(b), QoS fairness also increases with decreasing aggregate QSF. This suggests that tradeoffs exist between the aggregate QSF and throughput of STAs, as well as the QoS fairness between APs. To be more specific, for every decrease in the aggregate QSF of STAs, there is a corresponding increase in the aggregate throughput of STAs and QoS fairness between APs. In other words, the QoS balancing algorithms *trade* the aggregate QSF of STAs

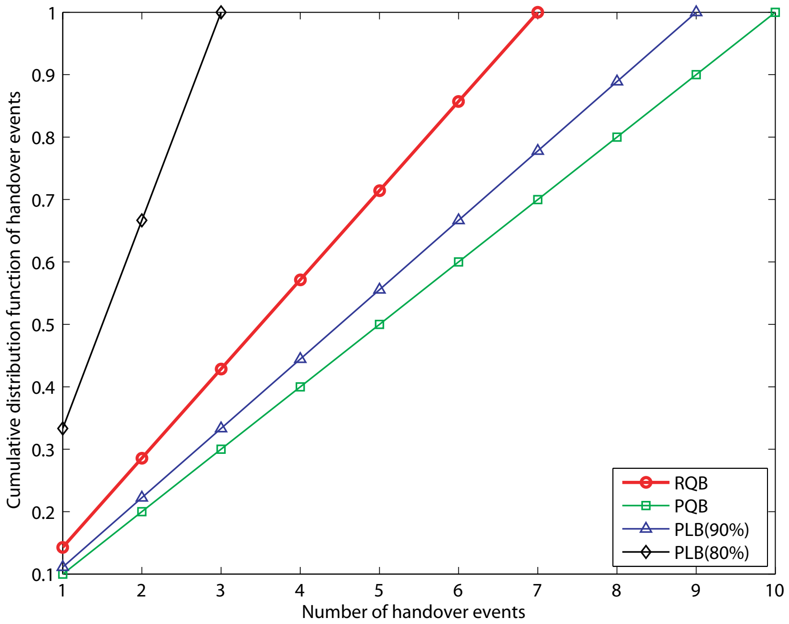


Figure 7.25: CDF of handover events in a multi-AP WLAN under error-prone channel conditions where the BER of wireless channels in BSS 1, BSS 2, and BSS 3 are 10^{-9}, 10^{-5}, and 10^{-6}, respectively.

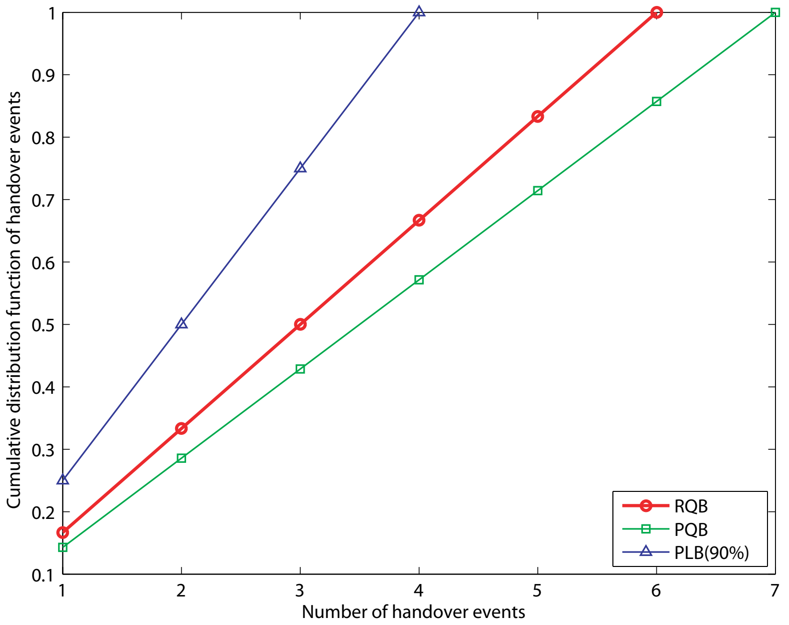


Figure 7.26: CDF of handover events in a multi-AP WLAN under ideal channel conditions.

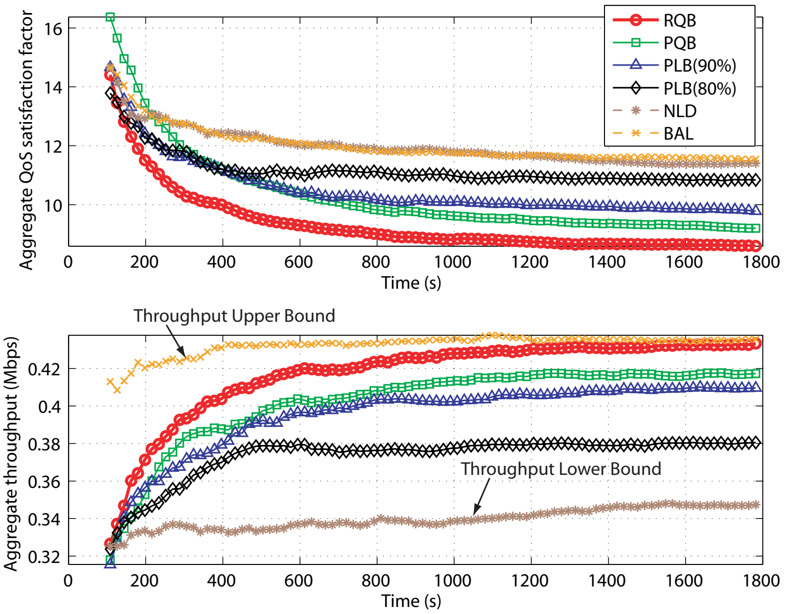


Figure 7.27: Average aggregate QSF and throughput of STAs (end-user throughput) under error-prone channel conditions where the BER of wireless channels in BSS 1, BSS 2, and BSS 3 are 10^{-9}, 10^{-5}, and 10^{-6}, respectively.

for QoS fairness between APs in order to maintain a QoS-balanced system which in turn yields higher aggregate throughput of STAs. The only exception is found in BAL which is initially balanced. As a result, BSS 1 of BAL is not as overloaded relatively, which explains for its highest aggregate QSF and throughput of STAs. Notice that RQB, after load distribution, approaches the throughput upper bound of BAL. Particularly, RQB achieves 4% higher aggregate throughput of STAs than PQB owing to its improvement of QoS fairness which accentuates the importance of maintaining a QoS-balanced system.

When comparing between the two QoS balancing algorithms, it is clear that RQB is able to achieve higher QoS fairness between APs and aggregate throughput of STAs, in addition to generate fewer handover events as compared to PQB, but at the expense of lower aggregate QSF of STAs. Although RQB results in lower aggregate QSF, it is important to emphasize that both RQB and PQB have similar average DL PD and aggregate PLR as shown in Figure 7.24(a) and Figure 7.24(b), respectively from the composite system perspective. It is also observed that both RQB and PQB share similar QoS performance with

BAL, whereas NLD has the worst QoS performance as expected. Clearly, this validates that RQB preserves the *baseline QoS* property as defined in Section 7.5.3. This baseline QoS property is highly desirable and it reiterates the advantage of using the measurement-based soft admission control. Specifically, soft admission control improves its flexibility in the presence of dynamic network conditions by exploiting spare capacity in an opportunistic manner while allowing occasional violations. Note that this favorable property is not found in PLB, albeit, it also belongs to the class of measurement-based approach.

7.6.4 Discussions

It is observed that the performance of all the three dynamic load distribution algorithms tend to achieve higher network utilization, QoS fairness, and generate fewer handover events under ideal channel conditions, which comes as no surprise. Moreover, the performance of all the three dynamic load distribution algorithms, which largely depends on their load metrics and decision triggers, has various tradeoffs. The load balancing algorithm which uses the mean CU as the load metric has the advantages of lower complexity and fewer handover events. However, it results in lower network utilization due to the required admission threshold for RT flows, which creates buffered capacity that may not be efficiently utilized. Furthermore, how to choose an admission threshold for RT flows optimally or adaptively is non-trivial as it is very difficult to obtain an accurate characterization of RT flows as a priori knowledge in the presence of dynamic network conditions. On the other hand, the QoS balancing algorithms which utilize QoS metrics as the load metric have the advantages of higher network utilization, QoS fairness, and aggregate throughput of STAs since their load metrics directly optimize the expected PD and PLR of the system. However, they tend to be more complex, generate more handover events, and result in lower aggregate QSF of STAs which, in fact, is a favorable tradeoff for achieving higher network utilization.

Between the two QoS balancing algorithms, it is important to note that the measurement-based soft admission control employed in RQB has evident advantages over the hard admission control found in PQB. Particularly, RQB yields higher QoS fairness and aggregate throughput of STAs with fewer handover events while preserving the baseline QoS property and maintaining similar overall composite capacity. These are attributed to the Bayesian learning process which reliably captures the historical variations of network conditions for use in the soft admission control to exploit any available capacity opportunistically and adapt to dynamic network conditions. Another significant advantage of RQB is that it is based on the technology agnostic approach as explained in Section 3.3.2. In particular, it provides a generic measurement-based approach which can be deployed in any wireless networks since it requires only link layer measurements of QoS metric to quantify traffic variations explicitly and wireless channel variations implicitly. On the contrary, PQB employs a model-based approach where generalization for different wireless networks is challenging and generally requires remodeling efforts. It is important to highlight that the recent IEEE 1900.4 standard discussed in Section 7.1, which has recently gained much attention, is an example of such measurement-based system.

7.7 Chapter Summary

This chapter has illustrated the intricate relationships between the TONA handover architecture and the recent IEEE 1900.4 system architecture, which are similar in many aspects. Additionally, the relevance of the LAS framework, developed based on the concept of RQB, to the IEEE 1900.4 functional architecture has been exemplified from the LAP. The LAP has been implemented based on the distributed radio resource usage optimization use case, and it is easily extensible to the dynamic spectrum assignment and dynamic spectrum sharing use cases. The simulation results have shown that the IEEE 1900.4 RRM which is based on the LAP can effectively exploit the cooperative exchange of QoS context information, as part of the CLM, between network-terminal entities to facilitate better utilization of radio resources. Such QoS awareness is supported by the generalized CCRRM architecture for coordinating informed VHO and dynamic load distribution to harness overall composite capacity and QoS improvements. On the other hand, the network-distributed RRM based on the LBM, which performs load balancing based on a single load metric such as the CU, is QoS-agnostic. Furthermore, it is incapacitated by wireless channel variations which could arise due to NLOS transmissions in legacy WLAN and/or dynamic spectrum assignment/sharing supported by reconfigurable RANs and terminals. As a consequence, the IEEE 1900.4 RRM has outperformed the network-distributed RRM in terms of the aggregate throughput of system by 15%, QoS balance by 23%, throughput fairness by 48%, and handover reduction by 72%. Interestingly, these results are consistent to the studies in [25] (see, e.g., §3.2.4 – 3.2.5) which have also shown that the IEEE 1900.4 RRM yields better performances for network capacity and user perceived QoS.

This chapter has also presented the comparative performance analysis between the three dynamic load distribution algorithms, viz., PLB, PQB, and RQB. The simulation results have indicated that the performance of all the three algorithms largely depends on their load metrics and decision triggers. In addition, they generally achieve better performance under ideal channel conditions. In particular, RQB achieves higher (significantly higher) QoS fairness, similar (higher) overall composite capacity, higher (much higher) end-user throughput, similar (much better) QoS performance, and moderate number of handover events as compared to PQB (PLB evaluated at both admission thresholds of 80% and 90%). Clearly, the generalized measurement-based approach employed in RQB, which reflects the recent IEEE 1900.4 RRM, is adaptive to dynamic network conditions which could arise from both traffic and wireless channel variations. As a result, RQB yields the best improvement in QoS fairness and aggregate end-user throughput while preserving its baseline QoS property. It is important to note that such attractive features are the remarkable result of maintaining a QoS-balanced system. Moreover, the baseline QoS property is not found in PLB, albeit, it is also based on the measurement-based approach.

This chapter has further underpinned that RQB has the highly desirable intrinsic property of preserving baseline QoS, in addition to the other favorable intrinsic properties of: (i) providing statistical QoS guarantee for multimedia traffic; (ii) maximizing overall composite capacity by maintaining QoS and throughput fairness; (iii) precluding unnecessary handovers; and (iv) mitigating the rate anomaly problem, which jointly achieve the end-to-end goal of promoting a QoS-balanced system. Thus, this thesis has corroborated that QoS balance should be employed as the criterion to quantify the state of balance in future wireless networks from the terminal perspective based on perceived QoS instead of from the network perspective based on absolute network load. Such end-to-end goal of effectuating a QoS-balanced system can be fulfilled by adopting the generalized CCRRM architecture embodied in this thesis.

CHAPTER 8

CONCLUSIONS

Since the release of the ITU-R M.1645 recommendation [1] in June 2003, *heterogeneity* and *convergence* have become the buzz words associated with future wireless networks. This has created a brand new problem space that has attracted vested interests from various research communities to pursue *seamless mobility* in future wireless networks, envisaged as an IP-based multi-RAT environment, leading to the recent advances in cooperative networks and cognitive networks. However, despite the significant progress in the cooperative and cognitive networks, the possibilities and benefits of their *cross-fertilization* have not received equal attention. Hence, vast research opportunities still remain to be explored. This thesis has aimed to provide a concrete framework to investigate the relationships between cooperation and cognition from a user-centric perspective in the design of advanced RRM solutions for future wireless networks. In particular, this thesis has leveraged on the *CoRe methodology* to provide guidelines in identifying possible synergetic interactions and cross-fertilization opportunities in the development of a generalized CCRRM architecture. The generalized CCRRM architecture has amalgamated the benefits of IP convergence together with the advances in cooperative and cognitive networks into a unifying whole to deliver the end-to-end goal of promoting a *QoS-balanced system*. This chapter is devoted to a summary of the main contributions reported in this thesis, which include the development and implementation of the TONA handover architecture, DANS algorithm, RQB algorithms, and unified analytical model, as well as the associated key results, followed by a discussion on prospective research directions for future work.

8.1 Summary of Contributions

One of the significant contributions of this thesis is the development of the CoRe methodology, which is poised as a concrete framework for lateral and in-depth investigations into the relation of cognition within a cooperative environment, to provide guidelines for innovative solutions toward the formulation of the generalized CCRRM architecture.

This thesis has defined, developed, implemented, and analyzed a novel unifying generalized CCRRM architecture for future wireless networks, which is anchored on the *technology agnostic approach*. One of the two key enablers of this technology agnostic approach is attributed to the proposed TONA handover architecture that has enabled *inter-network cooperation* by leveraging on the all-IP core network convergence to facilitate the cooperative exchange of QoS context information and dissemination of RRM policy. Additionally, the TONA handover architecture supports the notions of *network-assisted discovery* and *terminal-oriented decision* which have further stimulated *inter-entity cooperation* to support distributed decision making process between network-terminal entities. The TONA handover architecture has the principal benefits of supporting fast handover and power conservation to encourage 'green' terminals. These features will provide a firm baseline for incorporating future multi-mode, SDR-based devices as it is improbable to perform service discovery by exhaustive scanning procedures and operate multiple air interfaces at the same instant due to handover latency and battery lifetime constraints, respectively. An efficient implementation of the QoS broadcast mechanism in the TONA handover architecture to ensure interoperability with the existing standard has been presented. In addition, an evaluation of the system cost associated with the generalized CCRRM architecture, and the tradeoffs between QoS performance and QoS broadcast intervals have demonstrated that the QoS broadcast required in the generalized CCRRM architecture will not impose heavy load on the network, which further accentuate its benefits. Moreover, the exclusion of scanning operations in the fast handover design will lead to significant power conservation in the terminal despite its additional role in performing network selection.

The other key enabler of the technology agnostic approach is credited to an optimal measurement-based network selection technique which has been developed to coordinate VHO, based on dynamic QoS parameters, in a multi-RAT environment. The formulation of the dual-stage *QoS parameters estimation* process is a *cornerstone* of the technology agnostic approach which has provided a generic way to characterize the quality of wireless network and its channel, and provide link layer triggers. To be more specific, bootstrap approximation enables the estimation of dynamic QoS information from link layer measurements in a pragmatic way to provide *technology abstraction*. In addition, Bayesian learning facilitates as *link layer cognition* to filter unnecessary handover triggers which is a remarkable improvement over the shortcoming of existing cost function approach in terms of system stability, i.e., handover frequency, QoS performance, and system capacity. Moreover, the link layer cognition is augmented by the joint optimization of network-terminal distributed decision making process to provide link layer triggers for informed VHO and dynamic load distribution. Collectively, the link layer measurement and context awareness submodule and the QoS parameters estimation submodule have formed the technology abstraction and link cognition module which is the core of the generalized CCRRM architecture excogitated in this thesis. These key advancements in network selection have provided an avenue to take a step closer to ABC services and realize seamless mobility in future IP-based multi-RAT environment.

The novel concept of *RQB* has been coined to promote a long-term *QoS-balanced system* which is defined as the *end-to-end goal* of the generalized CCRRM architecture. A suite of RQB algorithms, augmented by different domains of cooperation, has been developed to exploit the heterogeneity of access networks and distribute load opportunistically. Specifically, the iLB scheme is based on bi-domain cooperation which has formed the *baseline design* of the generalized CCRRM architecture. The QLO framework is based on tri-domain cooperation which has featured an additional *intra-layer cooperation* between different RRM functional blocks to optimize load distribution in a single rate WLAN. The LAS framework is based on quad-domain cooperation which has introduced a supplementary *inter-layer cooperation* to establish synergetic interactions between link adaptation and load adaptation on-demand. The benefits of such synergetic exploitations have presented a novel way to treat the rate anomaly problem, which is inevitable in multirate WLAN-based cognitive networks, from a QoS perspective. Extensive simulation studies performed with comprehensive pragmatic scenarios have revealed several attractive intrinsic properties of RQB, viz., (i) providing statistical QoS guarantee; (ii) maximizing overall composite capacity by maintaining QoS and throughput fairness; (iii) precluding unnecessary handovers; (iv) mitigating the rate anomaly problem; and (v) preserving baseline QoS.

An elegant unified analytical model has been developed to obtain the key performance metrics of MAC delay, PLR, and throughput efficiency for the IEEE 802.11 DCF infrastructure BSS. The analytical model has integrated both Markov chain model and finite queueing model to capture *non-saturation* operating conditions. Additionally, it has considered *non-homogeneous* conditions by modeling the asymmetric traffic load between an AP and its associated STAs, heterogeneous flows between STAs, and heterogeneous wireless channel conditions between BSSs. These key performance metrics served as bounds for reliable capacity analysis from which a model-based PQB algorithm has been developed to benchmark the performance of the RQB algorithm. Extensive analyses and simulations have indicated that backoff freezing for an infrastructure BSS should be properly modeled in order to derive accurate performance metrics and consequently tight bounds for capacity analysis. Furthermore, the results have shown that ignoring backoff freezing for an infrastructure BSS will result in overly conservative bounds and eventually lead to low network utilization when deployed for admission control.

The intricate *similarities* between the proposed TONA handover architecture and the recent IEEE 1900.4 system architecture have been established. Moreover, the relevance of the LAS framework to the IEEE 1900.4 functional architecture has been exemplified through the implementation of the LAP based on the distributed radio resource usage optimization use case. Comprehensive simulation studies have corroborated that the IEEE

1900.4 RRM, based on the LAP, can effectively exploit the cooperative exchanges of context information between network-terminal entities to facilitate the orchestrated use of radio resources which in turn harness overall composite capacity and QoS improvements. As a consequence, the IEEE 1900.4 RRM has outperformed the network-distributed RRM in terms of the aggregate throughput of system by 15%, QoS balance by 23%, throughput fairness by 48%, and handover reduction by 72%. Remarkably, these results are consistent to the studies in [25] (see, e.g., §3.2.4 – 3.2.5) which have also demonstrated that the IEEE 1900.4 RRM yields better performances for network capacity and user perceived QoS. In order to benchmark the performance gains of the RQB algorithm, the comparative performance evaluation between the three dynamic load distribution algorithms has also been examined. Exhaustive simulations under diverse conditions have concluded that the RQB algorithm outperforms both the PLB and PQB algorithms to achieve higher QoS fairness and end-user throughput while preserving *baseline QoS* which is a desirable property not observed in the PLB algorithm. These studies are among the first which served as an early investigation effort to provide insights on the performance benefits of the baseline IEEE 1900.4 standard, and they will contribute toward the emerging IEEE 1900.4.a and IEEE 1900.4.1 standards. Finally, this thesis has corroborated that *QoS balance* should be employed as the criterion to quantify the state of balance in future wireless networks from the terminal or end-user perspective based on perceived QoS instead of from the network perspective based on absolute network load, particularly, when dynamic network conditions will be prevalent.

8.2 Future Research Directions

Given that this research is one of the first investigations devoted to the cross-fertilization of cooperative and cognitive principles within a generalized CCRRM architecture, a number of potential areas could be extended for future study as discussed in what follows.

8.2.1 Further Exploitation of Cooperation Domains

In Section 2.2.1, it is shown that several potential tasks could be associated with various stages of the COGNITION process. Particularly, the descriptions such as: (i) forming and leaving collaborative groups in the relate stage would enable interactions between peers and mentor; (ii) developing social relationships in the create stage could allow nomination and selection of mentor; and (iii) contribution of resources in the donate stage may result in the greater good of community, e.g., a terminal helping its peers to relay packet would benefit from lower channel access delay, which aim to solicit the cooperation between terminals have not been exploited in this thesis. On the other hand, this thesis implicitly assumes that inter-operation cooperation exists in Section 2.2. However, the level of inter-operator cooperation could be categorized as non-cooperative, i.e., competitive, semi-cooperative, or fully cooperative. Hence, it would be beneficial to study the impact on system performance with different degrees of cooperation between operators. In fact, the dynamic composition of roaming agreements between different networks operators or administrative domains is a new feature introduced in [9]. Other domains of cooperation such as the cooperation between the link layer and network layer to reduce the overall handover latency for seamless inter-technology mobility currently addressed by [20] and the cooperation between available resources of networks and terminals within a resource-trading framework suggested in [14] remain as interesting open research issues.

8.2.2 Relating Probability Theory with Fuzzy Set Theory

In Section 3.4, it is mentioned that the application of Bayesian learning and fuzzy logic as the VHO decision mechanism should be regarded as complementary rather than competitive. Interestingly, there are studies which promote the synergy of fuzzy set theory with probability theory within a Bayesian framework. To be more specific, it is argued in [72] that the membership function of a fuzzy set can be represented as a likelihood function owing to the fact that the former is an non-negative function defined in a subjective manner. Hence, the membership function can be used in place of the likelihood function to combine with a probability-based prior distribution to produce a probability-based posterior. The key advantage of such hybrid approach is the possibility to combine the benefits from: (i) the probability theory when knowledge is extensive, e.g., under nominal conditions; and (ii) the fuzzy set theory when knowledge is lacking, e.g., under anomalous conditions while the Bayesian method provides a framework to combine information from different sources through its updating process. This fuzzy-probability hybrid approach would be particularly attractive for improving the robustness of the cognition process within cognitive networks where the operating environment is highly non-linear and often unpredictable.

8.2.3 Integration of Reactive and Predictive Algorithms

In Section 7.6.3, it is demonstrated that the RQB algorithm is capable of exploiting spare capacity opportunistically which maximizes the capacity utilization of a CWN whilst preserving baseline QoS. This is a remarkable result of achieving the end-to-end goal of effectuating a QoS-balanced system. However, the measurement-based approach of the RQB algorithm implies that system cost would be inherently higher than that of the model-based technique used in the PQB algorithm. Therefore, the integration of both reactive and predictive dynamic load distribution algorithms could result in favorable trade-offs between system cost and modeling complexity. In particular, the predictive algorithm could be deployed under low to medium load conditions to regulate traffic load while the reactive algorithm may be invoked only during heavy load situations to exploit higher capacity utilization through opportunistic load distributions. Such an integrated scheme would also reduce the power consumption of the AP or base station and lead to ‘green’ networks just by invoking the respective algorithms appropriately. Furthermore, the advent of reconfigurable multi-mode, SDR-based devices with the enhanced capability to access multiple network (at the same time) raises the issues of power consumption and signaling load, which require further investigations.

Appendix

A-1 Discrete Event Simulation Package

OPNET™ Modeler® [164] is the de-facto standard for network research and development, modeling and simulation used by defense organization, research communities, and leading network equipment manufacturers. It is based on discrete event simulations which generate a sequence of states for a given system model. The model then evolves through these states over time, in which state variables change only at discrete points in time, based on the behavior of model components and their interactions. These points in time correspond to the occurrences of events which are defined as instantaneous occurrences that may either change the state of the system or make some decisions. Such evolution is representative of the way an actual system functions over time, provided that the model specifications are accurate.

OPNET™ Modeler® adopts a hierarchical modeling concept that consists of three domains, viz. network domain, node domain, and process domain. The network domain at the top level defines the network topology consisting of communication entities known as nodes and links to allow communications between nodes. The node domain at the second level comprises of a set of modules that describes the node's functionality and connections, which allow information to flow between modules. The modules in the nodes are implemented using process models of the process domain at the lowest level. The process models provide behavioral modeling for programmable modules using Proto C based on a combination of state transition diagrams, a library of high-level commands known as kernel procedures, and general facilities of the C/C++ programming language.

Furthermore, OPNET™ Modeler® also provides a suite of the IEEE 802.11 WLAN models that accurately emulate the operation of STA by implementing the complete CSMA/CA protocol and parameters such as packet transmission times, propagation delays, turnaround times, and timer values in accordance to the IEEE 802.11 standard. In this context, OPNET™ Modeler® Release 12.0 PL5 (Build 4523) to Release 14.5 PL8 (Build 7808) with Wireless Module is the discrete event simulator employed throughout

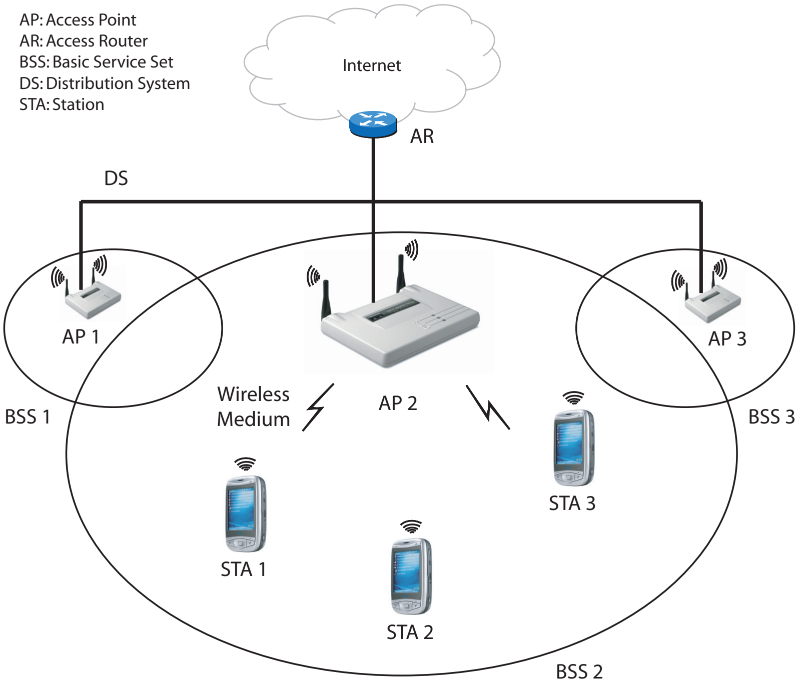


Figure A-1: IEEE 802.11 infrastructure BSS and extended service set.

this thesis to analyze the performance of the generalized CCRRM architecture whereby its proof of concept is developed based on the existing IEEE 802.11 WLAN models. Additionally, a suite of custom models comprising of DAPU, DANS, fast handover, service prioritization, admission control, load control, and link adaptation functions are implemented due to the lack of cooperation and cognition mechanisms in the existing WLAN models.

A-2 IEEE 802.11 WLAN Model

The proof of concept is developed and implemented based on the IEEE 802.11 WLAN [24], [184] with the basic access scheme of DCF. Figure A-1 illustrates the topology of an infrastructure BSS and extended service set considered in the performance evaluation of the generalized CCRRM architecture throughout this thesis.

The IEEE 802.11 networks consist of four major components, viz., STA, AP, wireless medium, and distribution system as shown in Figure A-1. The BSS is the basic building block of an IEEE 802.11 network in which the smallest possible IEEE 802.11 network is an IBSS with two STAs. An IBSS, which is not considered in this thesis, is typically set up among a small group of STAs for a short time and specific purpose, also known as ad hoc networks. On the contrary, this thesis considers an infrastructure BSS which is a commonplace for hotspot deployments in enterprises, airports, and campuses among others. An infrastructure BSS is differentiated from IBSS through the use of an AP which relays all traffic within a BSS including the traffic between STAs. This implies that no communication is possible between STAs[1] and any communication between STAs within the same BSS requires two-hop transmissions via the AP.

In order to provide a larger network coverage, BSSs can be linked via a distribution system to form an extended service set in which all APs in an extended service set are identified by the same service set identity. Note that, however, the distribution system is not part of an extended service set. The key concept of an extended service set is that it appears as a single BSS to the logical link control layer. Hence, STAs within an extended service set may communicate with each other and STAs may move from one BSS to another within the same extended service set by treating the backbone network as a single link layer domain where APs operate as bridges. Accordingly, BSS 2 in Figure A-1 represents an infrastructure BSS while the union of BSS 1, BSS 2, and BSS 3 represents an extended service set.

The DCF is based on CSMA/CA, whereas ethernet, its wireline counterpart, is based on carrier sense multiple access with collision detection. As part of the CSMA/CA mechanism shown in Figure A-2, a STA is required to perform backoff procedure before starting a transmission to reduce the probability of collisions. STA with a MSDU to deliver is allowed to initiate its transmission provided that the wireless medium is sensed as idle for

---

[1] Note that this restriction will be lifted with the emerging IEEE 802.11e standard which offers direct link setup to enable direct STA-to-STA frame transfer within a BSS.

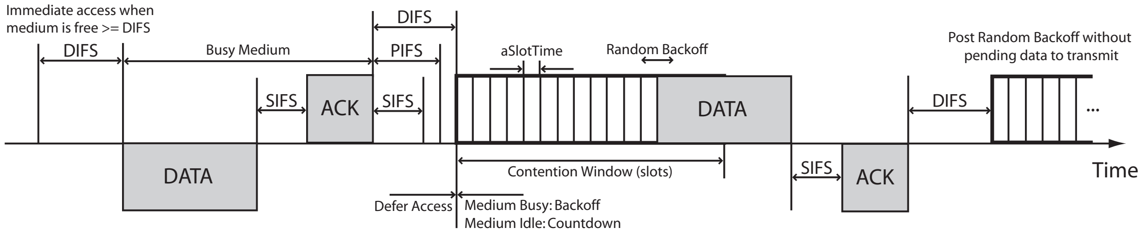


Figure A-2: DCF access procedure.

a minimum duration of DIFS, in addition to the random backoff time. The duration of random backoff time is determined as a multiple of a random function and a slot duration (aSlotTime). The random function produces a uniformly distributed pseudo-random integer over the interval $[0, CW]$ where $CW_{min} \leq CW \leq CW_{max}$. Each STA maintains a CW, which dictates the number of slot times it has to wait prior to transmission, with an initial value of CW_{min}. Upon an unsuccessful transmission, i.e., in the absence of a positive ACK, the subsequent backoff is performed with a doubled CW size which further reduces the probability of collision should there be multiple STAs attempting to access the channel. The CW size is upper bounded by CW_{max} and will be reset to CW_{min} upon successful transmission. Figure A-3 illustrates the exponential increase of CW, also known as the binary exponential backoff.

After each successful transmission, a post random backoff is mandatory, even though there is no other pending MSDUs, to guarantee that any frame (with the exception of the first MSDU arriving at an empty queue with an idle medium) will be delivered with backoff. For unsuccessful transmissions, a short retry counter will be incremented for frame size shorter than the RTS threshold and a long retry counter will be incremented for frame size longer than the RTS threshold before retransmissions. The frame would be discarded when either the short retry counter or long retry counter exceeds the short retry limit of seven and long retry limit of four, respectively. The short retry counter would be reset upon the successful transmission of frame shorter than the RTS threshold while the long retry counter would be reset upon the successful transmission of frame longer than

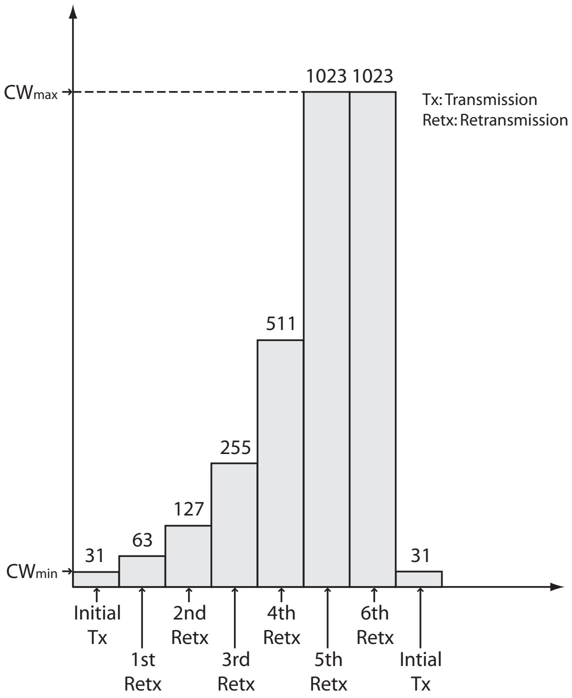


Figure A-3: Binary exponential backoff.

the RTS threshold. The time between two MAC frames is called an interframe space from which the short interframe space (SIFS) is used for the highest priority transmissions such as ACK frames. The point coordination function interframe space is used by the point coordination function during contention-free operation. Lastly, the DIFS is the minimum medium idle time for contention-based services.

A-2.1 Network Configurations

A wireline-to-wireless network topology is adopted throughout this thesis in order to focus on the achievable QoS metrics within each BSS. The QoS metrics of interest are PD and PLR that could characterize the 'goodness' of multimedia traffic which is expected to dominate in future wireless network, envisaged as an IP-based multi-RAT environment. Given that RT traffic such as VoIP and video requires one-way end-to-end delay of less than 150 ms according to the ITU-T's recommendation in [88], a codec delay of 40 ms, packetization delay of 20 ms at both sender and receiver, and backbone network delay of 30 ms are assumed. Therefore, the PD of WLAN in both UL and DL should be less than 60 ms in order to meet the one-way end-to-end delay requirement of RT packets. In addition, RT traffic such as VoIP can tolerate some PLR of up to 2% [89].

The network interface parameters are modeled according to the Cisco Aironet 1130AG series AP and CB21AG series wireless client reference interface specifications to improve the realism of the simulations. Note that the considered multi-AP WLAN or extended service set has at most three BSSs which operate in non-overlapping channels, and hence inter-cell interference does not exist. In addition, interferences from non-802.11 sources are not considered. Furthermore, all STAs are roaming capable to support handover events without multi-homing capability. Handover events are coordinated to one event every ten QoS broadcast intervals to prevent handover synchronization problem discussed in [5] as the objective is to investigate the performance gains of the generalized CCRRM architecture by its virtue of precluding unnecessary handovers.

In cases when the IEEE 802.11g AP exists, it is assumed that at least one legacy STA is associated with the IEEE 802.11g AP since it is unlikely to operate in the IEEE 802.11g only mode given the vast IEEE 802.11b deployments. However, the legacy STA does not transmit any traffic so that all the system resources are available for the IEEE 802.11g STAs. Additionally, the CTS-to-self protection mechanism is used by the IEEE 802.11g STAs. The MSDU lifetime for voice, video, and data packets are chosen as 50 ms, 150 ms, and 1 s, respectively in cases when MSDU lifetime limit mechanism is incorporated to discard MSDUs from the transmitter queue if they exceed the MSDU lifetime before successful transmission. This feature is favorable for delay-sensitive RT traffic as it minimizes unnecessary bandwidth consumption for transmission of MSDUs that have exceeded their useful lifetime, i.e., late frames or ACK frames in attempt to acknowledge these late frames.

A-2.2 Wireless Channel Models

The default OPNET™ PHY model considers wireless channels as AWGN channels and utilizes a fixed value of path loss exponent (with a value of 2 to represent free space path loss), which ignore the effects of fading and different propagation environments, respectively. Unless otherwise stated, the default OPNET™ PHY model is used for performance evaluation. In occasions where appropriate, the OPNET™ PHY model is enhanced to include shadow fading, Rayleigh flat fading, or the capability to simulate NLOS transmissions by varying the path loss exponent.

Log-normal Path Loss Model

An error-prone channel is considered by including shadow fading and variable path loss exponent to capture different propagation environments which may contain obstacles to cause NLOS transmissions. Accordingly, the log-normal path loss model is derived in [165] as

$$L_{path}\left(dB\right) = 10\log\left[\frac{\left(4\pi\right)^2 d^n}{\lambda^2}\right] + \mathrm{X}_\sigma \qquad \text{(A-1)}$$

where λ is the wavelength associated with the carrier frequency, d is the distance between the transmitters and receivers, n is the path loss exponent which determines the rate of loss, and X_σ is a zero-mean Gaussian distributed random variable with standard deviation σ (in dB). NLOS transmissions due to obstructions that are prevalent in hotspots and indoor environments are simulated by varying n uniformly $U\left[3,4\right]$ and shadow fading is accounted by varying σ uniformly $U\left[6,9\right]$. Additionally, background noise is modeled as

$$N = FkT_0B \qquad \text{(A-2)}$$

where k is the Boltzmann's constant, T_0 is the ambient room temperature of $290K$, B is the equivalent bandwidth of the receiver, and F is the noise figure of the receiver. Note that F is simulated with a normal distribution $N(5, 0.1)$ for the performance evaluation of QLO framework and $N(10, 0.1)$ for the performance evaluation of LAS framework.

Exponential Channel Model

The exponential channel model, adopted by the IEEE 802.11 Task Group-b for its simplicity and accuracy, is modeled in [184] as a finite impulse response filter where k_{max} is the maximum tap length. The taps are independent complex Gaussian distributed random variables with an exponentially decaying power delay profile in which the coefficient of kth tap is given by

$$h_k = N\left(0, \frac{1}{2}\sigma_k^2\right) + jN\left(0, \frac{1}{2}\sigma_k^2\right), \quad k = 0, \ldots, k_{\max}, \tag{A-3a}$$

$$k_{\max} = \left\lceil 10\frac{T_{RMS}}{T_s} \right\rceil, \quad \beta = e^{T_s/T_{RMS}}, \quad \sigma_k^2 = \frac{(1-\beta)\,\beta^k}{1-\beta^{k_{\max}+1}} \tag{A-3b}$$

where $N\left(0, \frac{1}{2}\sigma_k^2\right)$ is a zero-mean Gaussian distributed random variable with variance $\frac{1}{2}\sigma_k^2$. T_s is the sampling period and T_{RMS} is the root mean square delay spread of the channel. Thus, Rayleigh flat fading is simulated as a special case of the exponential channel model by considering a single tap channel with zero root mean square delay spread, i.e, setting $k_{max} = 0 \Rightarrow \sigma_k^2 = 1$ such that

$$h_0 = N\left(0, \frac{1}{2}\right) + jN\left(0, \frac{1}{2}\right). \tag{A-4}$$

A-2.3 Traffic Models

Simple Traffic Sources

In general, traffic is generated based on simple traffic sources which are available as standard built-in OPNET™ models. The simple traffic source characterizes the packet size and packet interarrival time using a particular probability distribution. Both CBR and VBR traffic sources are considered in this thesis. While the interarrival time for CBR traffic source is deterministic, the derivation of interarrival time for VBR traffic source is more involved. Hence, VBR traffic source is typically modeled as a Markov modulated Poisson process [161], commonly known as an ON-OFF source. Accordingly, the packet stream from a single voice source is modeled as a renewal process in which packets arrive at fixed packetization intervals of T_p during talk spurts and no packet is generated during silence periods. The interarrival time distribution of this renewal process is given by

$$R\left(t\right)=\left[\left(1-\frac{T_p}{T_{on}}\right)+\frac{T_p}{T_{on}}\left(1-e^{-\left(t-T_p\right)/T_{off}}\right)\right]U\left(t-T_p\right) \qquad \text{(A-5)}$$

where $U\left(t\right)$ is the unit step function. The talk spurt is exponentially distributed with a mean of T_{on} and the silence period is exponentially distributed with a mean of T_{off}, both of which alternate according to a continuous time Markov chain. Taking Laplace transform then gives

$$\tilde{R}\left(s\right)=\int_0^{\infty}e^{-st}dR\left(t\right)=\left[1-\frac{T_p}{T_{on}}+\frac{T_p}{T_{on}\left(1+sT_{off}\right)}\right]e^{-sT_p} \qquad \text{(A-6)}$$

where the mean packet arrival rate is $\left(T_p+T_pT_{off}/T_{on}\right)^{-1}$ and T_p is the fixed packet interarrival time. The peak packet arrival rate of $1/T_p$ which corresponds to that of a CBR traffic source is found by setting $T_{on}\to\infty$ and $T_{off}\to 0$.

For the generation of VoIP packets, it is assumed that header compression is not used. Hence, an additional 40 bytes RTP/UDP/IP header is added to the voice payload. Two types of voice packet generation are considered using either CBR or VBR traffic source. In the case when VoIP traffic is simulated with CBR source, each voice session generates one packet per packetization interval continuously. In the case of when VoIP traffic is simulated with VBR source using ON-OFF model, the voice packets are generated only during the ON period. According to the ITU-T's recommendation in [185], ON and OFF time can be approximated by an exponential distribution with mean values of 1.004 s and 1.587 s, respectively for a speech activity of 39%.

Application Traffic Models

Applications are the predominant sources of traffic in the network. Often, complex models are required to characterize network applications as they cannot be characterized by simple distributions, and hence must allow protocol interactions. OPNET™ provides a whole array of application traffic models including pre-configured standard application models of commonly used network applications such as voice, video conferencing, and FTP. In order to simulate realistic scenarios, the standard application models are employed in the performance evaluation of Chapter 3 which constitute the core component of the generalized CCRRM architecture.

A-2.4 Hidden Terminal Problem

The hidden terminal problem is not considered, and hence RTS/CTS mechanism is excluded from the simulation as the proof of concept in this thesis is based on a multi-AP WLAN which is essentially an extended service set of interconnected infrastructure BSSs. This is a reasonable assumption as the impact of hidden terminal will be minimal in an infrastructure BSS given that all traffic must be relayed via an AP and no traffic can flow between STAs. Moreover, commercial chipsets typically have a carrier sensing range to transmission range ratio R_{cs}/R_{tx} of greater than 1. E.g., Lucent Orinoco WaveLAN chipset and Intersil chipset have a carrier sensing range to transmission range ratio of 2.2 and 1, respectively. In the former, there will be no hidden terminals since all STAs are able to sense one another transmissions. In the latter, it could happen that one STA is a hidden terminal to another transmitting STA only if both are at the edge of the cell. However, the probability of this occurring will be low. As a result, there will be no hidden

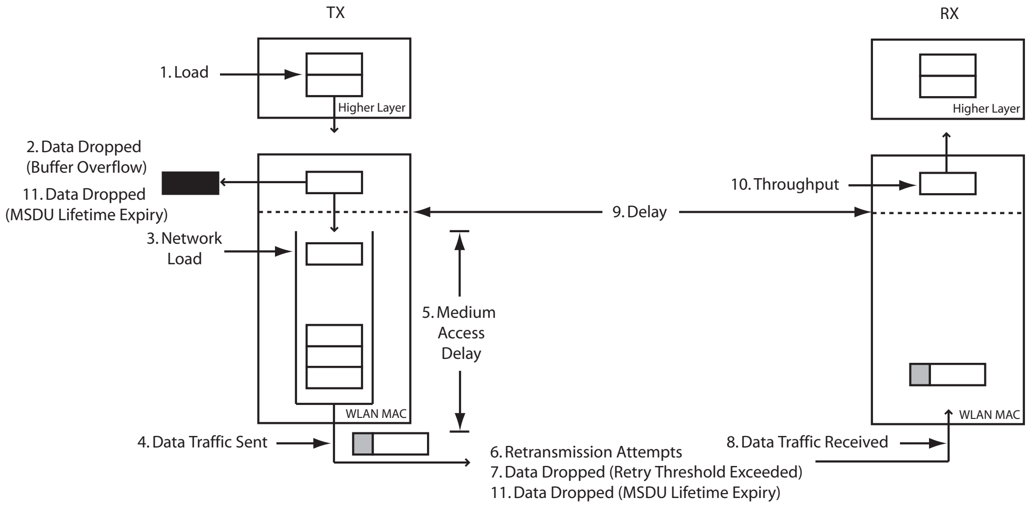

TX
RX
1. Load
Higher Layer
Higher Layer
2. Data Dropped (Buffer Overflow)
11. Data Dropped (MSDU Lifetime Expiry)
3. Network Load
10. Throughput
9. Delay
5. Medium Access Delay
WLAN MAC
WLAN MAC
4. Data Traffic Sent
6. Retransmission Attempts
7. Data Dropped (Retry Threshold Exceeded)
11. Data Dropped (MSDU Lifetime Expiry)
8. Data Traffic Received

Figure A-4: Common performance indicators.

terminals to the AP and all collisions due to hidden terminals, which occur with a remote probability as $R_{cs}/R_{tx} \geq 1$, would happen only at the AP.

A-2.5 Mobility

Although handover is crucial for both mobility and load balancing, the focus of this thesis is on load distribution through VHO or QoS-based handover to redistribute load to a better quality or less loaded AP opportunistically. Therefore, the end-users are assumed to have a quasi-static mobility pattern. In other words, the end-users could move from place to place, but they tend to remain in the same physical location for an extended period of time. This assumption is plausible and consistent with the recent studies in [186] and [187] on the behavior of mobile users in different environments.

A-3 Performance Indicators

The common performance indicators used throughout this thesis are depicted in Figure A-4 and defined as follows:

1. Load: is the total offered load submitted to the WLAN MAC by the higher layer of a particular node.

2. Data dropped (buffer overflow): is the total higher layer data traffic dropped by the WLAN MAC due to: (i) the full higher layer data buffer; or (ii) the size of the higher layer packet being greater than the maximum allowed data size defined in the IEEE 802.11 standard.

3. Network load: is the total data traffic received by the entire WLAN BSS from the higher layers of the MACs that is accepted and queued for transmission. This metric does not include any higher layer data traffic that is rejected without queueing due to buffer overflow. In addition, any data traffic that is relayed by the AP from its source to its destination within the BSS is counted twice (once at the source node and once at the AP).

4. Data traffic sent: is the total WLAN data traffic transmitted by the MAC. The PHY and MAC headers of the packet are included when computing the size of transmitted packets.

5. Medium access delay: is the total queueing and contention delays of the data, management, delayed Block-ACK, and Block-ACK Request frames transmitted by the WLAN MAC. For each frame, this metric is calculated as the duration from the time when it is inserted into the transmission queue, which is the arrival time for higher layer data packets and creation time for all other frames types, until the time when the frame is sent to the PHY for the first time.

6. Retransmission attempts: is the number of retransmission attempts until either packet is successfully transmitted or discarded as a result of reaching the short or long retry limit.

7. Data dropped (retry threshold exceeded): is the total higher layer data traffic dropped by the WLAN MAC due to consistently failing retransmissions. This metric reports

the number of the higher layer packets that are dropped because the MAC could not receive any ACKs for the (re)transmissions of those packets or their fragments, and the packets' short or long retry count has reached the MAC's short or long retry limit, respectively.

8. Data traffic received: is the total WLAN data traffic successfully received by the MAC from the PHY. This metric includes all data traffic received, regardless of the destination of the received frames. The PHY and MAC headers of the packet are included when computing the size of the received packets.
9. Delay: is the end-to-end delay of all the data packets that are successfully received by the WLAN MAC and forwarded to the higher layer.
10. Throughput: is the total data traffic successfully received and forwarded to the higher layer by the WLAN MAC.
11. Data dropped (MSDU lifetime expiry): is the total higher layer data traffic dropped by the WLAN MAC due to MSDU lifetime expiry during: (i) the queueing process; or (ii) the retransmission process.

A-3.1 Key Performance Indicators

Apart from the common performance indicators, a set of KPIs are defined to evaluate the performance of the generalized CCRRM architecture from three critical aspects as follows:

1. The QoS performance of STA or end-user is quantified as a function of two QoS metrics. Each QoS element is the ratio of the required QoS metric threshold and measured QoS value. Accordingly, QSF is defined as the minimum between the two QoS elements given by

$$QSF_{i,j} = \min_{k \in Links} \left(\frac{PD_i^t}{PD_{i,j}^{m,k}}, \frac{PLR_i^t}{PLR_{i,j}^{m,k}} \right), \tag{A-7a}$$

$$\overline{QSF} = \frac{1}{cn}\sum_{i=1}^{c}\sum_{j=1}^{n} QSF_{i,j} \tag{A-7b}$$

where PD_i^t and PLR_i^t are the PD threshold and PLR threshold, respectively. $PD_{i,j}^{m,k}$ and $PLR_{i,j}^{m,k}$ are the measured PD and measured PLR of ith service class and jth STA of k links, respectively. $\overline{QSF}$ is the average QSF of all STAs. $QSF < 1$ when the QoS requirements of STAs cannot be met. For the purpose of QSF computation, PD thresholds for voice and video are 50 ms while data is 300 ms, and PLR thresholds for all service classes are 2%. It is important to note that for infrastructure BSS, particularly, in the presence of many two-way communications such as VoIP connections, DL becomes the capacity bottleneck since each VoIP connection has duplex traffic which will eventually result in higher DL load and asymmetric traffic load on both links (see, e.g., Figure 6.11 of Section 6.4.1, [82], and [83]). Therefore, unless otherwise stated, the PD of interest is taken as the MAC delay experienced by the AP. On the other hand, the PLR is a function of data dropped due to buffer overflow, MSDU lifetime expiry, and retry threshold exceeded. Note that QSF in the performance evaluation of QLO framework considers only DL QoS metrics while QSF in the performance evaluation of LAS framework considers both UL and DL QoS metrics.

2. The effect of load distribution on the achievable throughput of STA or end-user is quantified by TFI which is defined as

$$TFI_{i,j} = \frac{\left|\bar{S}_{i,j}^m - S_i^t\right|}{S_i^t}, \tag{A-8a}$$

$$\overline{TFI} = \frac{1}{cn}\sum_{i=1}^{c}\sum_{j=1}^{n} TFI_{i,j} \tag{A-8b}$$

where $\bar{S}_{i,j}^m$ and S_i^t are the measured average throughput and target (maximum) throughput for ith service class and jth STA, respectively. $\overline{TFI}$ is the average TFI of all STAs. $TFI \in [0, 1]$ reflects the amount of deviation from target throughput and serves as a normalized measure of throughput fairness in the network. It has a value of 0 when the throughput of STA is fair and a value of 1 when the throughput of STA is tremendously unfair.

3. The Jain's fairness index [188] is adopted to quantify the effect of load distribution on the QoS fairness among the APs (STAs). Suppose x_i is the QoS metric (QSF) of AP (STA) i, then QBI can be defined as

$$QBI\left(x\right)=\left(\sum_{i=1}^{n}x_i\right)^2\bigg/n\left(\sum_{i=1}^{n}x_i^2\right) \tag{A-9}$$

   where n is the number of APs (STAs.) over which the load, i.e., STAs will be redistributed. $QBI\in[0,1]$ is a continuous function which is independent of scale. It has a value of 1 when all APs (STAs) have exactly the same QoS metric (QSF) and a value of $1/n$ when the QoS metric (QSF) of each AP (STA) is extremely unbalanced, which is 0 in the limit as $n\to\infty$.

4. The overall composite capacity as a result of deploying different dynamic load distribution algorithms is evaluated by computing the composite capacity efficiency η_{cc} which can be expressed as

$$\eta_{cc}=\frac{\sum_{n=1}^{APs}CU_{total}^{n}\times(1-PLR_n)}{\sum_{n=1}^{APs}CU_{max}^{n}},\quad CU_{max}^{n}=1.0 \tag{A-10}$$

   where CU_{total}^{n}, CU_{max}^{n}, and PLR_n of nth AP are defined in (5.2a), (5.2b), and (6.34), respectively. $\eta_{cc}\in[0,1]$ is a dimensionless performance measure of the effective composite capacity as a ratio to the maximum composite capacity, which ranges from 0 to 100%.

A-4 Statistical Validity of Discrete Event Simulations

A simulation result is considered statistically significant if it is unlikely to have occurred by pure chance. In general, system models that include stochastic behavior have results that are dependent on the initial seeding of the random number generator. Since a particular random seed selection can potentially result in an anomalous or non-representative behavior, it is important for each model configuration to be exercised with several random number seeds in order to determine a typical, representative behavior. The basic principle of statistical validity applied here is that if a typical behavior exists and many independent trials are performed, it is likely that a significant majority of these trials will fall within a close range of the representative behavior. Therefore, all results presented throughout this thesis are computed from multiple simulation runs in which each seed value of random number generator is different. For every set of simulation runs, the mean of a performance indicator is first calculated followed by its standard error which is essentially the standard deviation of the sampling distribution of the mean. Consequently, all the results are obtained with 95% confidence level where the confidence intervals are derived from the standard error that is in general less than 2%. The error bars are presented for critical results or results with a standard error of more than 2%. Otherwise, it can be taken that the standard error of these results is statistically insignificant when the error bars are not shown.

BIBLIOGRAPHY

[1] ITU-R M.1645. Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000. June 2003.

[2] E. Gustafsson and A. Jonsson. Always best connected. *IEEE Wireless Communications*, 10(1):49–55, February 2003.

[3] S. Seshan. *Low-Latency Handoff for Cellular Data Networks*. PhD thesis, University of California at Berkeley, Berkeley, CA, USA, 1995.

[4] M. Stemm and R. H. Katz. Vertical handoffs in wireless overlay networks. *Mobile Networks and Applications*, 3(4):335–350, 1998.

[5] H. J. Wang, R. H. Katz, and J. Giese. Policy-enabled handoffs across heterogeneous wireless networks. In *Proc. Second IEEE Workshop on Mobile Computing Systems and Applications, 1999. WMCSA '99*, pages 51–60, February 1999.

[6] FCC. Spectrum policy task force report. In *ET Docket No. 02-135*, November 2002.

[7] J. Mitola. *Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio*. Doctor of Technology, Royal Institute of Technology (KTH), Stockholm, Sweden, May 2000.

[8] S. Haykin. Cognitive radio: Brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, 23(2):201–220, February 2005.

[9] N. Niebert, A. Schieder, J. Zander, and R. Hancock, editors. *Ambient Networks: Co-operative Mobile Networking for the Wireless World*. John Wiley & Sons, Ltd, West Sussex, England, 2007.

[10] MobiLife Project Website. http://www.ist-mobilife.org/. September 2004.

[11] SPICE Project Website. http://www.ist-spice.org/. January 2006.

[12] WINNER Project Website. http://www.ist-winner.org. January 2004.

[13] E2R Project Website. http://www.e2r.motlabs.com. January 2004.

[14] F. H. P. Fitzek and M. D. Katz, editors. *Cognitive Wireless Networks: Concepts, Methodologies and Visions Inspiring the Age of Enlightenment of Wireless Communications*. Springer, Dordrecht, The Netherlands, 2007.

[15] E. Hossain, editor. *Heterogeneous Wireless Access Networks: Architectures and Protocols*. Springer, New York, NY, USA, 2008.

[16] IEEE 802.21-2008. Part 21: Media independent handover services. January 2009.

[17] R. Koodli and C. Perkins. *Mobile IPv4 fast handovers*. IETF RFC 4988, October 2007.

[18] S. Krishnan (editor), N. Montavont, E. Njedjou, S. Veerepalli, and A. Yegin (editor). *Link-layer event notifications for detecting network attachments*. IETF RFC 4957, August 2007.

[19] K. E. Malki, editor. *Low-latency handoffs in mobile IPv4*. IETF RFC 4881, June 2007.

[20] HURRICANE Project Website. http://www.ict-hurricane.eu. January 2008.

[21] S. Gundavelli (editor), K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil. *Proxy mobile IPv6*. IETF RFC 5213, August 2008.

[22] J. Kempf, editor. *Problem statement for network-based localized mobility management (NETLMM)*. IETF RFC 4830, April 2007.

[23] L. Sarakis, G. Kormentzas, and F. M. Guirao. Seamless service provision for multi heterogeneous access. *IEEE Wireless Communications*, 16(5):32–40, October 2009.

[24] IEEE 802.11-2007. Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications. June 2007.

[25] S. Jeux, G. Mange, P. Arnold, and F. Bernardo, editors. *E3 Deliverable D3.3: Simulation based recommendations for DSA and self-management*. July 2009.

[26] IEEE 1900.4-2009. IEEE standard for architectural building blocks enabling network-device distributed decision making for optimized radio resource usage in heterogeneous wireless access networks. February 2009.

[27] G. Wu, M. Mizuno, and P. J. M. Havinga. MIRAI architecture for heterogeneous network. *IEEE Communications Magazine*, 40(2):126–134, February 2002.

[28] C. Perkins, editor. *IP mobility support for IPv4*. IETF RFC 3344, August 2002.

[29] D. Johnson, C. Perkins, and J. Arkko. *Mobility support in IPv6*. IETF RFC 3775, June 2004.

[30] A.T. Campbell, J. Gomez, S. Kim, A.G. Valko, Chieh-Yih Wan, and Z.R. Turanyi. Design, implementation, and evaluation of cellular IP. *IEEE Personal Communications*, 7(4):42–49, August 2000.

[31] R. Ramjee, T.F. La Porta, L. Salgarelli, S. Thuel, K. Varadhan, and L. Li. IP-based access network infrastructure for next-generation wireless data networks. *IEEE Personal Communications*, 7(4):34–41, August 2000.

[32] R. Tafazolli, editor. *Technologies for the Wireless Future: Wireless World Research Forum (WWRF)*, volume 1. John Wiley & Sons, Ltd, West Sussex, England, 2005.

[33] S. Frattasi, H. Fathi, F.H.P. Fitzek, R. Prasad, and M.D. Katz. Defining 4G technology from the users perspective. *IEEE Network*, 20(1):35–41, January/February 2006.

[34] F. H. P. Fitzek and M. D. Katz, editors. *Cooperation in Wireless Networks: Principles and Applications: Real Egoistic Behavior is to Cooperate!* Springer, Dordrecht, The Netherlands, 2006.

[35] D. D. Clark, C. Partridge, J. C. Ramming, and J. T. Wroclawski. A knowledge plane for the internet. In *Proc. ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, 2003. SIGCOMM '03*, pages 3–10, August 2003.

[36] R. W. Thomas, L. A. DaSilva, and A. B. MacKenzie. Cognitive networks. In *Proc. First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005*, pages 352 –360, November 2005.

[37] R. W. Thomas, D. H. Friend, L. A. Dasilva, and A. B. Mackenzie. Cognitive networks: Adaptation and learning to achieve end-to-end performance objectives. *IEEE Communications Magazine*, 44(12):51–57, December 2006.

[38] IEEE 1900.1-2008. IEEE standard definitions and concepts for dynamic spectrum access: Terminology relating to emerging wireless networks, system functionality, and spectrum management. September 2008.

[39] W. Tuttlebee, editor. *Software Defined Radio: Enabling Technologies*. John Wiley & Sons, Inc., West Sussex, England, 2002.

[40] B. Shneiderman. *Leonardo's Laptop: Human Needs and the New Computing Technologies*. MIT Press, Cambridge, MA, USA, 2002.

[41] V. Kawadia and P.R. Kumar. A cautionary perspective on cross-layer design. *IEEE Wireless Communications*, 12(1):3 –11, February 2005.

[42] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J. Makela, R. Pichna, and J. Vallstron. Handoff in hybrid mobile data networks. *IEEE Personal Communications*, 7(2):34–47, April 2000.

[43] J. McNair and F. Zhu. Vertical handoffs in fourth-generation multinetwork environments. *IEEE Wireless Communications*, 11(3):8–15, June 2004.

[44] N. Nasser, A. Hasswa, and H. Hassanein. Handoffs in fourth generation heterogeneous networks. *IEEE Communications Magazine*, 44(10):96–103, October 2006.

[45] F. Bari and V. Leung. Service delivery over heterogeneous wireless systems: Networks selection aspects. In *Proc. International Conference on Wireless Communications and Mobile Computing, 2006. IWCMC '06*, pages 251–256, July 2006.

[46] E. H. Ong and J. Y. Khan. Dynamic access network selection with QoS parameters estimation: A step closer to ABC. In *Proc. IEEE 67th Vehicular Technology Conference, 2008. VTC 2008-Spring*, pages 2671–2676, May 2008.

[47] M. Kassar, B. Kervella, and G. Pujolle. An overview of vertical handover decision strategies in heterogeneous wireless networks. *Comput. Commun.*, 31(10):2607–2620, June 2008.

[48] Y. Chen and Y. Yang. A new 4G architecture providing multimode terminals always best connected services. *IEEE Wireless Communications*, 14(2):36–41, April 2007.

[49] E. H. Ong and J. Y. Khan. Cooperative radio resource management framework for future IP-based multiple radio access technologies environment. *Comput. Netw.*, 54(7):1083–1107, May 2010.

[50] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda. Performance anomaly of 802.11b. In *Proc. IEEE Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. INFOCOM 2003*, volume 2, pages 836–843, March/April 2003.

[51] G. Tan and J. Guttag. Time-based fairness improves performance in multi-rate WLANs. In *Proc. USENIX Annual Technical Conference, 2004. USENIX 2004*, Boston, MA, June 2004.

[52] W. Chen, J. Liu, and H. Huang. An adaptive scheme for vertical handoff in wireless overlay networks. In *Proc. Tenth International Conference on Parallel and Distributed Systems, 2004. ICPADS 2004*, pages 541–548, July 2004.

[53] Q. Nguyen-Vuong, Y. Ghamri-Doudane, and N. Agoulmine. On utility models for access network selection in wireless heterogeneous networks. In *Proc. IEEE Network Operations and Management Symposium 2008. NOMS 2008*, April 2008.

[54] Q. Song and A. Jamalipour. An adaptive quality-of-service network selection mechanism for heterogeneous mobile networks. *Wirel. Commun. Mob. Comput.*, 5(6):697–708, August 2005.

[55] E. Stevens-Navarro and V. W. S. Wong. Comparison between vertical handoff decision algorithms for heterogeneous wireless networks. In *Proc. IEEE 63rd Vehicular Technology Conference, 2006. VTC 2006-Spring*, volume 2, pages 947–951, May 2006.

[56] W. Zhang. Handover decision using fuzzy MADM in heterogeneous networks. In *Proc. IEEE Wireless Communications and Networking Conference, 2004. WCNC 2004*, volume 2, pages 653–658, March 2004.

[57] F. Bari and V. C. M. Leung. Automated network selection in a heterogeneous wireless network environment. *IEEE Network*, 21(1):34–40, January/February 2007.

[58] C. Guo, Z. Guo, Q. Zhang, and W. Zhu. A seamless and proactive end-to-end mobility solution for roaming across heterogeneous wireless networks. *IEEE Journal on Selected Areas in Communications*, 22(5):834–848, June 2004.

[59] Q. Nguyen-Vuong, N. Agoulmine, and Y. Ghamri-Doudane. Terminal-controlled mobility management in heterogeneous wireless networks. *IEEE Communications Magazine*, 45(4):122–129, April 2007.

[60] E. Stevens-Navarro, Y. Lin, and V. W. S. Wong. An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks. *IEEE Transactions on Vehicular Technology*, 57(2):1243–1254, March 2008.

[61] S. Lee, K. Sriram, K. Kim, Y. H. Kim, and N. Golmie. Vertical handoff decision algorithms for providing optimized performance in heterogeneous wireless networks. *IEEE Transactions on Vehicular Technology*, 58(2):865–881, February 2009.

[62] E. Piri and K. Pentikousis. Towards a GNU/Linux IEEE 802.21 implementation. In *Proc. IEEE International Conference on Communications, 2009. ICC '09*, pages 1–5, June 2009.

[63] Y. Yu, B. Yong, and C. Lan. Utility-dependent network selection using MADM in heterogeneous wireless networks. In *Proc. IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007*, pages 1–5, September 2007.

[64] K. Yang, I. Gondal, and B. Qiu. Multi-dimensional adaptive SINR based vertical handoff for heterogeneous wireless networks. *IEEE Communications Letters*, 12(6):438–440, June 2008.

[65] D. Chalmers and M. Sloman. A survey of quality of service in mobile computing environments. *IEEE Communications Surveys and Tutorials*, 2(2):2 –10, quarter 1999.

[66] K. Park and W. Willinger, editors. *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons, Inc., New York, NY, USA, 2000.

[67] J. F. Gibbon. *Real-Time Scheduling for Multimedia Services using Network Delay Estimation*. PhD thesis, Boston University, Boston, MA, USA, 1994.

[68] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, USA, 1993.

[69] A. M. Zoubir and B. Boashash. The bootstrap and its application in signal processing. *IEEE Signal Processing Magazine*, 15(1):56–76, January 1998.

[70] E. H. Ong and J. Y. Khan. On optimal network selection in a dynamic multi-RAT environment. *IEEE Communications Letters*, 14(3):217–219, March 2010.

[71] P. M. L. Chan, R. E. Sheriff, Y. F. Hu, P. Conforto, and C. Tocci. Mobility management incorporating fuzzy logic for heterogeneous a IP environment. *IEEE Communications Magazine*, 39(12):42–51, December 2001.

[72] T. J. Ross, J. M. Booker, and W. J. Parkinson, editors. *Fuzzy Logic and Probability Applications: Bridging the Gap*. ASA-Siam, Alexandria, VA, 2002.

[73] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, USA, 1986.

[74] H. Cramer. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[75] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition, 2003.

[76] F. Gustafsson. *Adaptive Filtering and Change Detection*. John Wiley & Sons, Ltd, West Sussex, England, 2000.

[77] S. Liu. An adaptive kalman filter for dynamic estimation of harmonic signals. In *Proc. 8th International Conference on Harmonics And Quality of Power, 1998*, volume 2, pages 636–640, Athens, Greece, October 1998.

[78] P. S. Maybeck. *Stochastic Models, Estimation, and Control*, volume 1. Academic Press, Inc., New York, NY, USA, 1979.

[79] S. McBurney, M. H. Williams, N. K. Taylor, and E. Papadopoulou. Managing user preferences for personalization in a pervasive service environment. In *Proc. The Third Advanced International Conference on Telecommunications, 2007. AICT 2007*, pages 32–32, May 2007.

[80] T. L. Saaty. How to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, 48(1):9–26, September 1990.

[81] B. Golany and M. Kress. A multicriteria evaluation of methods for obtaining weights from ratio-scale matrices. *European Journal of Operational Research*, 69(2):210–220, September 1993.

[82] S. Shin and H. Schulzrinne. Balancing uplink and downlink delay of VoIP traffic in WLANs using adaptive priority control (APC). In *Proc. The Third International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks, 2006. QShine '06*, volume 191, August 2006.

[83] F. Anjum, M. Elaoud, D. Famolari, A. Ghosh, R. Vaidyanathan, A. Dutta, P. Agrawal, T. Kodama, and Y. Katsube. Voice performance in WLAN networks: An experimental study. In *Proc. IEEE Global Telecommunications Conference, 2003. GLOBECOM '03*, volume 6, pages 3504–3508, December 2003.

[84] C. M. Chou and C. Huang. Dynamic vertical handover control algorithm for WLAN and UMTS. In *Proc. IEEE Wireless Communications and Networking Conference, 2006. WCNC 2006*, volume 1, pages 606–610, April 2006.

[85] F. H. P. Fitzek and M. Reisslein. MPEG-4 and H.263 video traces for network performance evaluation. *IEEE Network*, 15(6):40–54, November/December 2001.

[86] R. Skehill, M. Barry, W. Kent, M. O'Callaghan, N. Gawley, and S. Mcgrath. The common RRM approach to admission control for converged heterogeneous wireless networks. *IEEE Wireless Communications*, 14(2):48–56, April 2007.

[87] IEEE 802.11n 2009. Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications, amendment 5: Enhancements for higher throughput. October 2009.

[88] ITU-T G.114. One-way transmission time. May 2003.

[89] C. Shim, L. Xie, B. Zhang, and C. Sloane. How delay and packet loss impact voice quality in VoIP. Technical report, Qovia, Inc., December 2003.

[90] A. Mishra, M. Shin, and W. Arbaugh. An empirical analysis of the IEEE 802.11 MAC layer handoff process. *SIGCOMM Comput. Commun. Rev.*, 33(2):93–102, April 2003.

[91] H. Velayos and G. Karlsson. Techniques to reduce the IEEE 802.11b handoff time. In *Proc. IEEE International Conference on Communications, 2004. ICC '04*, volume 7, pages 3844–3848, June 2004.

[92] S. Garg and M. Kappes. An experimental study of throughput for UDP and VoIP traffic in IEEE 802.11b networks. In *Proc. IEEE Wireless Communications and Networking Conference, 2003. WCNC 2003*, volume 3, pages 1748–1753, March 2003.

[93] H. Zhai, X. Chen, and Y. Fang. How well can the IEEE 802.11 wireless LAN support quality of service? *IEEE Transactions on Wireless Communications*, 4(6):3084–3094, November 2005.

[94] X. Chen, H. Zhai, X. Tian, and Y. Fang. Supporting QoS in IEEE 802.11e wireless LANs. *IEEE Transactions on Wireless Communications*, 5(8):2217–2227, August 2006.

[95] G. Bianchi and I. Tinnirello. Improving load balancing mechanisms in wireless packet networks. In *Proc. IEEE International Conference on Communications, 2002. ICC '02*, volume 2, pages 891–895, May 2002.

[96] S. Garg and M. Kappes. Admission control for VoIP traffic in IEEE 802.11 networks. In *Proc. IEEE Global Telecommunications Conference, 2003. GLOBECOM '03*, volume 6, pages 3514–3518, December 2003.

[97] H. Zhai, J. Wang, and Y. Fang. Providing statistical QoS guarantee for voice over IP in the IEEE 802.11 wireless LANs. *IEEE Wireless Communications*, 13(1):36–43, February 2006.

[98] A. Bazzi, M. Diolaiti, and G. Pasolini. Measurement based call admission control strategies in infrastructured IEEE 802.11 WLANs. In *Proc. IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications, 2005. PIMRC 2005*, volume 3, pages 2093–2098, September 2005.

[99] R. Daher and D. Tavangarian. QoS-oriented load balancing for WLANs. In *Proc. 1st Workshop on Operator-Assisted (Wireless Mesh) Community Networks, 2006*, pages 1–12, September 2006.

[100] A. Balachandran, P. Bahl, and G. M. Voelker. Hot-spot congestion relief in public-area wireless networks. In *Proc. Fourth IEEE Workshop on Mobile Computing Systems and Applications, 2002*, pages 70–80, June 2002.

[101] H. Velayos, V. Aleo, and G. Karlsson. Load balancing in overlapping wireless LAN cells. In *Proc. IEEE International Conference on Communications, 2004. ICC '04*, volume 7, pages 3833–3836, June 2004.

[102] E. H. Ong and J. Y. Khan. QoS provisioning for VoIP over wireless local area networks. In *Proc. IEEE 11th International Conference on Communication Systems, 2008. ICCS 2008*, pages 906–911, November 2008.

[103] E. H. Ong and J. Y. Khan. An integrated load balancing scheme for future wireless networks. In *Proc. IEEE Global Telecommunications Conference Workshops, 2008. GLOBECOM Workshops '08*, pages 1–6, November/December 2008.

[104] I. Ramani and S. Savage. SyncScan: Practical fast handoff for 802.11 infrastructure networks. In *Proc. IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies. INFOCOM 2005*, volume 1, pages 675–684, March 2005.

[105] A. Nafaa. Provisioning of multimedia services in 802.11-based networks: Facts and challenges. *IEEE Wireless Communications*, 14(5):106–112, October 2007.

[106] I. Tinnirello and S. Choi. Temporal fairness provisioning in multi-rate contention-based 802.11e WLANs. In *Proc. Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks, 2005. WoWMoM 2005*, pages 220–230, June 2005.

[107] A. Banchs, A. Azcorra, C. Garcia, and R. Cuevas. Applications and challenges of the 802.11e EDCA mechanism: An experimental study. *IEEE Network*, 19(4):52–58, July/August 2005.

[108] J. Lee, W. Liao, J. Chen, and H. Lee. A practical QoS solution to voice over IP in IEEE 802.11 WLANs. *IEEE Communications Magazine*, 47(4):111–117, April 2009.

[109] G. Bianchi, I. Tinnirello, and L. Scalia. Understanding 802.11e contention-based prioritization mechanisms and their coexistence with legacy 802.11 stations. *IEEE Network*, 19(4):28–34, July/August 2005.

[110] P. E. Engelstad and O. N. Osterbo. Non-saturation and saturation analysis of IEEE 802.11e EDCA with starvation prediction. In *Proc. 8th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems, 2005. MSWiM '05*, pages 224–233, October 2005.

[111] X. Yang and N. H. Vaidya. Priority scheduling in wireless ad hoc networks. In *Proc. 3rd ACM International Symposium Mobile Ad Hoc Networking and Computing. MOBIHOC 2002*, pages 71–79, June 2002.

[112] W. Pattara-Atikom, P. Krishnamurthy, and S. Banerjee. Distributed mechanisms for quality of service in wireless LANs. *IEEE Wireless Communications*, 10(3):26–34, June 2003.

[113] P. Wang, H. Jiang, and W. Zhuang. IEEE 802.11e enhancement for voice service. *IEEE Wireless Communications*, 13(1):30–35, February 2006.

[114] D. Gao, J. Cai, and K. N. Ngan. Admission control in IEEE 802.11e wireless LANs. *IEEE Network*, 19(4):6–13, July/August 2005.

[115] D. Deng and H. Yen. Quality-of-service provisioning system for multimedia transmission in IEEE 802.11 wireless LANs. *IEEE Journal on Selected Areas in Communications*, 23(6):1240–1252, June 2005.

[116] A. Nyandoro, L. Libman, and M. Hassan. Service differentiation using the capture effect in 802.11 wireless LANs. *IEEE Transactions on Wireless Communications*, 6(8):2961–2971, August 2007.

[117] J. Yu and S. Choi. Comparison of modified dual queue and EDCA for VoIP over IEEE 802.11 WLAN. *Euro. Trans. Telecomms.*, 17(3):371–382, May 2006.

[118] E. H. Ong and J. Y. Khan. A unified QoS-inspired load optimization framework for multiple access points based wireless LANs. In *Proc. IEEE Wireless Communications and Networking Conference, 2009. WCNC 2009*, pages 1–6, April 2009.

[119] A. Kamerman and L. Monteban. Wavelan-II: A high-performance wireless lan for the unlicensed band. *Bell Labs Technical Journal*, 2(3):118–133, 1997.

[120] Y. Xiao and J. Rosdahl. Throughput and delay limits of IEEE 802.11. *IEEE Communications Letters*, 6(8):355–357, August 2002.

[121] FCC 05-57. Facilitating opportunities for flexible, efficient, and reliable spectrum use employing cognitive radio technologies. In *ET Docket No. 03-108*, March 2005.

[122] P. Bhagwat, P. Bhattacharya, A. Krishna, and S. K. Tripathi. Enhancing throughput over wireless LANs using channel state dependent packet scheduling. In *Proc. IEEE Fifteenth Annual Joint Conference of the IEEE Computer Societies. INFOCOM 1996*, volume 3, pages 1133–1140, March 1996.

[123] V. Bharghavan, S. Lu, and T. Nandagopal. Fair queuing in wireless networks: Issues and approaches. *IEEE Personal Communications*, 6(1):44–53, February 1999.

[124] A. Banchs and X. Perez. Distributed weighted fair queuing in 802.11 wireless LAN. In *Proc. IEEE International Conference on Communications, 2002. ICC '02*, volume 5, pages 3121–3127, May 2002.

[125] T. Joshi, A. Mukherjee, Younghwan Yoo, and D. P. Agrawal. Airtime fairness for IEEE 802.11 multirate networks. *IEEE Transactions on Mobile Computing*, 7(4):513–527, April 2008.

[126] E. H. Ong and J. Y. Khan. On load adaptation for multirate multi-AP multimedia WLAN-based cognitive networks. In *Proc. IFIP Wireless Days Conference, 2009. WD 2009*, pages 1–6, December 2009.

[127] K. Yang, I. Gondal, B. Qiu, and L. S. Dooley. Combined SINR based vertical handoff algorithm for next generation heterogeneous wireless networks. In *Proc. IEEE Global Telecommunications Conference, 2007. GLOBECOM '07*, pages 4483 – 4487, December 2007.

[128] Q. Pang, V. C. M. Leung, and S. C. Liew. An enhanced autorate algorithm for wireless local area networks employing loss differentiation. *IEEE Transactions on Vehicular Technology*, 57(1):521–531, January 2008.

[129] Y. Azar, A. Z. Broder, and A. R. Karlin. On-line load balancing. In *Proc. 33rd Annual IEEE Symposium on Foundations of Computer Science, 1992. FOCS 1992*, pages 218–225, October 1992.

[130] V. Gazis, N. Houssos, N. Alonistioti, and L. Merakos. On the complexity of "always best connected" in 4G mobile networks. In *Proc. IEEE 58th Vehicular Technology Conference, 2003. VTC 2003-Fall*, volume 4, pages 2312–2316, October 2003.

[131] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547, March 2000.

[132] P. Chatzimisios. *Performance Modelling and Enhancement of Wireless Communication Protocols*. PhD thesis, Bournemouth University, Poole, UK, December 2004.

[133] H. Zhai, Y. Kwon, and Y. Fang. Performance analysis of IEEE 802.11 MAC protocols in wireless LANs. *Wirel. Commun. Mob. Comput.*, 4(8):917–931, November 2004.

[134] Q. Ni, T. Li, T. Turletti, and Y. Xiao. Saturation throughput analysis of error-prone 802.11 wireless networks. *Wirel. Commun. Mob. Comput.*, 5(8):945–956, November 2005.

[135] E. Ziouva and T. Antonakopoulos. CSMA/CA performance under high traffic conditions: Throughput and delay analysis. *Comput. Commun.*, 25(3):313–321, February 2002.

[136] C. H. Foh and J. W. Tantra. Comments on IEEE 802.11 saturation throughput analysis with freezing of backoff counters. *IEEE Communications Letters*, 9(2):130–132, February 2005.

[137] H. Wu, Y. Peng, K. Long, S. Cheng, and J. Ma. Performance of reliable transport protocol over IEEE 802.11 wireless LAN: Analysis and enhancement. In *Proc. IEEE Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. INFOCOM 2002*, volume 2, pages 599–607, June 2002.

[138] P. Chatzimisios, A. C. Boucouvalas, and V. Vitsas. Influence of channel BER on IEEE 802.11 DCF. *Electronics Letters*, 39(23):1687–1689, November 2003.

[139] G. Bianchi and I. Tinnirello. Remarks on IEEE 802.11 DCF performance analysis. *IEEE Communications Letters*, 9(8):765–767, August 2005.

[140] Y. Xiao. Performance analysis of IEEE 802.11e EDCF under saturation condition. In *Proc. IEEE International Conference on Communications, 2004. ICC '04*, volume 1, pages 170–174, June 2004.

[141] M. Ergen. *I-WLAN: Intelligent Wireless Local Area Networking*. PhD thesis, University of California at Berkeley, Berkeley, CA, USA, Fall 2004.

[142] P. E. Engelstad and O. N. Osterbo. Analysis of the total delay of IEEE 802.11e EDCA and 802.11 DCF. In *Proc. IEEE International Conference on Communications, 2006. ICC '06*, volume 2, pages 552–559, June 2006.

[143] D. Yang, T. Lee, K. Jang, J. Chang, and S. Choi. Performance enhancement of multirate IEEE 802.11 WLANs with geographically scattered stations. *IEEE Transactions on Mobile Computing*, 5(7):906–919, July 2006.

[144] I. Tinnirello, G. Bianchi, and Y. Xiao. Refinements on IEEE 802.11 distributed coordination function modeling approaches. *IEEE Transactions on Vehicular Technology*, 59(3):1055–1067, March 2010.

[145] D. Malone, K.Duffy, and D. Leith. Modeling the 802.11 distributed coordination function in nonsaturated heterogeneous conditions. *IEEE/ACM Transactions on Networking*, 15(1):159–172, February 2007.

[146] G. R. Cantieni, Q. Ni, C. Barakat, and T. Turletti. Performance analysis under finite load and improvements for multirate 802.11. *Comput. Commun.*, 28(10):1095–1109, June 2005.

[147] N. T. Dao and R. A. Malaney. A new markov model for non-saturated 802.11 networks. In *Proc. 5th IEEE Consumer Communications and Networking Conference, 2008. CCNC 2008*, pages 420–424, January 2008.

[148] Q. Zhao, D. H. K. Tsang, and T. Sakurai. A simple model for nonsaturated IEEE 802.11 DCF networks. *IEEE Communications Letters*, 12(8):563–565, August 2008.

[149] K. Ghaboosi, M. Latva-aho, and Y. Xiao. A new approach on analysis of IEEE 802.11 DCF in non-saturated wireless networks. In *Proc. IEEE 67th Vehicular Technology Conference, 2008. VTC 2008-Spring*, pages 2345–2349, May 2008.

[150] R. P. Liu, G. Sutton, and I. B. Collings. A 3-D markov chain queueing model of IEEE 802.11 DCF with finite buffer and load. In *IEEE International Conference on Communications, 2009. ICC '09*, pages 1–5, June 2009.

[151] Y. C. Tay and K. C. Chua. A capacity analysis for the IEEE 802.11 MAC protocol. *Wirel. Netw.*, 7(2):159–171, March 2001.

[152] A. Kumar, E. Altman, D. Miorandi, and M. Goyal. New insights from a fixed point analysis of single cell IEEE 802.11 WLANs. In *Proc. IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies. INFOCOM 2005*, volume 3, pages 1550–1561, March 2005.

[153] O. Tickoo and B. Sikdar. A queueing model for finite load IEEE 802.11 random access MAC. In *Proc. IEEE International Conference on Communications, 2004. ICC '04*, volume 1, pages 175–179, June 2004.

[154] L. X. Cai, X. Shen, J. W. Mark, L. Cai, and Y. Xiao. Voice capacity analysis of WLAN with unbalanced traffic. *IEEE Transactions on Vehicular Technology*, 55(3):752–761, May 2006.

[155] K. Medepalli and F. A. Tobagi. System centric and user centric queueing models for IEEE 802.11 based wireless LANs. In *Proc. 2nd International Conference on Broadband Networks, 2005. BroadNets 2005*, pages 612–621, October 2005.

[156] J. V. Sudarev, L. B. White, and S. Perreau. Performance analysis of 802.11 CSMA/CA for infrastructure networks under finite load conditions. In *Proc. 14th IEEE Workshop on Local and Metropolitan Area Networks, 2005. LANMAN 2005*, pages 1–6, September 2005.

[157] A. Banchs, P. Serrano, and A. Azcorra. End-to-end delay analysis and admission control in 802.11 DCF WLANs. *Comput. Commun.*, 29(7):842–854, April 2006.

[158] H. M. K. Alazemi, A. Margolis, J. Choi, R. Vijaykumar, and S. Roy. Stochastic modelling and analysis of 802.11 DCF with heterogeneous non-saturated nodes. *Comput. Commun.*, 30(18):3652–3661, December 2007.

[159] G. Kuriakose, S. Harsha, A. Kumar, and V. Sharma. Analytical models for capacity estimation of IEEE 802.11 WLANs using DCF for internet applications. *Wirel. Netw.*, 15(2):259–277, February 2009.

[160] P. Raptis, V. Vitsas, P. Chatzimisios, and K. Paparrizos. Voice and data traffic analysis in IEEE 802.11 DCF infrastructure WLANs. In *Proc. Second International Conference on Advances in Mesh Networks, 2009. MESH 2009*, pages 37–42, June 2009.

[161] H. Heffes and D. Lucantoni. A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, 4(6):856–868, September 1986.

[162] L. Kleinrock. *Queueing Systems - Volume II: Computer Applications*. John Wiley & Sons, Inc., USA, 1976.

[163] G. Sharma, A. Ganesh, and P. Key. Performance analysis of contention based medium access control protocols. In *Proc. 25th IEEE International Conference on Computer Communications. INFOCOM 2006*, pages 1–12, April 2006.

[164] OPNET Technologies Inc. OPNET Modeler 14.5 PL8, http://www.opnet.com. Bethesda, MD, USA, August 2008.

[165] T. Rappaport. *Wireless Communications: Principles and Practice*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 2001.

[166] L. V. Fausett. *Numerical Methods: Algorithm and Applications*. Prentice Hall, Englewood Cliffs, NJ, 2003.

[167] H. Zhai and Y. Fang. Performance of wireless LANs based on IEEE 802.11 MAC protocols. In *Proc. 14th IEEE Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003*, volume 3, pages 2586–2590, September 2003.

[168] A. Zanella and F. De Pellegrini. Statistical characterization of the service time in saturated IEEE 802.11 networks. *IEEE Communications Letters*, 9(3):225–227, March 2005.

[169] M. J. Karam and F. A. Tobagi. Analysis of delay and delay jitter of voice traffic in the internet. *Comput. Netw.*, 40(6):711–726, December 2002.

[170] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory*. John Wiley & Sons, Inc., New York, NY, USA, 3rd edition, 1998.

[171] P. J. Denning. Virtual memory. *ACM Comput. Surv.*, 28(1):213–216, March 1996.

[172] M. A. Blumrich, C. Dubnicki, E. W. Felten, K. Li, and M. R. Mesarina. Virtual-memory-mapped network interfaces. *IEEE Micro*, 15(1):21–28, February 1995.

[173] R. Sharmila, M. V. Lakshmi Priya, and R. Parthasarathi. An active framework for a WLAN access point using Intel's IXP1200 network processor. In *Proc. High Performance Computing, HiPC 2004*, pages 71–80, December 2004.

[174] C. Heegard, J. T. Coffey, S. Gummadi, P. A. Murphy, R. Provencio, E. J. Rossin, S. Schrum, and M. B. Shoemake. High performance wireless ethernet. *IEEE Communications Magazine*, 39(11):64–73, November 2001.

[175] E. Akay and E. Ayanoglu. Low complexity decoding of bit-interleaved coded modulation for m-ary QAM. In *Proc. IEEE International Conference on Communications, 2004. ICC '04*, volume 2, pages 901–905, June 2004.

[176] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang. A measurement-based admission control algorithm for integrated service packet networks. *IEEE/ACM Transactions on Networking*, 5(1):56–70, February 1997.

[177] S. Shankar, J. del Prado Pavon, and P. Wienert. Optimal packing of VoIP calls in an IEEE 802.11 a/e WLAN in the presence of QoS constraints and channel errors. In *Proc. IEEE Global Telecommunications Conference, 2004. GLOBECOM '04*, volume 5, pages 2974–2980, November/December 2004.

[178] P. Magnusson, J. Lundsjo, J. Sachs, and P. Wallentin. Radio resource management distribution in a beyond 3G multi-radio access architecture. In *Proc. IEEE Global Telecommunications Conference, 2004. GLOBECOM '04*, volume 6, pages 3472–3477, November/December 2004.

[179] A. Tolli, P. Hakalin, and H. Holma. Performance evaluation of common radio resource management (CRRM). In *Proc. IEEE International Conference on Communications, 2002. ICC '02*, volume 5, pages 3429–3433, May 2002.

[180] O. Holland, M. Muck, P. Martigne, D. Bourse, P. Cordier, S. B. Jemaa, P. Houze, D. Grandblaise, C. Klock, T. Renk, J. Pan, P. Slanina, K. Mobner, L. Giupponi, J. P. Romero, R. Agusti, A. Attar, and A. H. Aghvami. Development of a radio enabler for reconfiguration management within the IEEE P1900.4 working group. In *Proc. 2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007*, pages 232–239, April 2007.

[181] E. H. Ong and J. Y. Khan. Distributed radio resource usage optimization of WLANs based on IEEE 1900.4 architecture. In *Proc. IFIP Wireless Days Conference, 2009. WD 2009*, pages 1–6, December 2009.

[182] E. H. Ong, J. Y. Khan, and K. Mahata. Comparative performance analysis of dynamic load distribution algorithms in a multi-AP wireless network. In *Proc. Annual IEEE India Conference, 2009. INDICON 2009*, pages 1–4, December 2009.

[183] E. H. Ong, J. Y. Khan, and K. Mahata. On dynamic load distribution algorithms for multi-AP WLAN under diverse conditions. In *Proc. IEEE Wireless Communications and Networking Conference, 2010. WCNC 2010*, pages 1–6, April 2010.

[184] B. O'Hara and A. Petrick. *The IEEE 802.11 Handbook: A Designer's Companion*. Standards Information Network, IEEE Press, New York, NY, USA, 1999.

[185] ITU-T P.59. Artificial conversational speech. March 1993.

[186] A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan. Characterizing user behavior and network performance in a public wireless LAN. *SIGMETRICS Perform. Eval. Rev.*, 30(1):195–205, June 2002.

[187] D. Kotz and K. Essien. Analysis of a campus-wide wireless network. *Wireless Networks*, 11(1-2):115–133, January 2005.

[188] D. Chiu and R. Jain. Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. *Comput. Netw. ISDN Syst.*, 17(1):1–14, June 1989.