Hybrid Radio Resource Management for Heterogeneous Wireless Access Network

Nagina Zarin

A Thesis
In the Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Electrical and Computer Engineering) at
Concordia University
Montreal, Quebec, Canada

October, 2021

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that thesis prepared

▷ By: Nagina ZARIN

▷ Entitled: Hybrid Radio Resource Management for Heterogeneous Wireless Access Network

and submitted in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY (Electrical and Computer Engineering)
complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Chun-Yi Su

_____External Examiner
Dr. Abraham Fapojuwo

_____ External to Program
Dr. Chadi Assi

_____ Examiner
Dr. Y. R. Shayan

_____Examiner
Dr. Dongyu Qiu

_____ Thesis Supervisor
Dr. Anjali Agarwal

Approved by: _____
Dr. R. Selmic, Graduate Program Director

October 1, 2021: _____
Dr. M. Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Hybrid Radio Resource Management for Heterogeneous Wireless Access Network

Nagina Zarin, PhD
Concordia University, 2021

Heterogeneous wireless access network (HWAN) is composed of fifth-generation (5G) and fourth-generation (4G) cellular systems and IEEE 802.11-based wireless local area networks (WLANs). These diverse and dense wireless networks have different data rates, coverage, capacity, cost, and QoS. Furthermore, user devices are multi-modal devices that allow users to connect to more than one network simultaneously. This thesis presents radio resource management for RAT selection, radio resource allocation, load balancing, congestion control mechanism, and user device (UD) energy management that can effectively utilize the available resources in the heterogeneous wireless networks and enhance the quality-of-service (QoS) and user quality-of-experience (QoE).

Recent studies on radio resource management in HWAN lead to two broad categories, 1) centralized architecture and 2) distributed model. In the centralized model, all the decision-making power confines to a centralized controller and user devices are assumed as passive transceivers. In contrast, user devices actively participate in radio resource management in the distributed model, resulting in poor resource utilization and maximum call blocking and call dropping probabilities.

In this thesis, we present a novel hybrid radio resource management model for HWAN that is composed of OFDMA based system and WLAN. In this model, both the centralized controller and the user device take part in resource management. Our hybrid mechanism considers attributes related to both user and network. However, these attributes are conflicting in nature. Moreover, a single RAT selection is performed based on user location and available networks, whereas UD with a multi-homing call receives the radio resource share from each network to fulfil its minimum data rate requirement. A novel approach is proposed for load balancing where an equal load ratio is maintained across all the available networks in HWAN. Performance evaluation through call blocking probability and network utilization will reveal the effectiveness of the proposed scheme.

The demand for more data rates is on the rise. The 5G heterogeneous wireless access network is a potential solution to tackle the high data rate demand. The 5G HWAN is composed of 5G new radio (NR) and 4G long-term evolution (LTE) base stations (BSs). In a practical system, the channel conditions fluctuate due to user mobility. We, therefore, investigate radio resource allocation and congestion control mechanism along with network-assisted distributive RAT selection in a time-varying 5G HWAN. This joint problem of radio resource allocation and congestion control management has signalling overhead and computational complexity limitations. Therefore, we use the Lyapunov optimization to convert the offline problem into an online optimization problem based on channel state information (CSI) and queue state information (QSI). The theoretical and simulation results evaluate the performance of our proposed approach under the assumption of network stability. In addition, simulation results are presented to depict our proposed scheme's effectiveness. Furthermore, our proposed RAT selection scheme performs better than the traditional centralized and distributive mechanisms.

Recently an increase in the usage of video applications has been observed. Therefore, we explore hybrid radio resource management video streaming over time-varying HWAN. Using the Lyapunov optimization technique, we decompose our two-time scale stochastic optimization problem into two main sub-problems. One of the sub-problems is related to radio resource allocation that operates at a scheduling time interval. The radio resource allocation policy is implemented at a centralized control node responsible for allocating radio resources from the available wireless networks using Lagrange dual method. The other sub-problem is related to the quality rate adaptation policy that works at a chunk time scale. Each user selects the appropriate quality level of the video chunks adaptively in a distributive way based on buffer state and channel state information. We analyze and compare the QoE of our proposed approach over an arbitrary sample path of channel state information with an optimal T-slot algorithm. Finally, we evaluate the performance analysis of our proposed scheme for video streaming over a time-varying heterogeneous wireless access network through simulation results.

Acknowledgements

First, I would like to express my sincerest gratitude to my supervisor, Professor Anjali Agarwal. I am thankful to her for giving me an opportunity to work in her research group. She helped me to lead a successful and enjoyable Ph.D. study. I would not be able to complete my Ph.D. without her immense support and knowledgeable guidance.

I would like to thank my thesis examination committee members Prof. Yousef R. Shayan, Prof. Dongyu Qiu, Prof. Chadi Assi and Prof. Abraham Fapojuwo. I appreciate their support and valuable comments and observations, which help me improve the quality of research work.

I would like to thank my parents and family for their love, massive support, and encouragement. Finally, I offer my special gratitude to my brother and mentor, Moh-u-din Bukhari, who encouraged and supported me during my Ph.D. program.

Contents

List of Figures

List of Tables

List of Abbreviations

AP	**A**ccess **P**oint
BS	**B**ase **S**tation
B5G	**B**eyonf **F**ifth Generation
CCN	**C**entral **C**ontroller **N**ode
CDF	**C**ommulative **D**istribution **F**unction
CSI	**C**hannel **S**tate **I**nformation
D2D	**D**evice to **D**evice
eMBB	**e**nhanced **M**obile **B**road**B**and
5G	**F**ifth **G**eneration
HWAN	**H**eterogenous **W**ireless **A**ccess **N**etwork
LB	**L**oad **B**alance
LTE	**L**ong **T**erm **E**velotion
ITU	**I**nternatioanl **T**elecommunication **U**nion
MEW	**M**ultiplicative **E**xponential **W**eighting
mMTC	**m**assive-**M**achine **T**ype **C**ommunication
OFDMA	**O**rthognal **F**requency **D**ivision **M**ultiple **A**ccess
QoE	**Q**uality **o**f **E**xperience
QoS	**Q**uality **o**f **S**ervice
QSI	**Q**ueue **S**tate **I**nformation
RAT	**R**adio **A**ccess Technology
RRM	**R**adio **R**esource **M**anagement
RSS	**R**eceived **S**ignal **S**trength
SAW	**S**imple **A**dditive **W**eighting
SNR	**S**ignal **N**oise **R**atio
TDMA	**T**ime **D**ivision **M**ultiple **A**ccess
3GPP	**3**rd **G**eneration **P**artnership **P**roject
UD	**U**ser **D**evice
URLLC	Ultra **R**eliable **L**o**w**-**L**atency **C**ommunication
WLAN	**W**ireless **L**ocal **A**rea **N**etwork

List of Symbols

B	Total bandwidth
W	RB bandwidth
$\mathcal{K}$	Set of users
N_o	Noise power spectral density
$\mathcal{L}$	Set of WLAN APs
M_l	Set of RBs at LTE BS
M_{5G}	Set of RBs at 5G NR BSs
$\rho_k^j(t)$	Association variable between RAT j and user k at time slot t
$x_{nk}(t)$	Allocation of RB n from LTE to user k at time slot t
$x_{mk}(t)$	Allocation of RB m from 5G NR BS i to user k at time slot t
$p_{nk}(t)$	Transmit power of RB n from LTE BS to user k at time slot t
$p_{imk}(t)$	Transmit power from RB m of 5G NR BS i to user k at time slot t
$P_{out}^w(t)$	Transmit power from WLAN access point at time slot t
$g_{nk}(t)$	Channel gain between user k and RB n from LTE BS at time slot t
$g_{imk}(t)$	Channel gain between user k and RB m from 5G NR BS i at time slot t
$g_{lk}(t)$	Channel gain between user k and AP l at time slot t
$r_{nk}(t)$	Transmit rate of user k from RB n of LTE BS at time slot t
$r_{mk}(t)$	Transmit rate of user k from RB m of 5G NR BSs at time slot t
$r_{lk}(t)$	Transmit rate of user k from WLAN AP l at time slot t
$R_k^w(t)$	Total transmit rate of user k from WLAN AP l at time slot t
$R_k^l(t)$	Transmit rate of user k from LTE BS at time slot t
$R_k^{5G}(t)$	Transmit rate of user k from 5G NR BSs at time slot t
$R_k(t)$	Total transmit rate of user k from LTE and 5G NR BSs at time slot t
$Q_k(t)$	Transmission queue of user k at time slot t
$D_k(t)$	Enqueue traffic rate of user k at time slot t
$B_k(t)$	Transmission queue of user k at time slot t
$T_{f,k}(t)$	Freezing time of user k at time slot t
$d_k(t)$	bitrate of the selected video chunk by user k at time slot t
C	Time duration of a video chunk
$E_{S_k}(t)$	Total power consumption of all active links of user device for user k at time slot t
$\gamma_k(t)$	Auxiliary variable indicating throughput of user k at time slot t
$\Gamma_k(t)$	Virtual queue for user k at time slot t

w_{nk}	Auxiliary variable for transmit power p_{nk} at time slot t
w_{imk}	Auxiliary variable for transmit power p_{imk} at time slot t
$A_k(t)$	Quality level of video chunks
$U_{S_k}(t)$	Video quality measure of the requested chunks
$b_k(t)$	Underrun indicator for each user$k$ at time slot t

Chapter 1

Introduction

The past few decades reveal tremendous growth in wireless and mobile communication. Statistics revealed by Cisco's virtual networking index forecast of internet users for the period 2018-2022 show a compound annual growth rate (CAGR) of 6% from 2018 to 2023. An 8% CAGR in the global mobile device and connection is expected for 2018-2023 [1], as shown in Fig. 1.1. Cellular mobile speed will be 43.9 Mbps in 2023, a three-fold increase from 2018. In contrast, the rate of 5G will reach 575 Mbps by 2023, and Wi-fi hotspots will grow four-fold from 2018 to 2023. Moreover, approximately 300 million applications will be globally downloaded by 2023 [1]. Future wireless communication is expected to support high data rates, global connectivity, high Quality of Service (QoS), and enhanced Quality of Experience (QoE) [2]. Heterogeneous wireless access networks (HWANs) can be a potential solution. In this chapter, we present HWAN, research motivation, and research contributions.

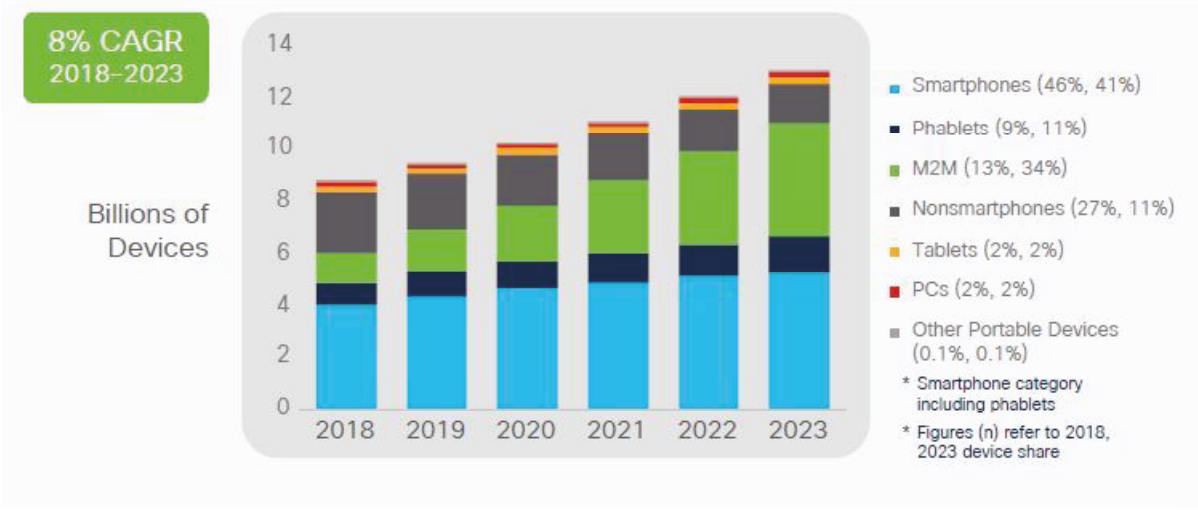

FIGURE 1.1: Global mobile device and connection growth [1]

1.1 The Heterogeneous Wireless Access Network

The vast deployment of wireless access technologies provides geographical locations to be covered by multiple wireless networks. The extensive implementation of wireless networks in a given area is related to the trends for free WLAN access, and as 4G, 5G and beyond 5G evolve, the number of legacy systems will grow. These different networks evolve as HWAN. HWAN can accommodate a massive amount of connections, provide a high data rate through multihoming connectivity, and maintain QoS and QoE requirements per user and application. HWAN is composed of 4G cellular networks (LTE/LTE-A), IEEE 802.11 Wireless Local Area Network (WLAN), IEEE 802.16 WiMAX, 5G, and beyond 5G (B5G) networks [2]. These different wireless access networks have their features, including coverage, QoS, capacity and cost. The WLANs, as well as 5G new radio (NR), provide high bandwidth in a limited coverage region, whereas both cellular and WiMAX provide broadband connectivity at a larger coverage zone.

HWAN consists of Base stations (BS), Access Points (AP), user devices (UDs), and IP backbone networks [3], as shown in Fig 1.2. HWAN has overlapping coverage zones and supports multi-RAT connections. Nowadays, UDs are provided with more than one radio interface. The multi-interface/multimodal UD can connect to any given RAT that satisfies its requirements. A communication network that allows the user device to transmit data and communicate over multiple radio access networks is termed as multi-radio access (MRA) system [4], or multihoming access system [5], [6]. Multi-RAT connectivity is subject to the number of available interfaces on the user device and the subscription of a user for network usage. HWAN, with overlapping coverage zones, provides a platform where UDs have "always best connection (ABC)" experience rather than "always connected experience" as in homogeneous access networks. HWAN supports bandwidth-hungry and high data rate applications via using features of multihoming by aggregating the allocated resources from different access networks.

1.2 Research Motivation

The diverse HWANs have different data rates, coverage, and capacity, which transform simple connectivity problems into a more challenging issue of radio resource management (RRM) [3]. Furthermore, the users in the coverage of HWAN with overlapping regions have a multi-RAT connectivity option since user devices are equipped with multi-radio interfaces. RRM is a mechanism required to efficiently utilize all the available resources and provide the required quality of service to users.

RRM involves network selection, radio resource allocation, radio resource utilization, congestion control, quality adaptation, energy management, mobility management and load balancing. It motivates us to propose a utility-based optimization technique for multi-RAT

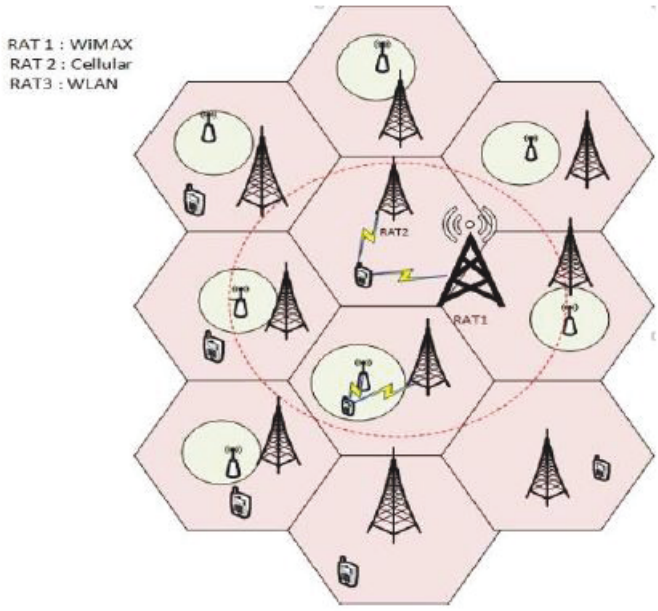

FIGURE 1.2: Heterogenous wireless access networks

allocation that takes attributes related to users and networks. Furthermore, it drives us to incorporate congestion control solutions with the resource allocation to provide maximum throughput to users while maintaining network stability. Furthermore, better video quality selection results in more power consumption. Therefore, it drives us to propose a solution for the energy efficiency of a multimodal user device to avail multi- homing facility during data and video streaming applications. Moreover, it strives us to investigate new radio resource management techniques in a time-varying condition that involves radio resources allocation, network congestion control, maintaining QoE of end-users, and managing user device energy consumption. The goal of RRM is to improve spectral efficiency, throughput, energy efficiency and overall performance of the network for multi-RAT connection and single connection in the HWAN. Therefore, we focus our research on exploring radio resource management schemes in HWAN that improve the spectral efficiency, network throughput and QoS, and user quality of experience (QoE).

1.3 Problem Statement

The literature survey shows that radio resource management in a heterogeneous network can be accomplished using either centralized management infrastructure or distributive management infrastructure. The centralized controller has a global view of the whole network. It has a single point of management and enhanced control over the entire network [7]-[9]. The centralized approach is more flexible and provides efficient resource utilization by balancing load across different RATs. However, in the centralized infrastructure, the central

controller communicates more frequently with all the network entities, results in congestion and may even increase the response time and delay. In addition to this, the centralized approach has the problem of single node failure causing convergence and instability problems in the network. It is assumed for the centralized approach that all the available networks belong to a single operator, which contradicts with reality where different RATs are run by different operators [10].

Another proposed solution for radio resource management is the distributed management infrastructure. Here the control is distributed among different network entities, including access routers [11], BS/AP and user device [10],[12]. However, in the distributed approach where the users perform network selection, the problem of load balancing arises as the users have a greedy approach without considering the actual load of the network. This user-centric approach results in poor resource utilization in coverage overlapping regions and increases call blocking probability and dropping probability. The network selection performed by a user does not guarantee a successful connection as the selected network may prefer to choose more valuable users. Cooperative distributed resource management has been proposed to efficiently improve the performance of radio resource management for heterogeneous wireless access networks [13]. However, this cooperation among networks and users increases signalling overheads.

Since the centralized and distributed management infrastructure has limitations, we propose a hybrid architecture, which takes the benefits of both centralized and distributed models. IEEE P1900.4 protocol-based hybrid approach is used for network selection based on low cost and mobility profile mobile users. They explored radio resource allocation for a single connection network using both best-effort service and differentiated traffic [35]. However, this approach is only limited to a single network connection. Another IEEE P1900.4 based hybrid RAT selection approach is proposed in [36], using IEEE P1900.4 protocol where the information about all available RATs and the decision of network selection is made. This approach considers mobility. The performance of the proposed method is compared with centralized and distributive mechanisms. Their proposed work gives better call blocking probability, vertical handoff probability and better user satisfaction. However, this technique is only limited to a single network connection. Finally, the authors in [37] proposed a hybrid network selection approach where the user device takes the final decision with network assistance. They used two mechanisms of the staircase and slope tuning technique, where it dynamically modulates the information broadcasted by a network (service cost and QoS information) and user preferences. The simulation results show enhanced network performance, higher network gain and better user QoE. However, this model is limited to a single network connection that takes consideration for elastic, inelastic traffic and streaming sessions.

We propose a radio resource management scheme that considers the attributes of both network and UD. Our proposed architecture has a central node termed central control node

(CCN) and distributed entities such as BS/AP and UDs. The central controller node works at a large time scale and manages, monitors, and associates UDs to the best connection. In contrast, the user devices work quickly, monitor changes in connection status, and update the BS about the channel state information. This hybrid architecture can manage resource allocation, network selection, load balancing and energy management of mobile users.

We consider a HWAN network layout that is composed of LTE and WLAN. The mathematical techniques used for network selection are utility theory, MADM [18], fuzzy logic [19], game theory [21], combinatorial optimization [24], and Markov chain [27]. All these techniques are used for network selection; however, we selected utility theory as our RAT selection model. This mathematical tool has simple implementation complexity, and the decision-making speed is fast. Moreover, it is suitable for our hybrid model as it takes multiple attributes from both network and user and evaluates its utility function. Utility functions translate network resources to user satisfaction, one of the key factors in 5G and future wireless communication networks. Resource allocation, performed by the central controller node, takes input from both network and users, ranks different RATs, and associates users to different RATs. The multi-RAT heterogeneous access network is composed of overlapping coverage zones. The overlaid smaller cells within the coverage of macrocells experience the issue of imbalance in load. The load balancing is required to balance the network load, increase network performance and enhance resource utilization. The authors in [66]- [68] explored an improvement in system performance via load balancing in the heterogeneous networks. In [69], the authors present two joint resource management schemes for determining system performance in a multi-RAT heterogeneous environment. We propose that CCN is equipped with a load balancing and congestion control mechanism in a heterogeneous environment. The load balancing is achieved by increasing or decreasing the number of associated users at different RATs per overload and underload conditions.

Considering a time-varying HWAN is a step towards a more realistic approach. The multi-modal devices equipped with multi-radio access or multihoming facility is an expected dominating connection option in 5G networks. Therefore, we consider hybrid RRM for time-varying 5G HWAN. 5G HWAN is composed of 4G LTE and 5G new radio (NR) BS. In 5G HWAN, the LTE BS act as an umbrella, where 5G NR BS act as hot spots. The LTE macro-BS provides maximum coverage to the users, whereas the 5G NR BSs provide maximum data rate to the users. In this hybrid RRM, we propose a network selection scheme, radio resource allocation policy and rate adaptation policy. We further investigate hybrid RRM for video streaming over a time-varying HWAN. Here we consider the QoE parameters related to video quality, freezing time and energy consumption of the UD. Radio resources are allocated, and video quality is selected without degrading the QoE of end-users.

1.4 Research Contributions

Our research focuses on a hybrid radio resource management in HWAN. Specially, we propose hybrid schemes related to the process of network selection, quality of service (QoS) based radio resource allocation, and congestion control in HWAN. The process of multi-criteria-based network selection helps in selecting the best network among the available networks. We propose optimal radio resource allocation, i.e. subcarrier and power allocation, that satisfies users QoS requirements and efficiently utilizes the available resources of different networks in the HWAN. Furthermore, we propose radio resource allocation and rate adaptation in a time-varying 5G HWAN that optimally allocates the radio resources and avoids network congestion, thereby maintain the QoS of mobile users. Moreover, we propose radio resource management for streaming sessions over a time-varying HWAN. The following discusses our research contributions.

- We propose a hybrid multicriteria-based RAT selection that takes attributes from both networks and users [14]. Moreover, both CCN and user devices make decisions about the appropriate RAT selection using a multiplicative exponential weighting (MEW) method. This selection can be single-homed or multi-homed depending on the location of users and the decision based on multi-attributes related to users and networks.

- We explore the QoS-based radio resource allocation in HWAN for multihoming calls [15]. Radio resource allocation is subject to the minimum data rate requirements of mobile users, thereby satisfying the QoS requirements of each user. As a result, we obtain an optimal solution for the allocation of power and subcarrier from the OFDM-based system and timeshare from WLAN.

- We explore load balancing in HWAN [16]. The performance of the propose scheme is evaluated in a cellular layout with overlapping coverage zones. Call blocking probability, and bandwidth utilization show the effectiveness of our propose approach.

- We investigate a hybrid congestion control and radio resource allocation algorithm in 5G HWAN for time-varying channels. The CCN utilizes the radio resource allocation policy to allocate the radio resources, i.e. resource blocks, and transmit power to the mobile users. In contrast, each user performs the congestion control policy in a distributive manner [17].

- We propose a hybrid radio resource management over a time-varying HWAN for a streaming session. The radio resources allocation is subject to the QoE of end-users.

1.5 Thesis Outline

We organize the thesis as follows. In Chapter 2, a novel hybrid approach for RAT selection in HWAN is investigated [14]. It is a multicriteria-based RAT selection and takes attributes from both networks and users. Moreover, both CCN and user devices are involved in deciding the appropriate RAT selection using a multiplicative exponential weighting (MEW) method. This selection can be single-homed or multi-homed depending on the location of users and the decision based on multi-attributes related to users and networks. The MEW method with the multicriteria is compared with the simple additive weighting (SAW) utility function approach in terms of accuracy. Finally, we compare the performance of our proposed plan with the conventional methods, i.e. centralized and distributive methods of RAT selection.

Chapter 3 explores the QoS-based radio resource allocation in HWAN [15]. We consider users with multihoming calls. We obtain an optimal solution for allocating power and subcarrier from the OFDM-based system and timeshare from WLAN. Load balancing in HWAN is proposed in Chapter 4. Our proposed load balancing scheme balances both the overload and underload conditions. The performance of the proposed approach is evaluated through call blocking probability and bandwidth utilization. Radio resource allocation and rate adaptation for time-varying 5G HWAN are investigated in Chapter 5. We consider cellular networks, i.e. LTE and 5G NR BSs, in our proposed system model and implement a novel hybrid congestion control and radio resource allocation algorithm. The resource allocation policy implemented at CCN allocates the radio resources, i.e. resource blocks, and transmit power to the mobile users. In contrast, the congestion control policy is performed at the user end in a distributive manner. Moreover, we propose a network-assisted RAT selection scheme, which helps in selecting the appropriate RAT based on the utilities of received signal strength and QoS [17].

We propose a hybrid radio resource management over a time-varying HWAN for video streaming sessions in Chapter 6. We consider video quality, freezing time and user device battery as QoE attributes. Optimal quality selection is proposed without compromising the QoE of end-users. The conclusion and future research work of our research are presented in Chapter 7.

Chapter 2

Hybrid Network Selection Scheme in HWAN

Heterogeneous wireless access network (HWAN), an integration of different RATs in an overlapping zone, supports bandwidth-hungry applications and fulfills high data rates' demands. One of the main challenges in HWAN is the selection of an appropriate RAT (single connection) or multi-RATs (multi-homing connection) depending on RATs availability and user requirement. There can be two possible solutions, 1) centralized approach and 2) distributive mechanism. However, these solutions have some serious limitations, which are explained in detail in this chapter. This chapter presents a novel hybrid scheme for RAT selection in HWAN, a two-step process in which both a CCN and UD are involved in network selection. Our key objective is to explore the role of each entity (CCN and UD) in the process of RAT selection. Furthermore, we consider multi-attributes related to both user and network. Therefore, it is important to explore the impact of different crucial criteria on RATs ranking results. The other main objective is to compare the precision of our proposed hybrid approach with traditional mechanisms. UD screens the available list of scanned networks based on the received signal strength and user mobility profile in our proposed approach. Next, we compare the RAT screening results using a multiplicative exponential weighting method (MEW) with a multi-criteria simple additive weighting (SAW) utility function. Finally, the CCN takes multi-criteria related to the application, UD, and network, generating a sorted list of the most appropriate RATs based on evaluating multiplicative exponential weighted utility function. The CCN then associates users to one (single connection) or more available RATs (multi-homed). RATs ranking and association are elaborated by calculating different networks final utilities.

2.1 Background and Introduction

The HWAN provides multi-RAT connectivity options due to the overlapping coverage zone. The selection of an appropriate RAT, an integral part of radio resource management, has become a challenging issue in HWAN. The literature survey shows that RAT selection can

be categorized as 1) single RAT selection, and 2) multi-RAT selection as nowadays, the user devices have multi-radio interfaces, and they can connect to more than one RAT simultaneously. The literature survey further shows that RAT selection can be performed using conventional methods, i.e. centralized method of RAT selection, distributive method of RAT selection, and collaborative method of RAT selection. These different network selection strategies are based on different mathematical models and theories. These models and theories include utility theory, multi-attribute decision making (MADM) [18], fuzzy logic [19], [20], game theory [21]-[23], combinatorial optimization [24]-[26], and Markov Chain [27], [28]. In general, the RAT selection in all these approaches is based on some pre-defined criteria. These criteria are related either to the user, network, or both [29]. User-related attributes are throughput, required Quality-of-Experience (QoE), and battery consumption, whereas network-related features include balance in load, revenue, service differentiation and network throughput. Received signal strength (RSS) is one of the most important criteria used for selecting the BS or AP of the available network [30]. Offered bandwidth is another important attribute used to associate users with the best network [31], [32]. The authors in [33], [34] proposed a multi-attribute RAT selection mechanism based on the utilities of cost, bandwidth and received signal strength to select the BS/AP of the most suitable RATs among the available networks. The available networks may not have enough resources to accommodate the incoming calls. Therefore, using multi-homing services allows an incoming user with high bandwidth demand to connect simultaneously to more than one network. The multi-homing facility aggregates the bandwidth from different networks to fulfil the high bandwidth user's application demand. Multi-homing provides global connectivity to mobile users. It has low call blocking probability and high system capacity.

Centralized and distributive mechanisms have both advantages and disadvantages. RAT selection's centralized approach balances the load across different RATs, thereby efficiently utilizing the available resources of different RATs [7], [9]. However, the centralized framework has a higher delay and response time due to congestion caused by increased signalling and communication among network elements. In the distributed management, the process of radio network selection is implemented at different network entities, including access routers [10], base station/access point (BS/AP) and user devices (UDs) [11], [12]. However, the distributed approach may lead to an imbalance load across HWAN as the mobile users greedily select the network without any prior knowledge of its load. Furthermore, the chosen network may not be able to accommodate. Thus, an increase in call blocking and dropping can be observed. The network selection may not guarantee a successful connection as the selected network may prefer to choose more valuable users.

In the literature survey, minimal work is related to the hybrid method of radio resource management. The hybrid approach using the cost function proposed in [35], and based on the IEEE P1900.4 protocol, presents network selection based on the low cost and mobility profile of mobile users, where resource allocation for a single connection network using both

best-effort service and differentiated traffic is proposed. This approach is only limited to a single network connection. Another IEEE P1900.4 based hybrid RAT selection approach is proposed in [36], where the decision of network selection is made by gathering information about all available RATs using the IEEE P1900.4 protocol. This approach considers mobility. The performance of the proposed method is compared with centralized and distributive mechanisms. Their proposed work gives better call blocking probability, vertical handoff probability and better user satisfaction. However, this technique is only limited to a single network connection. The authors in [37] propose a hybrid network selection approach where the user device takes the final decision with network assistance. The obtained simulation results reveal improved network performance, higher network gain and better user satisfaction. However, these proposed hybrid schemes are user-centric that takes network assistance, i.e., extending the existing approach. The user-centric network-assisted plan assumes that the networks broadcast their loading information to the users. However, this seems impractical as the network always has an intention to serve users. It is therefore required to mask the loading condition of a network from users.

To cope with the challenges in the centralized and distributive approach, we propose a hybrid multi-attribute network selection scheme, which takes input related to network and use, and distributes the decision-making process between user and CCN. Utility theory is one of the most popular strategies for network selection [38]. It is based on decision-making attributes related to the user, mobile device, application, and network. These attributes can be positive/negative, described with linear, logarithmic, exponential, and sigmoidal functions. Therefore, it is crucial to identify each related attribute's type and choose an appropriate utility function for each attribute. SAW, a method used to determine the total utility of a RAT, is the most popular network selection method. However, it has some limitations, and the authors in [39] suggested solutions in multiplicative exponential weighting (MEW). Therefore, we use the MEW method for RATs ranking and association. The obtained results highlight the comparison of both these schemes.

The rest of the chapter is organized as follows: Section 2.2 describes our proposed system model, which includes hybrid RAT selection architecture and proposed algorithm. Then, Section 2.3 discusses the simulation results. Finally, Section 2.4 summarizes the chapter.

2.2 System Model

This section presents our proposed system architecture in terms of wireless networks, UDs, and CCN. The detailed description is given as follows.

2.2.1 System Architecture

We consider two controlling entities in our proposed hybrid network selection scheme in HWAN: 1) CCN, and 2) controller at UD. Figure 2.1 depicts our proposed hybrid network selection architecture.

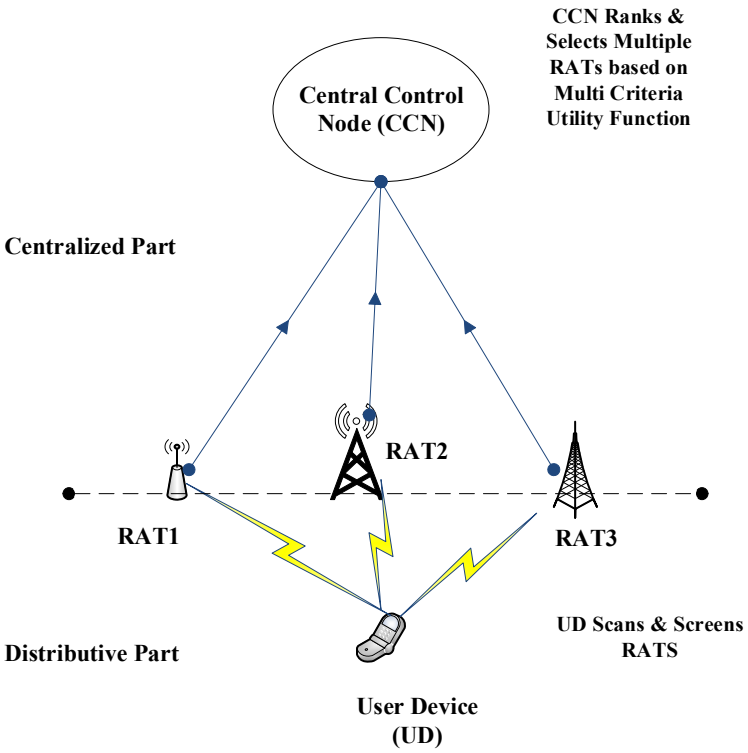

FIGURE 2.1: Proposed Hybrid Architecture for Network Selection

From the literature, we explored that decision of RAT selection is either made at the user end or at the centralized controller. However, our proposed algorithm distributes the process of RAT selection between the user device and CCN, which makes it novel from the traditional approaches of RAT selection. Furthermore, our multi-criteria-based RAT selection considers attributes related to network (load ratio), the user device (battery consumption), application requirements (throughput), and user (received signal strength and mobility). UDs are multimodal, and they can access a single network or multi-RATs simultaneously. CCN, located at the backend, performs access selection. The CCN takes input from both network and user and associates users to different RATs. UD communicates with CCN via different RATs based on the channel status. The operational process of network selection begins at a user device. UD first scans for the available networks in its vicinity. UD measures the received signal strength from the list of available networks in its service area. It screens the available set of scanned networks based on the received signal strength and user mobility. Mobile user with high speed removes the networks with low coverage from the list of scanned networks

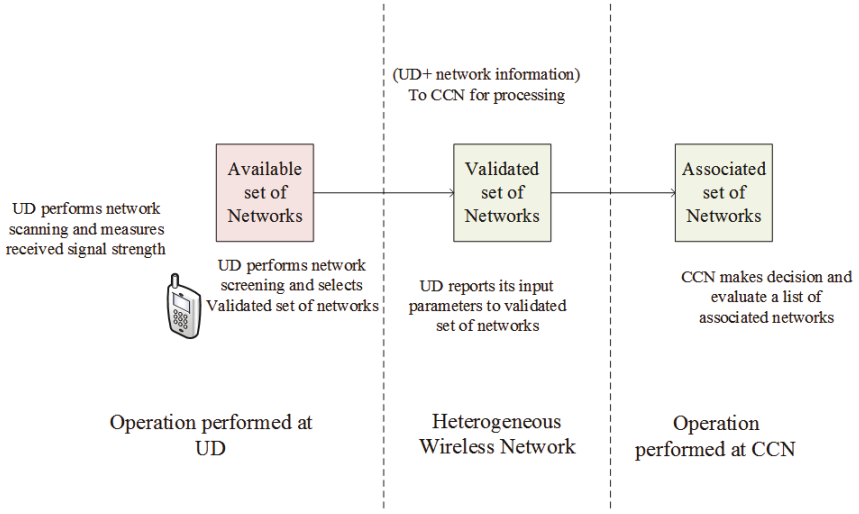

FIGURE 2.2: Network selection mechanism

to avoid frequent handoff. If the received signal strength is higher than a threshold such that $RSS_i \geq RSS_{Threshold}$, UD uses its multi-interface facility, reports its input parameter, i.e. application BW, user mobility and available resources (battery constraint) to the BS/AP of the available set of networks. We have assumed that the signalling information exchange among users and BS/AP takes place on time. Furthermore, we assumed perfectly accurate signalling information exchange and no delivery delay. The BS/APs of these available RATs add their own information with the user message and forward it to the CCN for further processing. Moreover, we assume no signalling delay between BS/AP and CCN. Available networks are ranked in descending order based on total utility calculation. The total utility function is based on a multi-criteria MEW utility function. CCN selects networks with high utility values. Thus, the CCN ranks the networks and provides an associated set of networks. Finally, CNN directs the users to associate with different RATs, either single RAT or multi-RATs. Figure 2.2 highlights the proposed network selection mechanism.

2.2.2 Proposed Algorithm

The proposed algorithm is based on the multi-criteria multiplicative utility function. The process of network selection begins at a user device. The involvement of UD in the process of RAT selection speeds up the decision-making process and reduces the burden on the central controller. During the first step of RATs screening, user utility function[38], [39] for different RATs based on received signal strength (RSS) and mobility is given by Eq. (2.1)

$$U_u^n(x) = (U_{RSS})^{w_{RSS}} * (U_m)^{w_m}, \quad w_{RSS} + w_m = 1 \tag{2.1}$$

where n corresponds to different available RATs, U_u is the user utility function and U_{RSS} and U_m are utility functions for RSS and user mobility, respectively. w_{RSS} and w_m are weights of

these selected criteria. Networks with lower utility values are ignored from the list of available RATs. In our second step of RATs ranking, the rest of the parameters related to user (monetary cost), device (battery consumption), and application (required BW and throughput) are forwarded to the CCN via a validated set of networks. The CCN makes the final decision of RAT selection by evaluating the total utility function for each network.

$$U_u^{vn}(x) = (U_c)^{w_c} * (U_e)^{w_e} * (U_\gamma)^{w_\gamma} * (U_L)^{w_L}, \tag{2.2}$$

where $U_u^{vn}(x)$ is the total utility of validated RAT vn for user u, U_c , U_e ,U_γ, and U_L corresponds to the utility function for monetary cost, battery consumption, throughput and network load, respectively. The sum of all weights w_c, w_e, w_γ, and w_L, is equal to 1. Based on Eq. 2.2, the CCN sorts networks in descending order and associates users to single or multi-RATs that best suits its requirements.

2.2.2.1 Utility function for RSS

The received signal strength P_r is a positive attribute and evaluated using the following equation

$$P_r = \frac{P_t * G_t * G_r}{P_L} \tag{2.3}$$

where P_t, G_t and G_r are the transmitted power, transmitting antenna gain and received antenna gain, respectively. The value of path loss P_L for the cellular network is calculated using Cost-231 Hatta Extended Model, whereas, for WLAN, a two ray path loss model is employed. The utility function for RSS is calculated as a linear function given by [38].

$$U_{RSS} = \begin{cases} 0, & \text{if } P \leq P_{Thresh} \\ \frac{P_r - P_{thresh}}{P_{max} - P_{thresh}} & \text{if } P_{thresh} < P \leq P_{max} \\ 1, & \text{if } P > P_{max} \end{cases} \tag{2.4}$$

where P_{thresh} and P_{max} are the threshold power and maximum power, respectively.

2.2.2.2 Utility function for Mobility

User mobility is related to two attributes, i.e. speed of the mobile user and range R of an AP/BS, which reflects the expected time of residence of a mobile user within the coverage of an AP/BS [40]. It is desirable to prevent high-speed mobile users from connecting to short-range AP/BS to avoid a high risk of interruption in the future. Users are classified into three main categories in terms of mobility, i.e. static users, pedestrians (moving with moderate

speed) and mobile users (moving with high speed). Pedestrian users of the WLAN network have a utility of 0.5. The utility for mobility is given by

$$U_m = \begin{cases} 0, & \text{if high speed users in WLAN coverage with } R > R_{max} \\ 0.5 & \text{if moderate speed users in WLAN coverage with } R \leq R_{max} \\ 1, & \text{for Cellular Network and static users} \end{cases} \tag{2.5}$$

2.2.2.3 Utility function for Cost

The monetary cost C, a negative attribute, can be given by the following utility function [41]

$$U_c = \begin{cases} 1, & \text{if } C \leq C_{min} \\ 1 - \frac{C - C_{min}}{C_{max} - C_{min}} & \text{if } C_{min} < C < C_{max} \\ 0, & \text{if } C > C_{max} \end{cases} \tag{2.6}$$

2.2.2.4 Utility function for Throughput

Using Shannon theory the maximum throughput achieved by user i from AP/BS j for OFDM based cellular network is given by [42]

$$\gamma_{ij} = B_j log_2(1 + SINR_{ij}) \tag{2.7}$$

where the bandwidth B_j is in Hz and the signal-to-interference and noise ratio $SINR_{ij}$ between user i and BS j is given by

$$SINR_{ij} = \frac{p_{ij}^r}{I + N_0} \tag{2.8}$$

where p_{ij} is the received signal power. the noise power present at the UD is given by N_0, whereas I corresponds to the interference power and is given by

$$I = I_{inter} + I_{intra} \tag{2.9}$$

where I_{inter} is the inter-cell interference, and I_{intra} is the intra-cell interference power. It is assumed that for cellular network $I_{intra} = 0$. Based on frequency re-use, the total interference power is given by

$$I = I_{inter} = \sum_{(k \neq j)} P_{(i,k)}, \tag{2.10}$$

where $P_{(i,k)}$is the power received at user i from interfering nodes k. The WLAN network is throughput fair network as CSMA/CA assures the same connection probability for users accessing the same AP [43]. However, users connected to different AP have different throughput [44]. The throughput of user i from Wi-Fi AP j is given by

$$\gamma_{ij} = \frac{L_{packet}}{\sum_{k \in N_j} \frac{L_{packet}}{R_{k,j}}} \tag{2.11}$$

where L_{packet} is the length of the packet, N_j corresponds to the number of users associated with the AP j, and $R_{(k,j)}$ is the rate of user k connected to AP j. The utility function for throughput is given by [45]

$$U_\gamma = \begin{cases} 0, & \text{if } \gamma \leq \gamma_{thresh} \\ e^{-\partial(\gamma - \gamma_{thresh})} & \text{if } \gamma_{thresh} < \gamma \leq \gamma_{max} \\ 1, & \text{if } \gamma > \gamma_{max} \end{cases} \tag{2.12}$$

where ∂ is the shape parameter, higher values of ∂ makes the graph steeper.

2.2.2.5 Utility function for Battery Consumption

Battery energy constraint is a negative attribute. It is desirable to associate users with a RAT with minimum energy consumption. If the energy consumption of a device is lower, the better is the utility. So the utility function for energy consumption is given by

$$U_e = \begin{cases} 1, & \text{if } E \leq E_{min} \\ 1 - \frac{E - E_{min}}{E_{max} - E_{min}} & \text{if } E_{min} < E \leq E_{max} \\ 0, & \text{if } E > E_{max} \end{cases} \tag{2.13}$$

The energy consumption for all active interfaces of a user is given as [46]:

$$E = \sum_{i=1}^{n} (\gamma_{req} * P_{ti}), \;\; i = 1, ...n \tag{2.14}$$

where E is the total energy consumption measured in joules, $\gamma_{req}$is the required throughput in kbps, which depicts data transmission by an interface, and P_{ti} describes power consumption of an interface.

2.2.2.6 Utility function for Network Load

Network load L is a crucial attribute for selecting an appropriate network as it reflects the amount of available bandwidth in a network. The utility function of load is given as

$$U_L = f(L) \tag{2.15}$$

$$L = \frac{B_{io}}{B_{max}} \tag{2.16}$$

B_{io} is the number of channels in use in the network, and B_{max} is the maximum number of channels in the network. L is the load ratio of the network under consideration. Transforming the load ratio L into utility, we consider two values of threshold, i.e. the upper bound $(L_{thresh2})$ and lower bound $(L_{thresh1})$. If a network load is less than the lower threshold value, then $U_L = 1$, whereas for a network load higher than the upper threshold value, the utility of load is equal to 0. The utility of load is given by

$$U_L = \begin{cases} 1, & \text{if } L < L_{thresh1} \\ 1 - e^{-\partial * A}, & L_{thresh1} \leq L \leq L_{thresh2} \\ 0, & \text{if } L > L_{thresh2} \end{cases} \tag{2.17}$$

$$A = (L - L_{thresh1}) \tag{2.18}$$

where $A = (L - L_{thresh1})$ and ∂ is the shape parameter, and the graph becomes steeper with higher values of ∂.

We approximated received signal strength, cost, and battery consumption using linear utility functions in our mathematical model. In contrast, throughput from a RAT is described by using "increasing marginal utility" and RAT load ratio as "decreasing marginal utility" as shown in Figure 2.3. The graph for load ratio shows that after the threshold of 0.8, the utility of network load becomes zero. It should be noted that users cannot differentiate between the services from networks with $L < 0.5$. The controller approximates the utility value for the load ratio of less than 0.5 as 1. The utility of positive attributes like throughput is the reverse replica of load ratio [45]. The utility of cost is one, i.e. $U_c = 1$ if the network services are available for free. However, the utility of cost is zero if the network monetary cost is higher than the user is willing to pay, as shown in Figure 2.3.

Once the utility functions of these criteria are calculated, it is then required to assign an appropriate weight to the individual criterion. Weights of these different criteria act as "tuning knobs" in the process of RAT selection. Users and service providers can set weights per their preferences. However, before assigning weights to different criteria, it should be noted that each criterion has its importance. The more critical a criterion, the higher the value of the assigned weight. UD considers the attributes of mobility and received signal strength

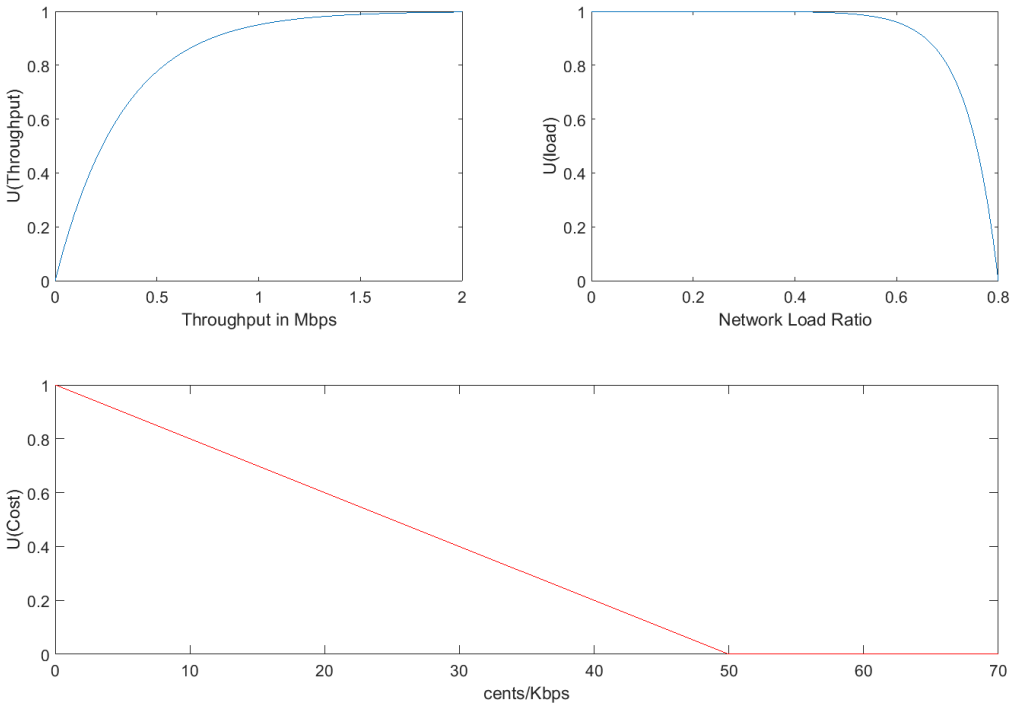

FIGURE 2.3: Utility Function of Throughput, Cost, and Network Load

and their related weights, as given in Eq. 2.1. We use $w_{RSS} = 0.55$ and $w_m = 0.45$, as the sum of their weights, $w_{RSS} + w_m = 1$. The CCN considers four attributes, i.e. throughput, network load, battery consumption, and cost, to make the final decision of RAT selection based on Eq. 2.2. The recent literature survey shows that the mobile user is more concerned about battery life, and we assigned $w_e = 0.3$. Similarly, the network performance is directly related to the network loading condition, and we assigned $w_L = 0.3$. Since we assumed that our users are less concerned about the service cost and throughput requirements, we assigned both w_c and w_γ a weight of 0.2. The sum of these weights $w_e + w_L + w_c + w_\gamma = 1$. The weights of these different criteria are shown in Table 2.2.

2.3 Simulation Results and Discussion

This section presents the results of our proposed hybrid network selection scheme. We consider four RATs located at different positions in our simulation area of 2000 square meters, shown in Figure 2.4. We randomly distribute the users in this area. The simulation area has overlapping zones where users receive services from more than one RAT. The characteristics and features of these different RATs are given in Table 2.1. MATLAB-based simulation results are obtained using four RATs with six criteria. Table 2.2 shows the characteristics of all

TABLE 2.1: RATs Main Features [45]

Features	WLAN1	WLAN2	LTE	WiMax
RSS Threshold	-60dBm	-55dBm	-90dBm	-130dBm
Coverage zone	100-500	50-100	500-1000	500-1000
PL Model	Two-ray Model	Two -ray Model	Cost-321 Hata Model	Cost-321 Hata Model
Fc	2.4GHz	2.4GHz	2.1GHz	2.3
User mobility support	Pedestrian (5m/hr)	Static -	Mobile (10km/hr)	Mobile (10km/hr)
Max Throughput	0.1-2Mbps	1-6Mbps	70Mbps	54Mbps
Cost(cent/kbps)	2	Free	6	8
Load	0-90%	0-90%	0-90%	0-90%

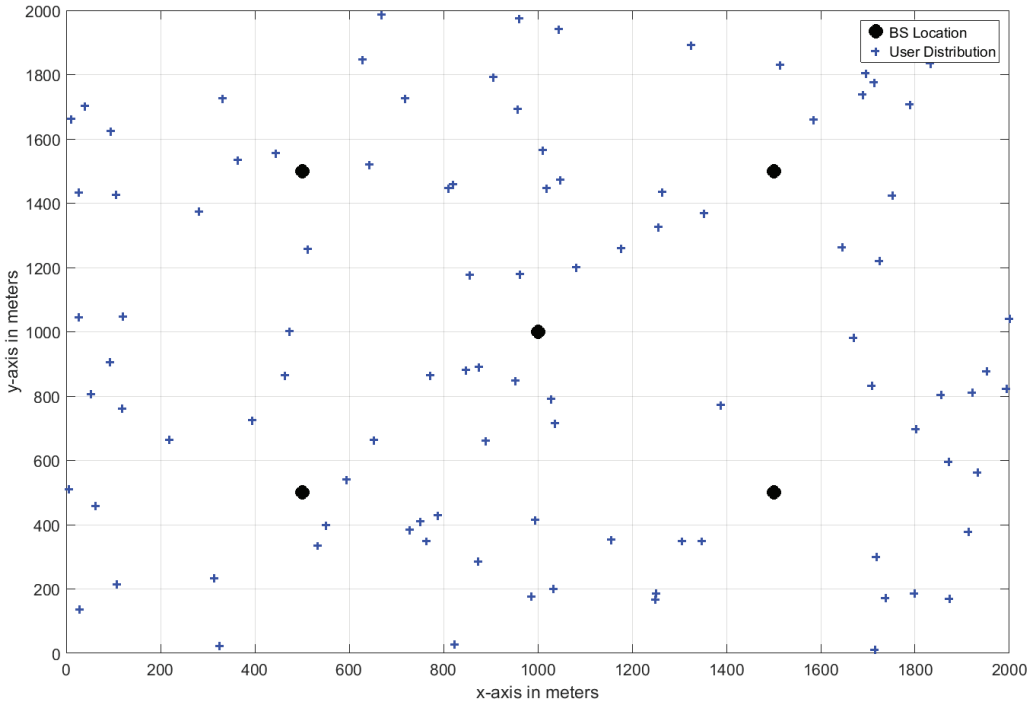

FIGURE 2.4: Simulation Topology

the six attributes.

The process of RAT ranking, selection and user association is a two-step process. Step 1 of the proposed hybrid multiplicative multi-criteria exponential weighted (MEW) utility-based network selection approach is compared with a simple additive weighting (SAW) method,

TABLE 2.2: Simulation Parameters (Attributes Vs RAT)

RAT	Cost	Speed	Application Req_Throughput	Battery Consumption	RSS	Load (L)
WLAN1	2	5m/hr	2Mbps	0.007(x)	-45	0-0.8%
WLAN2	0	-	-	0.006(x)	-45	0-0.8%
LTE	4-8	-	-	0.018(x)	-55	0-0.8%
WiMax	4-8	-	-	0.018(x)	-55	0-0.8%
Weights	0.2	0.45	0.2	0.3	0.55	0.3

whereas in step 2, simulation results show the list of ranked RATs. Our approach is unique such that the process of RAT's ranking and selection begins at the user device. At step 1, it is assumed that at time interval t=0, the randomly distributed mobile users start moving in the area. The user device scans the available networks' service area and calculates their utilities based on the received signal strength and user mobility. Figure 2.5 shows network ranking based on these two criteria. It is assumed that the user is mobile with moderate speed. The proposed approach screens and removes WLAN2 from the list based on user speed and RAT coverage. From Figure 2.5, we can see that with the MEW approach, WLAN2 has zero utility value, whereas the SAW method still gives a high utility for WLAN2. However, it is not a suitable RAT for a user moving at moderate speed. The user device forwards its input parameters only to the validated set of networks.

At step 2, the BS/APs of validated RATs send a processed message to the central controller node. Figure 2.6 shows the evaluation of the sorted set of RATs based on the final decision made by CCN. In step 2, we ranked the RATs by considering two scenarios, i.e. 1) overall utilities and RATs ranking without energy attribute, and 2) overall utilities and RATs ranking with energy attribute. It can be seen from Figure 2.6 that ignoring one attribute generates different utility values for RATs under consideration. Without considering the battery consumption metric, users can be associated with all the RATs available in its vicinity, as the final utility values of these RATs are higher than those of 0.5. However, considering the mobile user's energy consumption, only WLAN1 fulfills the nomadic user's requirements.

Figure 2.7 shows the comparison of our proposed hybrid RATs ranking and network selection with the existing schemes. Figure 2.7[a] reveals that the decision of RATs ranking of our proposed hybrid scheme is more precise than the traditional centralized and distributive mechanisms. The centralized mechanism provides WLAN2 in the list of ranked RATs as it does not consider user mobility. This scheme provides WiMax and LTE as the most appropriate RATs for connectivity, based on loading information and QoS requirements. This RATs ranking decision does not include user-related attributes. The distributive mechanism and our proposed approach give WLAN1 the most appropriate RAT for connectivity. However, our RATs ranking decision is more precise than the distributive mechanism as it considers

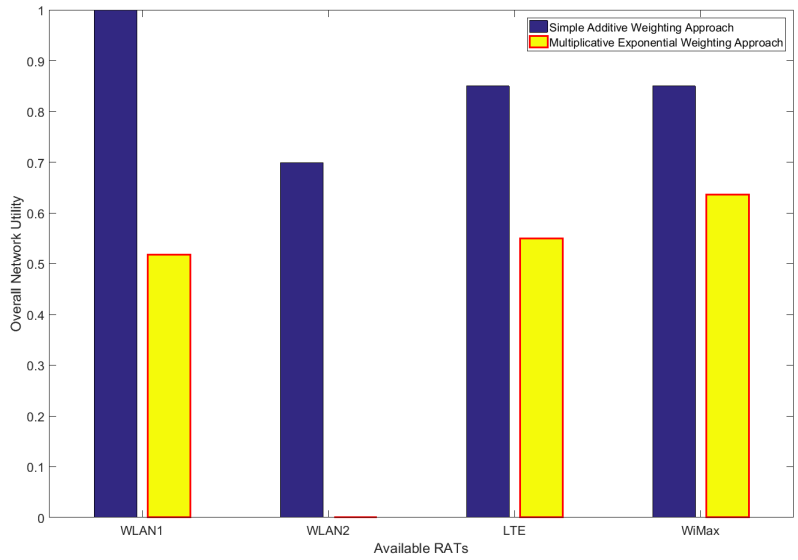

FIGURE 2.5: RAT Screening at UD based on Multiple Criteria

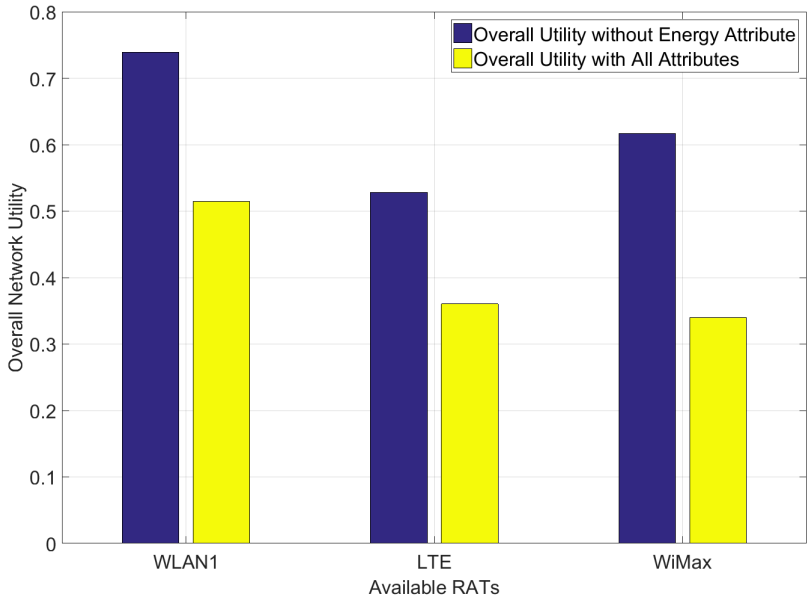

FIGURE 2.6: RAT Ranking Decision at CCN

metrics from both network and user. Therefore, this is further elaborated in Figure 2.7[b] by considering imbalance conditions across the networks where WLAN1 is overloaded and WiMAX and LTE are underloaded. Figure 2.7[b] shows that there is no change in the results of the distributive mechanism as it does not consider network load. However, our proposed hybrid approach gives a precise result by providing LTE and WiMax as the most suitable

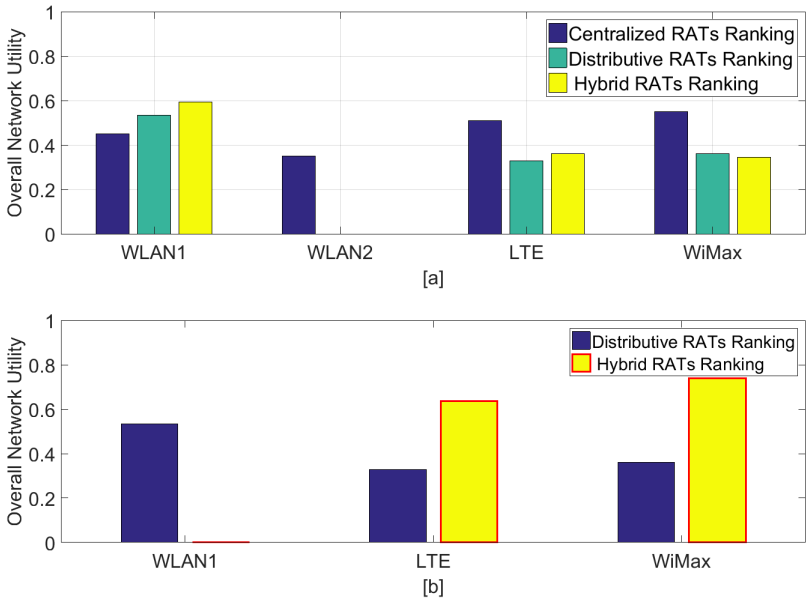

FIGURE 2.7: Comparison of RATs ranking

RATs for connectivity.

2.4 Summary

In this chapter, we presented a hybrid scheme for RATs ranking and selection in HWAN. Our approach involves two necessary steps, RATs screening, an operation performed at UD and RATs ranking and selection conducted at CCN. Our proposed scheme is different from the traditional plans for RAT selection, where only decision-making attributes are taken from either user or network. Our proposed RAT selection method distributes the decision-making process between the user and CCN, and it takes input from both network and user. The involvement of UD in the decision-making process reduces the processing burden on CCN by initiating multi-criteria RAT screening. It also helps in ignoring the inappropriate RAT from the list of available RATs. We compared our proposed hybrid scheme with the existing traditional schemes. The simulation results show that our approach gives a more precise decision of RATs ranking and selection than the current methods.

Chapter 3

QoS based Joint Radio Resource Allocation for Multi-Homing Calls in HWAN

In Chapter 2, we proposed RAT selection in HWAN. Now our objective is to allocate radio resources from the selected RATs. However, the challenge is to develop a radio resource allocation mechanism that can allocate radio resources to multi-homing calls from the chosen RATs. In this chapter, we propose a QoS-based radio resource allocation. Our optimization problem is based on system sum-rate maximization under the minimum data-rate constraint. We propose a joint radio resource allocation scheme for a HWAN composed of OFDMA based macro BS and WLAN APs. However, our objective is to decompose the radio resource allocation scheme, which can allocate radio resources (subcarrier and power) from the OFDMA system and optimal timeshare allocation from the WLAN system. Our main objective is to compare the overall sum-throughput of our proposed optimal resource allocation algorithm in HWAN for the multi-RAT approach with the single RAT approach that uses WLAN or OFDMA based system. Furthermore, in this chapter, we will investigate the impact of minimum data rate requirements of users on the proposed algorithm's convergence rate..

3.1 Background and Introduction

The problem of joint radio resource allocation has been widely explored in recent years. The authors in [49] proposed a joint resource (subcarrier and power) allocation for the downlink OFDMA based system. Radio resource allocation from allocating OFDMA and WLAN in HWAN has been explored in [50]-[54]. In [50], the authors proposed radio resource allocation based on sum-rate maximization constrained by the proportional user rate. They have considered optimal network selection and multi-homing resource allocation. Furthermore, they have considered network utility maximization (NUM) for radio resource allocation with

QoS constraints. However, they did not explore multi-homing calls. The authors in [51] proposed user association and data rate allocation based on maximizing the utility function. Maximizing mobile users' energy efficiency by providing QoS is explored in [52], where radio resource allocation for multi-homing calls is achieved. Joint bandwidth and power allocation in HWAN are studied by considering LTE and WLAN-based RATs [54]. The authors in [55] proposed a delay aware mechanism for allocating resources from WLAN and cellular BS. They considered energy-efficient transmission for multi-homing in HWAN, based on stochastic optimization. The authors of [56] addressed wireless resource management in HWAN by minimizing per bit energy consumption.

The authors in [57] proposed a simulator-based solution for the integration of different technologies in HWAN. Their work support 3GPP elements and protocols and provides multi-homing facility and resource management. In [58], the authors investigated sum-rate maximization for full-duplex OFDMA based channels. They proposed a polynomial-time algorithm that is nearly optimal under high SINR constraints. Finally, the authors in [59] explored resource allocation based on user QoS and user priority. However, their work is limited only to a cognitive network with primary and secondary users.

In Chapter 2, we performed multi-criteria-based RAT selection. In this chapter, we formulate the allocation of radio resources to mobile users from the selected networks. We explore QoS-based wireless resource management in HWAN for multi-homing calls based on optimization. Similar to [49], we consider radio resource allocation management based on sum-rate maximization. However, in our approach, we consider both cellular BS based on OFDMA and WLAN AP, which makes it different from the work done in [49]. The QoS constraint is related to the minimum data rate requirement. Using the Lagrange duality approach, we formulate the optimal subcarrier and power allocation from the OFDMA based network and timeshare allocation from WLAN.

This chapter is organized as follows. First, the system model is given in Section 3.2. Then, the problem is formulated in Section 3.3, where Lagrange decomposition is presented. Next, radio resource allocation, i.e. subcarrier and transmit power from OFDMA-based system, and timeshare from WLAN are presented in Section 3.4. Then, simulation results are given in Section 3.5. Finally, the chapter summary is presented in Section 3.6.

3.2 System Model

We consider a HWAN composed of a single macro BS (3GPP cellular network) located at the center of the cell, with L WLAN APs distributed in the cell, as shown in Figure 3.1. We assume no interference between the networks as they operate in different frequency bands. A central controller node controls and manages radio resource allocation of the HWAN. HWAN has a K number of users. User devices are multimodal, equipped with multi-homing features and can connect to cellular networks and WLAN simultaneously. An OFDMA based

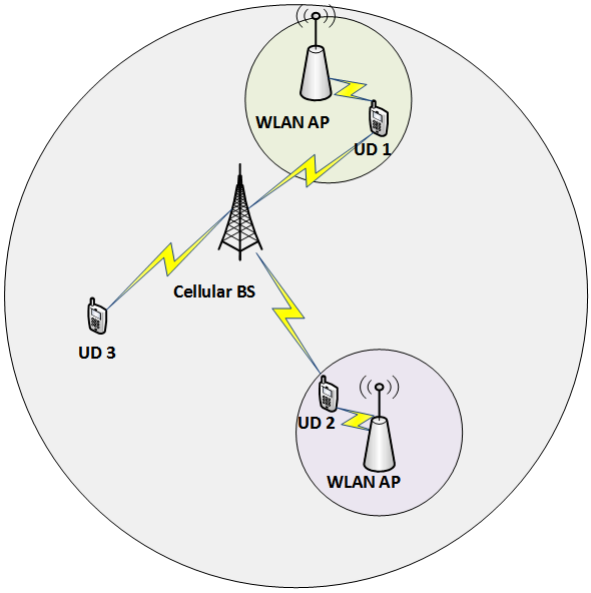

FIGURE 3.1: Heterogeneous Wireless Access Network with cellular BS and WLAN APs

system (i.e. LTE) comprises N subcarrier having bandwidth G given by $G = \frac{B}{N}$, where the total bandwidth is B. Let p_{nk} is the transmit power for the link between user k and subcarrier n from cellular BS and g_{nk} is the channel gain between user k and subcarrier n from cellular BS. We assume that multiple users could share one subcarrier in a time-sharing manner by using $x_{nk} \geq 0$ as the time-sharing fraction for the allocation of subcarrier n to user k. The transmit power allocated on subcarrier n to user k during the time-sharing slot is $\frac{p_{nk}}{x_{nk}}$.Then the maximum data rate r_{nk} approximated by Shannon theorem is given as follows:

$$r_{nk} = \begin{cases} x_{nk} G log_2(1 + \frac{p_{nk} g_{nk}}{x_{nk} G N_0}), & x_{nk} > 0, \\ 0, & x_{nk} = 0 \end{cases} \tag{3.1}$$

where N_0 is the noise spectral density. The subcarrier allocation is approximated by the set given in Eq. 3.2,

$$X = \{[x_{nk}]_{N \times K} | \sum_{k=1}^{K} x_{nk} \leq 1, \forall n, 0 \leq x_{nk} \leq 1, \forall n, k\}. \tag{3.2}$$

whereas the power allocation set is given by Eq. 3.3,

$$P = \{[p_{nk}]_{N \times K} | \, p_{nk} \geq 0, \forall n, k\}. \tag{3.3}$$

The total throughput obtained by user k from OFDMA based cellular c BS is given by Eq. 3.4,

$$R_k^c = \sum_{n=1}^{N} r_{nk}. \tag{3.4}$$

IEEE 802.11 WLAN equipped with an enhanced version of the distributive coordination function (DCF) allows users to share the entire bandwidth without any collision. It is considered that WLAN may use TDMA based approach where users can occupy the entire bandwidth in its allocated time slot t_{lk}^w. The data rate r_{lk} is approximated by determining the instantaneous signal-to-noise-ratio (SNR) between user k and AP l and is given by Eq. 3.5,

$$r_{lk}^w = f(\frac{P_{out}^w g_{lk}}{N_o}), \tag{3.5}$$

where P_{out}^w is the power transmitted by WLAN's AP. All the associated users receive the same amount of radiated power from the WLAN's AP. g_{lk} and N_0 are the link gain between user k and AP l and noise variance at the WLAN channel, respectively. The achievable data rate based on the SNR threshold is decided by $f(.)$ [61], given in Section 3.5 in Table 3.1. As per assumption, the user connects to a single access point per WLAN. The access point with the highest data rates is selected among all the available access points. The throughput received by user k from WLAN w AP l is given by Eq. 3.6,

$$R_k^w = \sum_{l=1}^{L} t_{lk} r_{lk}. \tag{3.6}$$

Based on our association assumption between APs and UDs, the sum expression includes only one term. However, to maintain the notation consistency between WLAN and OFDMA networks, the sum expression is introduced in Eq. 3.6. The maximum throughput is the sum of the throughputs of all active links and is given by Eq. 3.7,

$$R(X, P, T) = \sum_{k=1}^{K} (R_k^c + R_k^w), \tag{3.7}$$

where $X = [x_{nk}]_{N\times K}$, $P = [p_{nk}]_{N\times K}$ and $T = [t_{lk}]_{L\times K}$.

3.3 Problem Formulation

Multi-homing optimal resource allocation in HWAN is formulated in this section. The optimization problem P1 for total aggregate transmission rate for multi-homed users in heterogeneous wireless access network in the overlapped zones under QoS constraint is given by

$$\text{P1}: \quad \max R(X,P,T) = \sum_{k=1}^{K}(R_k^c + R_k^w) \tag{3.8a}$$

$$\text{Subject to } R_k^c + R_k^w \geq R_k^{min}, \ \forall k \tag{3.8b}$$

$$\sum_{k=1}^{K}\sum_{n=1}^{N} p_{nk} \leq P_{max}^c \tag{3.8c}$$

$$\sum_{k=1}^{K} t_{lk} \leq 1, \quad \forall l \tag{3.8d}$$

$$t_{lk} \geq 0 \quad \forall\, l,\, k \tag{3.8e}$$

Eq. 3.8a gives the objective function, and Eq. 3.8b is the QoS constraint that describes the minimum data rate requirement of the user k. The sum power constraint for OFDMA based cellular network is shown in Eq. 3.8c. Eq. 3.8d shows the time fraction constraint among K users at WLAN's APs, whereas Eq. 3.8e gives the physical explanation of the variable. This optimization problem is a convex optimization problem as the Hessian matrix of the function given in Eq. 3.1, i.e. the first term of the optimization problem is positive semi-definite, given by Eq. 3.9,

$$H = S \times \begin{bmatrix} -\frac{p_{nk}^2}{x_{nk}(x_{nk}+\frac{p_{nk}g_{nk}}{GN_0})^2} & \frac{p_{nk}}{(x_{nk}+\frac{p_{nk}g_{nk}}{GN_o})^2} \\ \frac{p_{nk}}{(x_{nk}+\frac{p_{nk}g_{nk}}{GN_o})^2} & -\frac{x_{nk}}{(x_{nk}+\frac{p_{nk}g_{nk}}{GN_0})^2} \end{bmatrix}, \tag{3.9}$$

where $S = \frac{g_{nk}^2}{ln(2)GN_0}$. The Hessian matrix shows that it is a concave function in terms of x_{nk} and p_{nk}. Furthermore, all the constraints are affine and linear [60]. Thus, it is possible to obtain polynomial time solution for the optimization problem given in Eqs. 3.8a-3.8e. Eq. 3.6, i.e. the second term of the optimization problem is a linear problem, since we have real-valued positive data rate r_{lk}^w and $t_{lk} \geq 0$, and the corresponding constraints related to R_k^w are linear and affine.

3.4 Radio Resource Allocation

We can solve our convex optimization problem using the Lagrange duality approach. The Lagrangian of problem P1 is given by

$$\begin{aligned} L(X,P,T,\lambda,\sigma,\vartheta,\mu) = & \sum_{k=1}^{K}\sum_{n=1}^{N}((1+\lambda_k)r_{nk} - I) \; + \\ & \sum_{k=1}^{K}\sum_{l=1}^{L}((1+\lambda_k)t_{lk}r_{lk} - J) \; - \\ & \sum_{k=1}^{K}\lambda_k R_k^{min} + \sigma P_{max}^{c} + \sum_{l=1}^{L}\vartheta_l \end{aligned} \tag{3.10}$$

where

$$I = \sigma p_{nk} \tag{3.11}$$

$$J = \vartheta_l t_{lk} - \mu_{lk} t_{lk} \tag{3.12}$$

The Lagrange multipliers $\lambda_k, \sigma, \vartheta_l, \mu_{lk}$ are non-negative. The dual function is given by

$$D(\lambda,\sigma,\vartheta,\mu) = \max_{X\in Y, P\in Z, T} L(X,P,T,\lambda,\sigma,\vartheta,\mu). \tag{3.13}$$

The dual function of problem P1 is given by

$$\underset{(\lambda,\sigma,\vartheta,\mu)\geq 0}{\text{minimize}} D(\lambda,\sigma,\vartheta,\mu). \tag{3.14}$$

The optimal values for the primal and dual problems are the same since the convex optimization problem P1 satisfies Slater's condition and holds duality [60]. Eq. 3.13 can be written as

$$
\begin{aligned}
D(\lambda,\sigma,\vartheta,\mu) = \max_{X\in Y,P\in Z} \sum_{k=1}^{K}\sum_{m=1}^{M}((1+\lambda_k)r_{nk} - I) \; + \\
\max_{T} \sum_{k=1}^{K}\sum_{l=1}^{L}((1+\lambda_k)t_{lk}r_{lk} - J) \; - \\
\sum_{k=1}^{K}\lambda_k R_k^{min} + \sigma P_{max}^{c} + \sum_{l=1}^{L}\vartheta_l
\end{aligned}
\tag{3.15}
$$

Optimal power and subcarrier allocation are achieved by maximizing the first term, and optimal timeshare is obtained by maximizing the second term of the Eq. 3.15. Thus, this problem is decomposed into two subproblems. From problem P1, it can be observed that constraint Eq. 3.8b is required to be fulfilled by both terms. However, these terms can be solved separately [56] as subproblems and are proved as follow.

3.4.1 Optimal Subcarrier and Power Allocation

The first term of Eq. 3.15 is jointly convex in x_{nk} and p_{nk}. We consider the first term of Eq. 3.15 as a subproblem 1 for radio resource allocation from OFDMA based system.

$$
\text{subproblem 1}: \quad \max_{X\in Y,P\in Z} \sum_{k=1}^{K}\sum_{n=1}^{N}((1+\lambda_k)r_{nk} - I). \tag{3.16}
$$

The optimal subcarrier and power allocation are obtained by differentiating Eq. 3.16 with respect to p_{nk} and equating it to zero.

$$
p_{nk} = x_{nk}G[\frac{(1+\lambda_k)}{\sigma \ln 2} - \frac{No}{g_{nk}}]^{+}, \tag{3.17}
$$

where $[x]^{+} = max(0,x)$. Putting Eq. 3.17 into Eq. 3.16 gives

$$
\Phi_{nk}(x_{nk},\alpha_{nk}) = x_{nk}\alpha_{nk}(\lambda_k,\sigma), \tag{3.18}
$$

where

$$
\begin{aligned}
\alpha_{nk}(\lambda_k,\sigma) = \; -\sigma G(\frac{(1+\lambda_k)}{\ln 2\sigma} - \frac{No}{g_{nk}}) + \\
(1+\lambda_k)Glog_2(\frac{g_{nk}}{No}\frac{(1+\lambda_k)}{\ln 2\sigma}.
\end{aligned}
\tag{3.19}
$$

For an ergodic channel with continues CDF [49], and for a given optimal values of λ_k^* and σ^* the optimal solution for Eq. 3.13 is given as

$$\begin{cases} x^*_{nk^*} = 1, \\ p^*_{nk^*} = G[\frac{(1+\lambda_k)}{\sigma \ln 2} - \frac{No}{g_{nk}}]^+, & \text{if} \quad k^* = \arg\max_k \ \alpha_{nk}(\lambda_k, \sigma) \\ x^*_{nk^*} = 0, \ \ p^*_{nk^*} = 0, & \text{otherwise.} \end{cases} \tag{3.20}$$

We have assumed that a subcarrier is shared among multiple users in a time-sharing manner. However, this ergodic channel with continuous CDF leads to no sharing. From Eq. 3.20 we can see that no subcarrier n is allocated to any user k if they suffer from severe channel fading, and thus no power allocation. Eq. 3.20 shows that subcarrier n is exclusively allocated to user k^* if it is the only maximizer of Eq. 3.19, and thus $x_{nk} = 1$. If multiple users maximize Eq. 3.19, then the allocation of a subcarrier n among multiple users is not possible as the variables λ_k and σ may not be optimal. However, this multi-maximizer case has zero probability due to the ergodic channel having continuous CDF. Thus, for the optimal values λ_k^* and σ^*, a unique solution x_{nk} can be achieved.

Proposition 1. The probability of multiple users maximizing Eq. 3.19 for a subcarrier n with optimal values of λ_k^* and σ^* is zero. Optimality with probability 1 (w.p.1) leads to a unique solution of x_{nk} for λ_k^* and σ^*.

Proof. Let assume that two users $k1$ and $k2$ maximize Eq. 3.19 such that $\alpha_{nk1}(\lambda_k^*, \sigma^*) = \alpha_{nk2}(\lambda_k^*, \sigma^*)$, where $k1 \neq k2$. Eq. 3.19 is a function of channel gain g_{nk} and its value increases for $g_{nk} > 0$. However, wireless fadding channel is continuous and the event $\alpha_{nk1}(\lambda_k^*, \sigma^*) - \alpha_{nk2}(\lambda_k^*, \sigma^*) = 0$, for two usersr $k1$ and $k2$ has Lebesgue measure zero i.e. the probability of the event is zero such that $P[\alpha_{nk1}(\lambda_k^*, \sigma^*) = \alpha_{nk2}(\lambda_k^*, \sigma^*)] = 0$. □

3.4.2 Optimal Time Share Allocation

The time fraction allocation is determined by considering the second part of Eq. 3.15, given by subproblem two as follows.

$$\begin{aligned} \text{subproblem 2} : \Lambda(\lambda, \vartheta, \mu) = \max_T \sum_{k=1}^{K} \sum_{l=1}^{L} ((1+\lambda_k) t_{lk} r_{lk} \\ - \vartheta_l t_{lk} + \mu_{lk} t_{lk}) \end{aligned} \tag{3.21}$$

UDs in the coverage of WLAN AP can connect to only one AP at most since, per our assumption, there are no overlapping coverage zones between WLAN APs. Thus, UDs with

$t_{lk} = 0$ have no connectivity to WLAN AP. However, it is possible to decompose this problem of timeshare allocation into L subproblems as these WLAN APs work independently.

$$\Lambda(\lambda, \vartheta, \mu) = \max_{T} \sum_{k=1}^{K} ((1 + \lambda_k) t_{lk} r_{lk} - \vartheta_l t_{lk} + \mu_{lk} t_{lk}). \tag{3.22}$$

Differentiating Eq. 3.22 with respect to t_{lk} gives

$$\frac{\partial \Lambda}{\partial t_{lk}} = (1 + \lambda_k) r_{lk} - \vartheta_l + \mu_{lk} \tag{3.23}$$

$$\begin{cases} t_{lk} = 0, \ , \ \text{if} \ \ (1 + \lambda_k) r_{lk} - \vartheta_l + \mu_{lk} = 0 \\ \infty \qquad , \ \text{otherwise} \end{cases} \tag{3.24}$$

This Eq. 3.24 helps in determining the equation of λ_k, a common multiplier that links the resources of cellular BS and WLAN APs. It is important to define a threshold value for λ_k^T, as it balances the wireless resource allocation of cellular BS and WLAN APs among users [56], Eq. 3.23 gives $\lambda_k = \frac{\vartheta_l - \mu_l k}{r_{lk}} - 1$ for $\mu_{lk} \geq 0$, whereas $\lambda_k^T = \frac{\vartheta_l}{r_{lk}} - 1$ for $\mu_{lk} = 0$. There is an inverse relationship between λ_k and r_{lk}, the power allocation p_{nk} from the cellular network BS is directly related to λ_k given in Eq. 3.20. For greater value of r_{lk}, a lower value of λ_k is established as threshold value and less power is allocated to the user from cellular network BS and vice versa. For user k to be connected to an AP l, the $t_{lk} > 0$ such that $t_{lk} = \frac{R_{min} - \sum_{n=1}^{N} r_{nk}}{r_{lk}}$, where $\mu_{lk} = 0$ and $\lambda_k = \lambda_k^T$, whereas, for a user k having no connection with an AP l, $t_{lk} = 0$ then $\mu_{lk} \geq 0$, and $\lambda_k < \lambda_k^T$. In this case, the user receives resources from a cellular BS, that is an example of single network connection. Furthermore, if a user is not in coverage of an WLAN AP then $t_{lk} = 0$. These conditions of connectivity with WLAN AP are given in Eq. 3.25.

$$\begin{cases} t_{lk^*} = \frac{R_{min} - \sum_{n=1}^{N} r_{nk}}{r_{lk}} \leq 1 \ , & \text{if} \quad \mu_{lk} = 0, \lambda_k = \lambda_k^T \\ t_{lk^*} = 0, & \text{if } \mu_{lk} \geq 0; \ \lambda_k < \lambda_k^T \\ t_{lk} = 0, & \text{if} \quad l \neq l^* \end{cases} \tag{3.25}$$

3.4.3 Multiplier Updates

The optimal values of Lagrangian multipliers λ^*, σ^* and ϑ^* are obtained from a differentiable dual function given in Eq. 3.14 by using gradient descent method, that is given as follows:

$$\lambda_k(j+1) = [\lambda_k(j) - \gamma(R_{min} - \sum_{n=1}^{N} r_{nk}(j) - \sum_{l=1}^{L} t_{lk}(j) r_{lk})]^+, \tag{3.26}$$

$$\sigma(j+1) = [\sigma(j) - \beta(\sum_{k=1}^{K}\sum_{n=1}^{N} p_{nk} - P_{max}^{c})]^{+}, \tag{3.27}$$

$$\vartheta_l(j+1) = [\vartheta_l(j) - \delta(\sum_{k=1}^{K} t_{lk} - 1)]^{+}, \tag{3.28}$$

where γ, β and δ are the gradient step size, and j is an iteration index. The convergence of Eqs. 3.26-3.28 is possible as the gradient of Eq. 3.13 satisfies the Lipchitz continuity condition. The value of the Lagrangian multiplier μ is updated [56] as

$$\mu_{lk}^{*} = \vartheta_l^{*} - (1 + \lambda_k^{*})r_{lk}. \tag{3.29}$$

The joint radio resource (optimal subcarrier, power and timeshare) allocation is depicted in algorithm 1.

Algorithm 1 Radio Resource Allocation

1: Initialize $\lambda[0]$, $\sigma(0)$, $\vartheta(0)$ and $j = 0$
2: Calculate x_{nk} and p_{nk} using Eq. 3.20
3: Calculate t_{lk} using Eq. 3.25
4: Update λ_k, σ, ϑ_l and μ_{lk} using Eqs. 3.26-3.29
5: if $|\lambda_k(j+1) - \lambda_k(j)| < \epsilon$, $|\sigma(j+1) - \sigma(j)| < \epsilon$, $|\vartheta_l(j+1) - \vartheta_l(j)| < \epsilon$, and $|\mu_{lk}(j+1) - \mu_{lk}(j)| < \epsilon$,
6: Then $[x_{nk}^{*}], [p_{nk}^{*}], [t_{lk}^{*}] = [x_{nk}(j)], [p_{nk}(j)], [t_{lk}(j)]$,
7: $[\lambda_k^{*}] = [\lambda_k(j)]$, $\sigma^{*} = \sigma(j)$, $[\vartheta_l^{*}] = [\vartheta_l(j)]$
8: else j=j+1; go back to step 2
9: end if

3.4.4 Complexity Analysis

The computation complexity of Algorithm 1 is $O(LK) + O(NK/\mu^2)$ since the CCN allocates radio resources. Here $O(NK/\mu^2)$ is the complexity of subcarrier and power allocation of the cellular BS, where $(1/\mu^2)$ corresponds to the number of iterations that achieves convergence accuracy of μ of Eq. (3.20). $O(LK)$ is related to the complexity of the time fraction allocation of WLAN.

3.5 Results and Discussion

The performance evaluation is carried out through Matlab-based simulation results. For our simulations, we consider a cellular macro BS located at the center (0,0) with a radius of 1000 meters. Four APs of WLAN at location (500,433.107), (-500, 433.107), (500,-433.107), (-500,-433.107) overlay in the coverage of macrocell BS. These APs have a coverage of 250 meters.

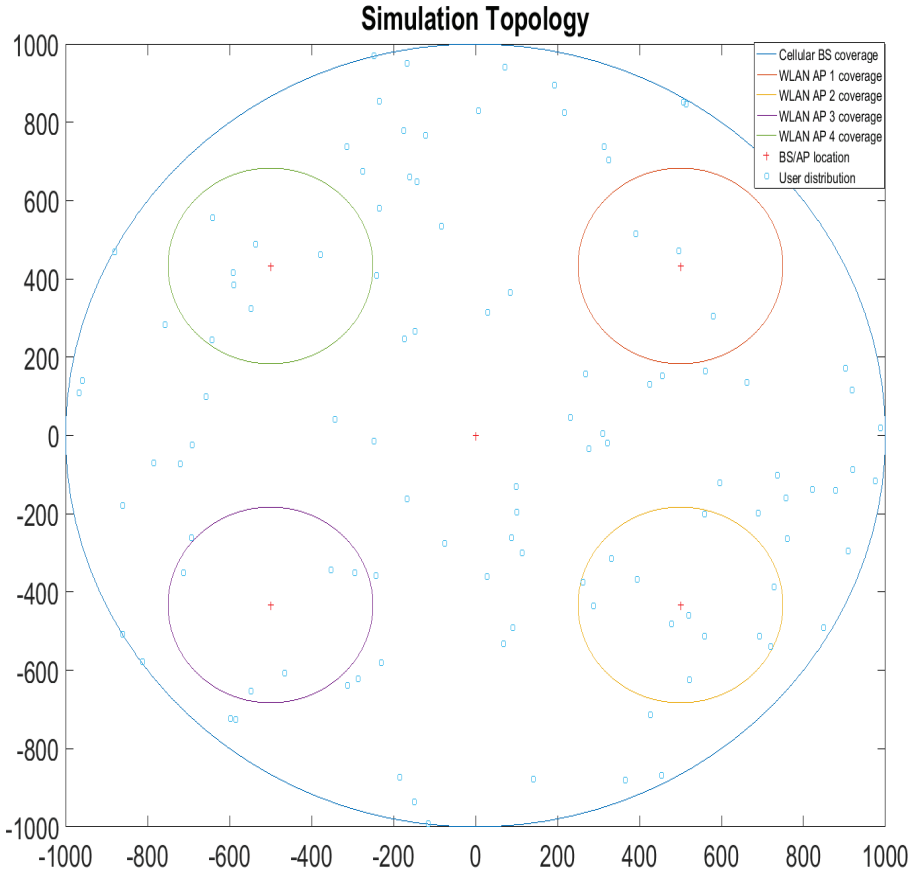

FIGURE 3.2: Simulation Topology

Random distribution of users is assumed in this simulation area. The minimum data rate requirement for each mobile user is 4 Mbps. The simulation topology is given in Figure 3.2. The rate adaptation mechanism described in [61] is used to determine the data rate for a UD k from an AP l. This rate adaptation mechanism involves mapping of SNR to data rates shown in Table 3.1. The rest of the simulation parameters related to the cellular network are given in Table 3.2.

Figure 3.3 shows the system sum-throughput performance of HWAN. From Figure 3.3, it can be seen that the system sum-throughput for LTE and our proposed optimal mechanism increases with the number of users in the system. In contrast, the system sum-throughput for WLAN decreases with the increase in the number of users. Initially, the sum rate for WLAN is high, but a gradual decrease can be observed with an increase in the number of users. WLAN performance degrades with an increase in the number of users. It can be seen from the graph that LTE has the lowest sum-throughput, whereas our proposed optimal mechanism has the highest sum-throughput. Our proposed approach maximizes total

TABLE 3.1: SNR Versus Rate [61]

SNR range (dB)	Rate (Mbps)
> 24.56	54
$[24.05, 24.56]$	48
$[18.8, 24.05]$	36
$[17.04, 18.8]$	24
$[10.79, 17.04]$	18
$[9.03, 10.79]$	18
$[7.78, 9.03]$	9
$[6.02, 7.78]$	6
<6.02	0

TABLE 3.2: Simulation Parameters

Features	Description
Total subcarriers	1024
Path Loss Model	Cost 231 Hata Model
Subcarrier BW	15KHz
Noise Spectral Density	-174dBm/Hz
Total transmit Power	50 Watts

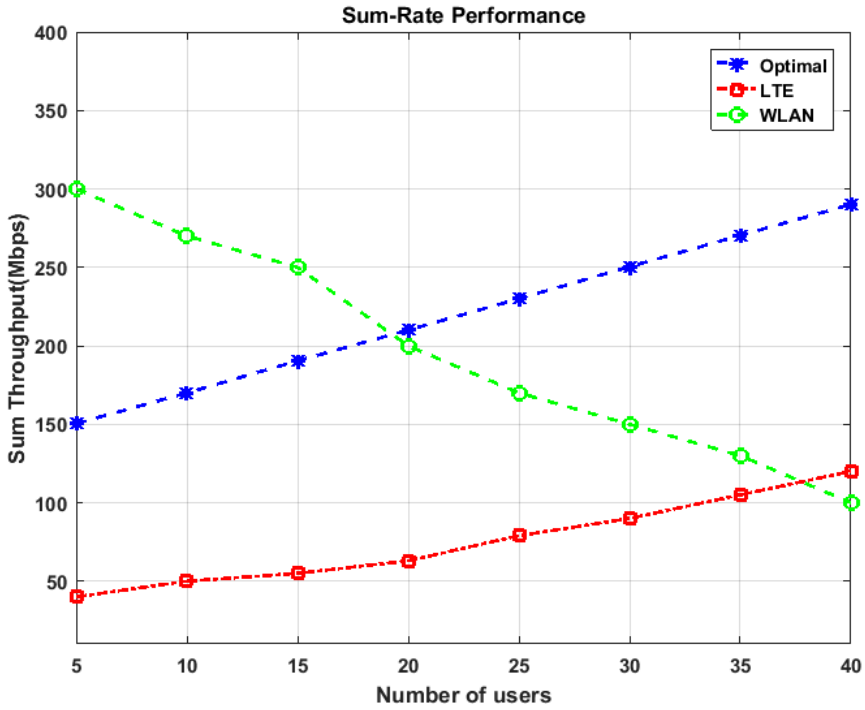

FIGURE 3.3: Sum-Throughput Vs. Number of users

system throughput under the constraint of transmit power and minimum QoS requirements in HWAN.

In Figure 3.4, we analyze the impact of the minimum data rate requirement on the convergence rate for our proposed optimal joint resource allocation mechanism. We consider 20 mobile users and three different values for the minimum data rate. The proposed algorithm has a fast convergence rate. The total system throughput almost converges after 20 iterations, as shown in Figure 3.4. It can be seen from the plot that the minimum data rate of mobile users has no impact on the convergence rate of the proposed algorithm.

3.6 Summary

In this chapter, we presented joint radio resource allocation for multi-homing calls under QoS constraint in HWAN. We considered the OFDMA based system (LTE or WiMAX) and WLAN in our HWAN. More specifically, we considered the system sum-throughput maximization for HWAN. We formulated our optimization problem as a convex optimization problem. By applying the Lagrangian duality, we decomposed our problem into two parts. The first subproblem gives optimal radio resources from the OFDMA based cellular system, whereas the second subproblem solution yields optimal timeshare allocation from WLAN AP. Less power is allocated to users from cellular BS if they receive high data rates from WLAN AP.

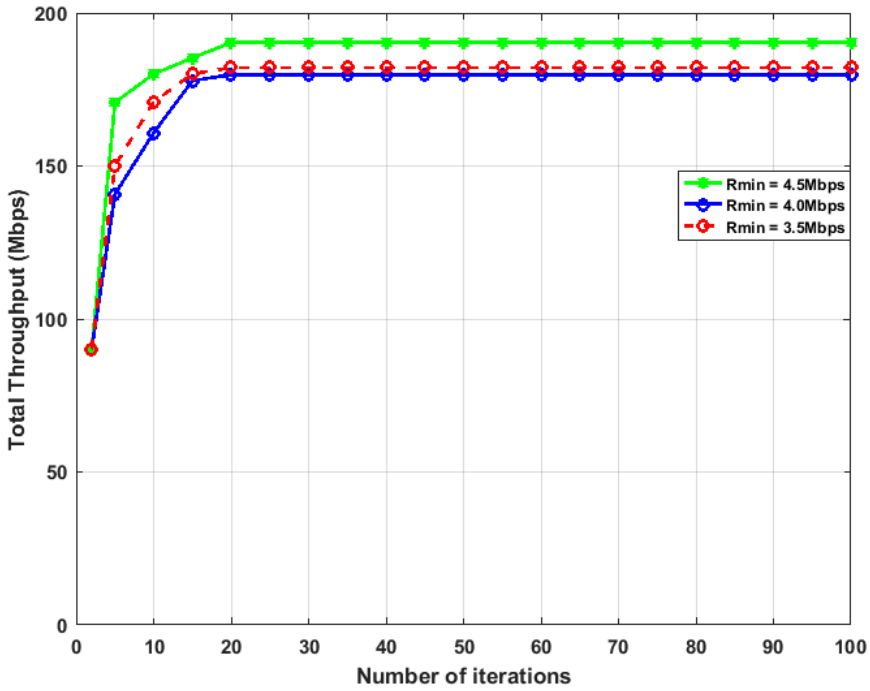

FIGURE 3.4: Impact of minimum data rate on convergence rate

It is proved mathematically in Section 3.3. Furthermore, the simulation results show that our proposed algorithm gives an enhanced system sum rate. Finally, the simulation results reveal that the convergence rate of our proposed algorithm is insensitive to the minimum data rate requirement of mobile users.

Chapter 4

Load Balancing in Heterogeneous Wireless Access Network

For stable network performance, it is crucial to balance the load across the available RATs. In this chapter, our objective is to investigate load balancing in heterogeneous wireless access networks. We equip the CCN with a load balancing mechanism. The CCN takes feedback from the BS of available RATs to balance the load across the network. In this chapter, our main objective is to consider different imbalance load conditions and then evaluate the performance of our proposed scheme through call blocking probability and network utilization.

4.1 Background and Introduction

The heterogeneous wireless access network is composed of WLAN, LTE, and 5G. These wireless networks have different data rates, coverage, and capacity [62], [63]. Smartphones equipped with multi-RAT interfaces can connect to single or multiple networks simultaneously [64]. The literature survey reveals that users have a greedy approach. They select the best network among the available options [65], as different wireless access networks operate autonomously without sharing any information on radio resource allocation. As a result, an imbalance in load across the HWAN is created that leads to the problems of network congestion and underutilization, which alternatively increases call blocking and call dropping, provides poor QoS to end-users and weak utilization of resources at a network level. It is therefore required to balance the network load, a key feature of radio resource management that increases network performance and enhances resource utilization. Load balancing can be achieved by using cooperation and collaboration among the different RATs of heterogeneous wireless access network [10]. However, this leads to an increase in signalling overhead. In the distributive approach, the networks assist the users in choosing the appropriate network according to their load by broadcasting their loading information to the users. However, different networks always have the intention to serve more users, as network revenue is directly related to the number of users. Therefore, it is required to mask the network loading information from users by implementing a centralized intelligent entity to

balance the load across different RATs of HWAN. The central controller has a global view of the entire HWAN, balances the load across AP/BS of different RATs. Improvement in system performance via load balancing in the heterogeneous network has been explored in [66]-[68]. In [69], the authors presented joint resource management schemes for determining system performance in a multi-RAT heterogeneous environment. The emphasis is focused on QoS aware and green coverage management achieved through load balancing and network selection procedures in the multi-RAT heterogeneous wireless network. The authors in [70] managed traffic flow through different radio networks using soft load balancing. The authors in [71] proposed a load balancing algorithm in a heterogeneous wireless access network based on a real-time network selection scheme. The authors in [77] proposed an algorithm for balancing the load and boosting the capacity of HWAN by improving the dynamic spectrum access. Their work involves the design of a new framework for the RRM/MAC layer to fulfil the need for dynamic spectrum access and advancement in HWAN. The authors in [73] proposed a relay-based device-to-device load balancing scheme to balance the load between macro BS and Femto BS in a multiuser HWAN environment. In this Chapter, we propose a load balancing mechanism, where a CCN balances the load by equally distributing the load among all the available RATs in the HWAN. The AP/BS of different RATs updates the CCN with their load ratios in a distributive way, thus making it a hybrid approach of load balancing. This simple and robust approach decreases call blocking probability and increase spectrum usage of underutilized RATs. The obtained results show the effectiveness of our proposed scheme.

The rest of the Chapter is organized as follows: Section 4.2 describes our proposed system model, which includes a network model and algorithm for load balancing. Section 4.3 discusses simulation results. Finally, Section 4.4 gives the conclusion of our research work in this Chapter.

4.2 System Model

4.2.1 Network Model

We consider load balancing in a multi-RATs heterogeneous wireless environment. Figure 4.1 shows heterogeneous wireless access network composed of a single macro-cell BS of RAT_1 that overlays M smaller randomly distributed $RATs$ i.e. $RAT_i, i = 2...M$. Further, it is assumed that macro-cell BS located at the center of the cell covers the whole service area, whereas the M randomly distributed $RATs$ with limited coverage provide services in overlapping zones only within the macro-cell coverage area. These M small $RATs$ are not deployed at the cell edges. There exists no interference as these $RATs$ are operated in different frequency bands. In the overlapping zones, users have multi-connectivity options. It is assumed that within the overlapping regions, the users receive the strongest signal R_{ss}

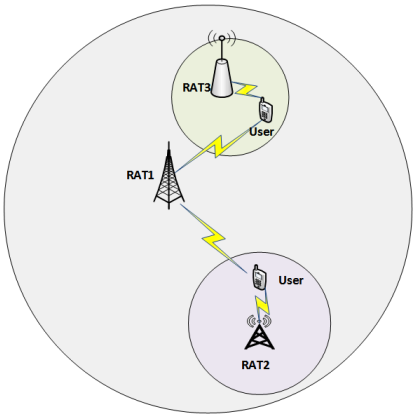

FIGURE 4.1: Heterogeneous Wireless Access Network Architecture

from the available options. It is assumed that the BS/AP of these different *RATs* has limited bandwidth. A central controller node is managing the resources of all the available *RATs*. It is assumed that these different *RATs* either belong to same operator[1]/different operators (with collaboration)[2] or an autonomous[3] system which maintains the SLA of different network operators. Our proposed architecture is comparable to IP multi-media (IMS) system architecture. The central controller node operates at a large time scale. The central controller node is equipped with a load monitor and handoff controller. The central controller node receives load updates from the available *RATs* of different networks based on some threshold value. As the central controller node has a global view of the whole network, once an imbalance in load is detected, the central controller node re-allocates the radio resources. The handoff manager is then invoked, and users are directed to handoff to the suitable target *RAT*. Our proposed load balancing scheme not only overcomes the condition of overload but also maintains an equal load ratio among all networks and avoids the underutilization of a network and its radio resources. It should be noted that the central controller node initiates re-allocation and handoff only upon imbalance in load, thus avoids the ping pong effect due to frequent handoff.

4.2.2 Proposed Algorithm for Load Balance

Load balancing in a wireless communication network is implemented to efficiently utilize the available radio resources of radio access technologies and prevent the occurrence of undesired conditions of congestion and spectrum shortage. In our proposed approach, the central controller node monitors the load condition of BS/AP of RATs within the cell, and the load balancing process is initiated upon imbalance in load. The central controller node

[1]http://www.o2.co.uk/apps/tu-go
[2]www.bitbuzz.com/index.html
[3]hhttp://fi.google.com

has a global view of the entire network, which makes load balancing possible in the heterogeneous wireless environment. The load ratio of a network is given by

$$L_R = B_{io} / B_{max} \tag{4.1}$$

where B_{io} corresponds to the number of channels in use, B_{max} is the maximum number of channels, and L_R is the load ratio of the BS/AP of a RAT under consideration within the cell. In our proposed approach, we define $L_{threshold}$ as an indicator for overload condition in the cell. The RATs within the cell updates the central controller node about its load status on a large time scale (i.e. in order of seconds to minutes). The load balancing process is initiated if the load ratio of a RAT becomes equal or greater than $L_{threshold}$, i.e. $L_R \geq L_{threshold}$. Based on our proposed network model, load balancing is initiated during the following three scenarios.

4.2.2.1 Scenario I

Let the macro-cell BS of RAT_1 suffers an overload condition. Upon updating the load monitor, the process of load balancing is initiated. From the network model, it can be observed that users belonging to both RAT_1 and M distributed $RATs$ are available in the overlapping regions. Load balancing is achieved by redistributing the resources of these $RATs$ within the overlapping zones, such that the radio resources of RAT_1 are released by transferring its users to the overlaid low coverage $RATs$. The imbalance in load is tackled by maintaining the same load ratio across all the available $RATs$ in the heterogeneous wireless network. CCN is updated with the new load ratio of the available networks. The mechanism of load balancing during Scenario I has been depicted in Figure 4.2. Here we assumed the BS/APs of only two overlaid $RATs$ i.e. RAT_2 and RAT_3 for $M = 2$. User2 is in the overlapping region 1 where RAT_1 and RAT_2 provides the services, whereas User1 is in the region shared by RAT_1 and RAT_3. Both User 1 and User 2 initially belong to RAT_1. After load balancing, both User1 and User2 are directed to switch their networks.

4.2.2.2 Scenario II

Suppose the load monitor observes an overload condition in one of the inner overlaid RAT_i. The load balancing mechanism will only be invoked within the overlapping region covered by both RAT_1 and congested RAT_i. Calls are moved from RAT_i to RAT_1 and same load ratio is maintained between RAT_1 and RAT_i. The central controller node is updated about the new load status. The load balancing procedure is highlighted in Figure 4.3. In this case, both User1 and User2 are initially connected to RAT_2, and after load balancing, they are moved to RAT_1.

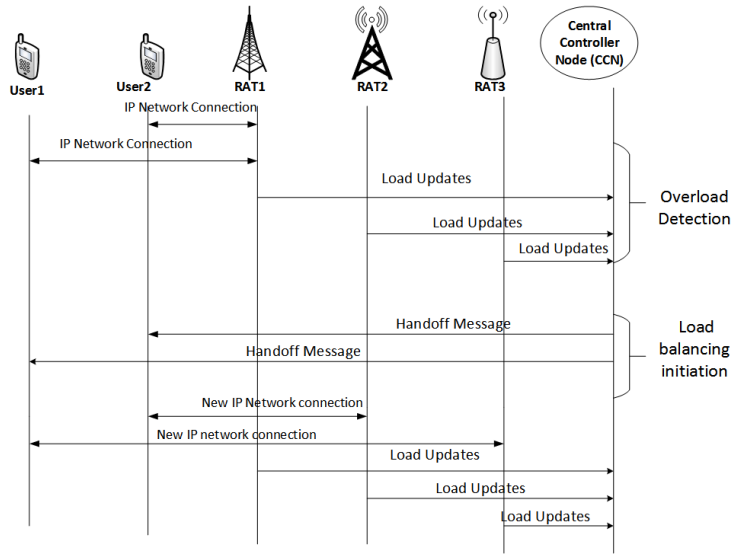

FIGURE 4.2: Load Balancing Procedure (Scenario I)

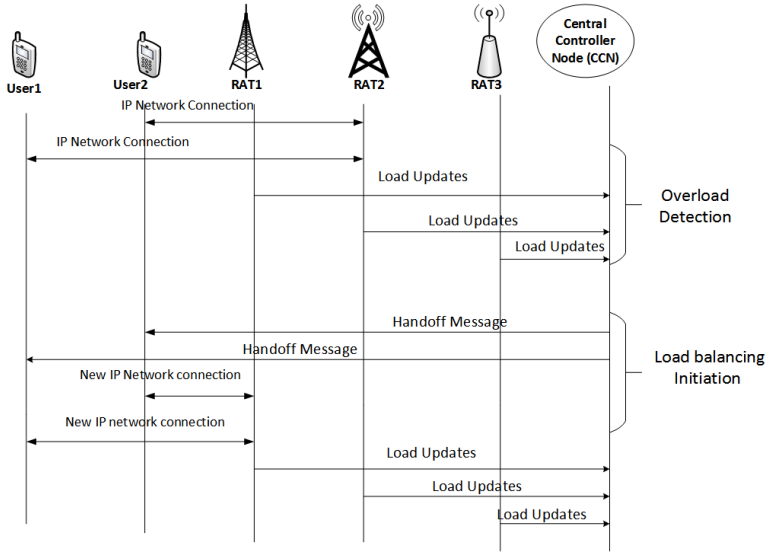

FIGURE 4.3: Load Balancing Procedure (Scenario II)

4.2.2.3 Scenario III

It is possible to expect a worse scenario where RAT_1 is unable to accommodate more load from RAT_i to maintain a load ratio less than the threshold, or there exists a congested overlapped region. Then the central controller considers the non-congested overlapping region. The load balancing is performed in two steps. Step 1) central controller node first releases RAT_1 resources by switching its calls to RAT_i in the non-congested region. Step 2) switch calls from RAT_i in the congested region to RAT_1, as a room in RAT_1 is created in step 1 to accommodate handoff calls. The imbalance in load is overcome by maintaining the same load ratio across all the available networks. This load balancing procedure is shown in Figure 4.4.

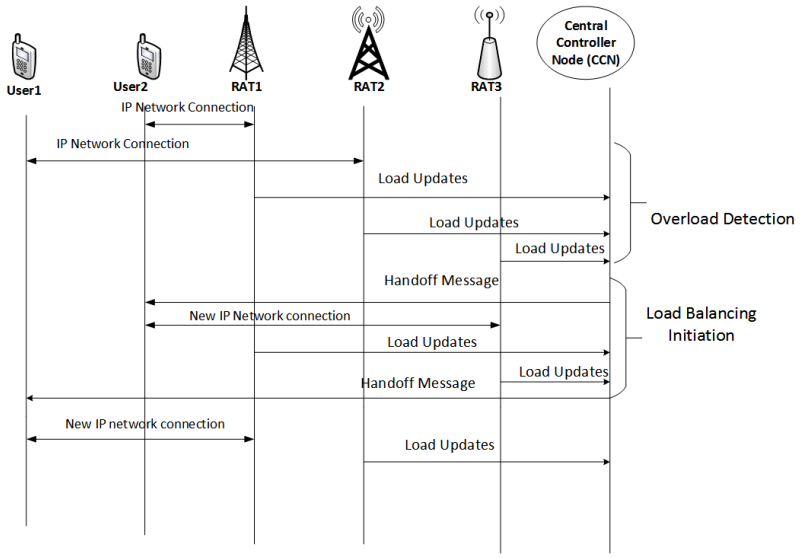

FIGURE 4.4: Load Balancing Procedure (Scenario III)

CCN is updated with the new load status of the heterogeneous wireless access networks.

4.3 Simulation Results

Simulations model is based on Figure 4.1 which considers three *RATs*, RAT_1 covering the entire service area with two inner *RATs*, i.e. RAT_2 and RAT_3. RAT_1, RAT_2, and RAT_3 are assigned with radio resources (channels) 20, 10 and 8 respectively. It is assumed that the inner *RATs* are not positioned at the edge of the RAT_1 coverage area. The arrival of calls in the network follows Poisson distribution, whereas call holding time follows exponential distributions. Mobile users are uniformly distributed in the entire service area. The simulation parameters are shown in Table I. Simulation results are carried out in the MATLAB simulator. The performance of the load balancer is evaluated by plotting "call blocking probability vs. offered load ($\Lambda = \frac{\lambda}{\mu}$)" of all the available *RATs* before and after load balancing. In the simulations, we implemented all the three scenarios described in Section II and are shown in Figure 4.5, 4.6 and 4.7. From Figure 4.5, we can see that initially RAT_1 is overloaded and has 100% blocking probability, whereas RAT_2 and RAT_3 are under-loaded, having low blocking probabilities. The central controller node has detailed information of the entire heterogeneous wireless access network. It will re-distribute the load of RAT_1 among RAT_2 and RAT_3 in the overlapping regions. A decrease in call blocking probability of RAT_1 after performing load balancing can be observed. The call blocking probabilities of RAT_2 and RAT_3 are increased. However, these blocking probabilities are below the threshold value of 1%. Note that after load balancing, both RAT_2 and RAT_3 have the same call blocking probability as shown in Figure 4.5.

TABLE 4.1: Simulation Parameters

Features	RAT1	RAT2	RAT3
Total Channels	20	10	8
Coverage zone	800m Circular	180m Circular	180 Circular
Call Holding Time ($1/\mu$)	120 s	120 s	120 s
Load Threshold	80%	80%	80%
Call arrival rate (λ) (Poision Distribution)	[0.02, 0.08, 0.14,....0.8]		

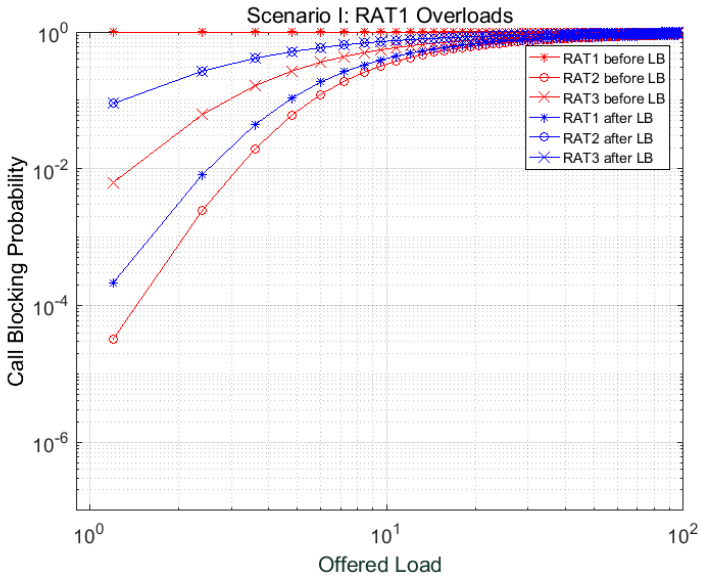

FIGURE 4.5: Call Blocking Probability Vs. Offered Load (Scenario I)

In Figure 4.6, inner overlaid RAT_2 only is considered as overloaded. In this situation, load balancing is applied only on the *RATs* available in the overlapping region where the resources of both RAT_1 and RAT_2 are available. The overload in RAT_2 is resolved by allocating the resources of RAT_1 to the users of RAT_2. Resources of RAT_3 remain the same. The decrease in call blocking probability in RAT_2 can be clearly observed, whereas RAT_1 call blocking is still under threshold value. Note that the call blocking probability remains the same before and after load balancing for RAT_3 as shown in Figure 4.6.

Figure 4.7 shows the worst scenario, where inner network RAT_2 is overloaded, and RAT_1 load ratio is above the threshold. High call blocking probabilities of RAT_1 and RAT_2 before invoking the load balancing process are shown in Figure 4.7. The central controller node

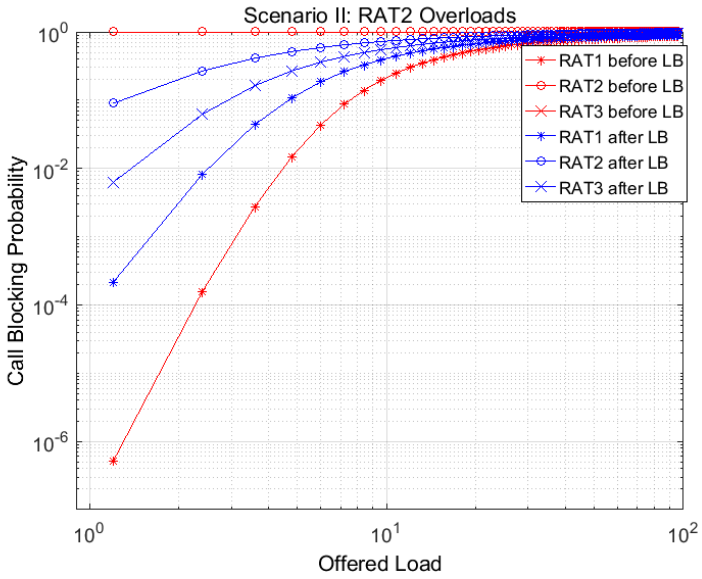

FIGURE 4.6: Call Blocking Probability Vs. Offered Load (Scenario II)

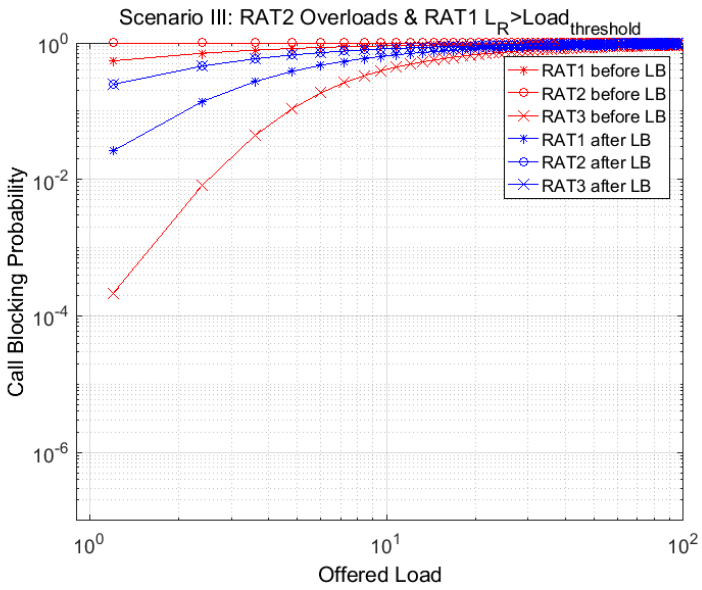

FIGURE 4.7: Call Blocking Probability Vs. Offered Load (Scenario III)

considers the load of all available *RATs* in the heterogeneous wireless access network. Overload in RAT_1 and RAT_2 is resolved by transferring calls from RAT_1 to RAT_3 and then from RAT_2 to RAT_1 while keeping their load ratios under threshold. The performance parameter, i.e. call blocking probability, for all the three *RATs* can be visualized in Figure 4.7. Both RAT_2 and RAT_3 have identical call blocking probability after load balancing shown in Figure 4.7.

Another set of graphs, Figures 4.8, 4.9, and 4.10, show bandwidth utilization before and

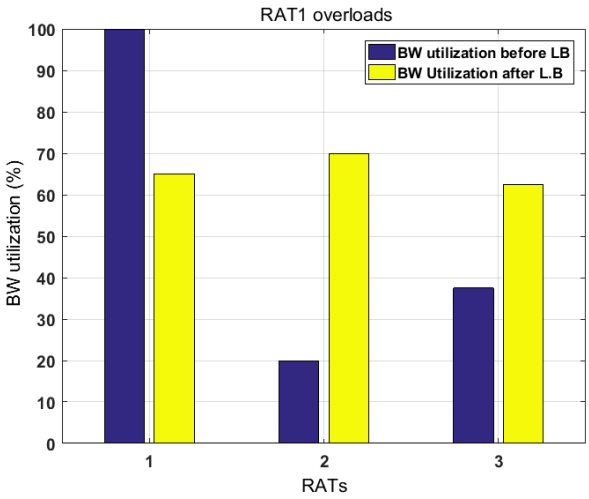

FIGURE 4.8: RAT BW utilization (Scenario I)

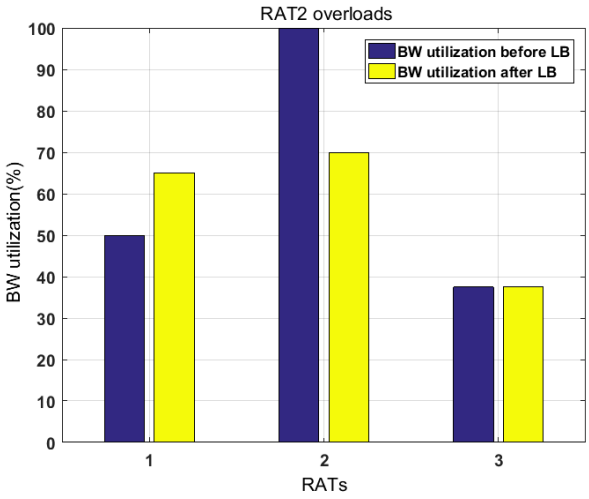

FIGURE 4.9: RAT BW utilization (Scenario II)

after load balancing at the available BS/AP of these different *RATs* for all the scenarios discussed in section II. Before load balancing, overload *RATs* show maximum BW utilization, whereas other *RATs* with enough resources are underutilized. However, after performing load balancing, an almost equal load is distributed among the *RATs*. Load balancing overcomes both overload and underload conditions in the heterogeneous wireless access network, as can be seen from Figures 4.8, 4.9, and 4.10.

4.4 Summary

In this chapter, the proposed load balancing scheme in a heterogeneous wireless environment efficiently utilizes the available radio resources by maintaining an equal load ratio

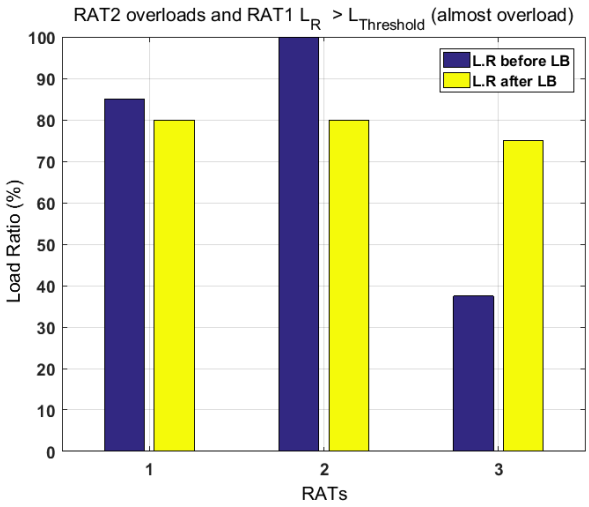

FIGURE 4.10: RAT BW utilization (Scenario III)

across all the *RATs*. Overloaded *RATs* have poor QoS due to an increase in call blocking and dropping, whereas underutilized *RATs* have low revenues as well as poor usage of the radio resources. Our proposed load balancing mechanism tackles the imbalance condition by resolving both overload and underload conditions. The performance of the proposed load balancing mechanism is evaluated in cellular cell layout with overlapping regions, a step towards a more realistic approach. Call blocking probability is used as a tool to analyze the performance of our proposed scheme. The *RATs* BW utilization graphs depict the effectiveness of the proposed scheme.

Chapter 5

Hybrid Radio Resource Management for Time-Varying 5G Heterogeneous Wireless Access Network

In practice, the mobility of the users causes fluctuation in the wireless channel conditions. In this chapter, our objective is to explore radio resource management for a time-varying 5G heterogeneous wireless access network that includes multi-RATs such as 5G new radio (NR) and long-term evolution (LTE). We incorporate congestion control management with radio resource allocation and RAT selection. However, the joint solution has challenges of signalling overhead and computational complexity. Therefore, it is required to decompose the process of radio resource management. We explore the RAT selection scheme that is performed by each user device with network assistance. Our objective is to maximize the average throughput utility subject to admission control and resource allocation. Radio resource allocation and congestion control are formulated as stochastic optimization problems. Using the Lyapunov optimization, we can decompose our radio resource management into two subproblems 1) optimal radio resource allocation and 2) congestion control. Radio resource allocation policy implemented at the central controller node allocates resources at each time slot using the Lagrange dual method. In contrast, congestion control is carried out at the user end based on throughput adaptation according to its current channel conditions. Therefore, it is crucial to investigate the theoretical and simulation results to evaluate the performance of our proposed approach under the assumption of network stability. Furthermore, it is desirable to depict our proposed scheme's effectiveness by simulating the individual user throughput and queue length and comparing the performance of the equal power and adaptive power allocation technique. Moreover, we need to compare the performance of our proposed RAT selection scheme with the traditional centralized and distributive mechanisms.

5.1 Background and Introduction

The 5G heterogeneous wireless access network (HWAN) is proposed as a prospective solution to tackle the high data rate demand. The International Telecommunication Union (ITU) categorized diverse 5G into three main groups to support services such as 1) enhanced mobile broadband (eMBB), 2) massive-machine type communication (mMTC), and 3) ultrareliable low-latency communication (URLLC) [74]. 5G new radio (NR) interface with multi-connectivity features is expected to support these services by connecting to different RATs in a heterogeneous environment [2]. This multi-RAT interwork is based on Long Term Evolution (LTE) dual connectivity introduced by LTE release 12 [76], approved by the 3rd generation partnership project (3GPP) [77]. 5G re-defines the concept of radio resource management, i.e., the scope of radio resource management is not only limited to the allocation of radio resources, but it considers the role of central controller node and radio interfaces. Therefore, it is essential to design efficient algorithms that satisfy application requirements related to bandwidth and quality of experience to efficiently use the 5G services. 5G HWAN composed of wireless networks such as 3GPP LTE and 5G NR base station is a solution to accommodate a large number of connections and high data rates by enabling multihoming features with coverage of overlapping zones and unprecedented throughput gains as compared to the traditional homogeneous wireless access network (LTE only or 5G NR only) [78].

Practical wireless radio networks support mobile users and operate at a time-varying channel condition. It is, therefore, crucial to provide smooth services over fluctuating radio channel conditions. For a stable 5G HWAN, it is essential to maintain system stability by avoiding traffic congestion. Throughput adaptation over time is proposed in order to cope with the time-varying nature of the radio channel. In the context of traffic congestion control, throughput can be defined as the rate of traffic admission to the 5G HWAN. Throughput adaptation allows a user to dynamically adapt its throughput level depending on its current channel condition, e.g. during a service session, a mobile user with better channel conditions selects high throughput, whereas a lower throughput is selected due to worse channel conditions. Better throughput provides better user satisfaction. However, it is possible that the available radio resources at 5G HWAN may not support the high throughput and may cause interruptions in the session, which further degrades the end-user experience.

The recent literature survey reveals that authors in [79] explored the algorithm for joint quality adaptation and rate allocation in orthogonal frequency division multiple access (OFDMA) based system for single home connections only. The rate adaptation for the device-to-device (D2D) communication has been investigated in [80], where the scheduling algorithm allows selecting different quality levels. Furthermore, in [79], the rate adaptation decision for all users is made at the network end, whereas in [80], the rate adaptation decision is made at

the transmitter end. Radio resource management's centralized solution involves high computational complexity and signalling overhead as it requires global information. It is further investigated that the centralized approach has limited spectral efficiency due to capacity and delay constraints [81].

Radio resource management for multihoming connectivity in a heterogeneous wireless network (composed of wireless local area network and cellular network) has been investigated in [82], where a centralized entity performs the joint resource allocation and quality adaptation. They implemented the Lyapunov technique to develop a quality-aware streaming algorithm. In [83], multi-RAT selection and radio resource allocation formulated as an optimization problem subject to the quality of service (minimum data rate) requirement of its connected users is studied. To overcome signalling overhead in a centralized approach, they proposed a distributive system for RAT selection. However, their work does not include optimal power and resource block allocation. Radio resource allocation for a multiuser environment has been formulated in [84], where the cognitive base station utilizes both the licensed cellular and primary user bands for transmission. More recently, the authors in [85] explored radio resource allocation and adaptive bitrate quality adaptation for orthogonal frequency division multiple access cellular networks. However, their work lacks the concept of multiconnection. In [14], the authors proposed RAT selection based on multi-criteria related to user and network. These decision-making parameters include received signal strength, network load, mobility of user and throughput. However, this RAT selection does not involve time-varying channel conditions. The authors in [15] explored radio resource allocation in a heterogeneous wireless access network composed of LTE base station and wireless local area network access points. It is related to optimal radio resource allocation, i.e. subcarrier, power, and time fraction allocation. The network utility maximization involves user quality of service as one of the main constraints. However, this work lacks time-varying channels and congestion control, which is related to throughput adaptation.

We propose radio resource management for OFDMA based 5G heterogeneous wireless access network that includes RAT selection, resource block and power allocation, and traffic congestion control. We decomposed radio resource management into three main parts, RAT selection, radio resource allocation, and congestion control. RAT selection is performed at the user end with the network assistance, whereas radio resource allocation combined with traffic congestion control is formulated by maximizing network utility subject to network-related constraints. We propose a hybrid solution for this problem by decomposing it into two components; the centralized part that performs radio resource allocation and the distributive part that relates to throughput adaptation. In our proposed approach, the central controller node (which can be a centralized entity or cloud) allocates radio resources (resource block and power) to the mobile users. In contrast, each user individually performs throughput adaptation based on the allocated radio resources.

The main contributions of this chapter are listed below:

1. Design of a mechanism for RAT selection performed by each user with network assistance per time slot.

2. Using the stochastic optimization problem formulation to maximize the network utility, which is the maximization of the time-averaged user's throughput.

3. Development of an online and simple solution based on Lyapunov optimization technique [86], [87] to decompose the joint problem of rate allocation, which is incorporated by congestion control into sub-policies such that congestion control policy takes place at user end and transmission rate allocation at central controller node that requires channel state information and queue state information. This leads us to the development of a hybrid algorithm where the central controller node performs radio resource allocation, and congestion control takes place at the user end.

4. Radio resource allocation algorithm for the allocation of resource blocks and power is based on the Lagrange duality approach and multiplier update technique.

This chapter is organized as follows. The system model is given in Section 5.2. The problem is formulated in Section 5.3, where Lyapunov optimization and problem decomposition are presented. Algorithms for RAT selection, radio resource allocation and congestion control are presented in Section 5.4. Performance analysis of the proposed algorithm is presented in Section 5.5. Simulation results are given in Section 5.6. Finally, the chapter's conclusion is presented in Section 5.7.

5.2 System Model

We consider a 5G HWAN shown in Fig. 5.1 [10]. It consists of LTE macro base station (BS) located at the center of the macrocell overlaid by S small cells of 5G NR BSs, where 5G NR BSs are grouped as a set $\mathcal{S} = \{1, 2, ..., S\}$. All these BSs of 5G HWAN are connected to the central controller node (CCN). 5G HWAN provides multihoming services. User devices (UDs) are multi-modal, where they can connect to more than one network simultaneously. We assume K users that are uniformly distributed in the coverage of HWAN and are grouped as a set $\mathcal{K} = \{1, 2, ..., K\}$. Each UD has a specific set of BSs of the available RATs depending on its macrocell location.

We consider resource blocks (RBs) as the minimum radio resource allocation unit in our proposed 5G HWAN. In traditional 4G multi-RAT HWAN (comprising WLAN/LTE), physical layer RB is defined differently in various RATs. Whereas in OFDM-based 5G HWAN, both 4G LTE and 5G NR maintain the same description for physical RB. Each RB has 12 subcarriers in the frequency domain and 14 symbols in the time domain as per the agreement of 3GPP for 5G NR. Moreover, 5G NR allows scaling of the subcarrier spacing, i.e. 15KHz $\times 2^q$, where $(q \in \{0, 1, .., 4\})$. Furthermore, 5G NR uses different transmission time intervals (TTI)

per 3GPP agreement. We assume no intra-cell interference as both LTE BS and 5G NR BSs maintain a separate set of RBs, i.e. M_l and M_{5G} are the set of RBs at LTE BS and 5G NR, respectively. It is assumed that users associated with 5G NR receive signals from multiple 5G NR BSs on allocated RBs. Furthermore, we assume that 5G NR allocates orthogonal RBs to different UDs, which leads to no inter-5G NR interference. Our proposed network operates in a time-slotted manner indexed t of duration t_s. We consider a slow time-varying scenario where users are mobile. The channel conditions are assumed to be changing at a time scale of $t_s = 0.1$ sec. During this time interval, the BSs and user devices share the real-time QSI and CSI. Furthermore, the BSs update the CCN about the QSI and CSI of all users. For implementing radio resource allocation and throughput adaptation decisions at CCN and user devices, we assumed a time slot of a duration of 0.1 sec.

5.2.1 Radio Resource Allocation in 5G HWAN

We consider a time-variant 5G HWAN with an LTE BS and 5G NR BSs. It is, therefore, essential to describe the downlink resource allocation from an OFDMA based system. In an OFDMA based 5G HWAN, we consider W as RB bandwidth, whereas B is the system bandwidth. A binary indicator $x_{nk}(t)$ is used for the allocation of RB n to user k from LTE BS, whereas $x_{mk}(t)$ indicates the allocation of RB m to user k from 5G NR at time slot t. The maximum transmit rate of mobile user k from LTE BS and 5G NR BSs on RB n and m, respectively, is approximated by Shannon theorem given as follows:

$$r^{l}_{nk}(t) = x_{nk}(t)\ W \log_2\left(1 + p_{nk}(t) g_{nk}(t)\right), \tag{5.1}$$

$$r^{5G}_{mk}(t) = x_{mk}(t)\ W \log_2\left(1 + \sum_{i \in \mathcal{S}} p_{imk}(t) g_{imk}(t)\right), \tag{5.2}$$

where $p_{nk}(t)$ and $p_{imk}(t)$ describe the transmit power for the link between user k and RB n from LTE BS, and between user k and RB m from 5G NR BS i, respectively. $g_{nk}(t)$ and $g_{imk}(t)$ describe the channel state information (CSI) of the link between user k and RB n from the LTE BS, and between user k and RB m from 5G NR BS i, respectively. It should be noted that the CSI includes antenna gain, path loss, noise, fast and shadow fading, and it is assumed to be constant over the TTI, with no correlation between the channel coefficients. The CSI is assumed independent and identically distributed (i.i.d.) over the time slots. The total transmit rate obtained by user k from the LTE BS and 5G NR BSs at time slot t is given by

$$R^{l}_{k}(t) \ = \ \sum_{n \in M_l} r^{l}_{nk}(t), \tag{5.3}$$

$$R^{5G}_{k}(t) = \sum_{m \in M_{5G}} r^{5G}_{mk}(t). \tag{5.4}$$

The total transmit rate received by user k from LTE BS and 5G NR BSs at time slot t is given by

$$R_k(t) = R_k^l(t) + R_k^{5G}(t). \tag{5.5}$$

We further modify our problem since mobile users are multihoming, where the CCN allocates the radio resources such that mobile users can get services either from both LTE BS and 5G NR BSs per scheduling time slot or from the BS of a single RAT, i.e. either LTE BS or 5G NR BS per scheduling time slot t. It is important to associate user with either a single RAT (LTE BS or 5G NR BS) or both per time slot t, we therefore, set a network association variable $\rho_k^j(t)$ which is related to the efficiency of RAT j for user k such that $\rho_k^j(t) > 0$, if user k is associated with RAT j, otherwise $\rho_k^j(t) = 0$. The modified Eq. (5.5) is given by

$$R_k(t) = \sum_{j \in J} \rho_k^j(t) R_k^j(t) \tag{5.6}$$

where $j \in J$ corresponds to the available RATs in the 5G HWAN, i.e. $J = \{l, 5G\}$, $\rho_k^j(0 \leq \rho_k^j(t) \leq 1)$. Multihoming is allowed as $\sum_{j=1}^{J} \rho_k^j(t) \leq 1, \forall\, k,\ t$. Therefore, we can conclude the following condition for user association per time slot,

$$\begin{cases} \text{Case I:} & \text{if } \rho_k^{j'}(t) > 0 \text{ and } \rho_k^j(t) > 0 \text{ where } j' \neq j \\ \text{Case II:} & \text{if } \rho_k^{j'}(t) > 0 \text{ and } \rho_k^j(t) = 0 \text{ where } j' \neq j \end{cases} \tag{5.7}$$

Eq. (5.6) is further modified as

$$R_k(t) = \rho_k^l(t) R_k^l(t) + \rho_k^{5G}(t) R_k^{5G}(t). \tag{5.8}$$

The evaluation of user association index $\rho_k^j(t)$ is further explained in detail in Section 5.4.2. The time-averaged transmit rate of user k in the HWAN is given by

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{R_k(t)\}. \tag{5.9}$$

The total transmit rate of the 5G HWAN is given by

$$R_{tot}(t) = \sum_{k \in \mathcal{K}} R_k(t). \tag{5.10}$$

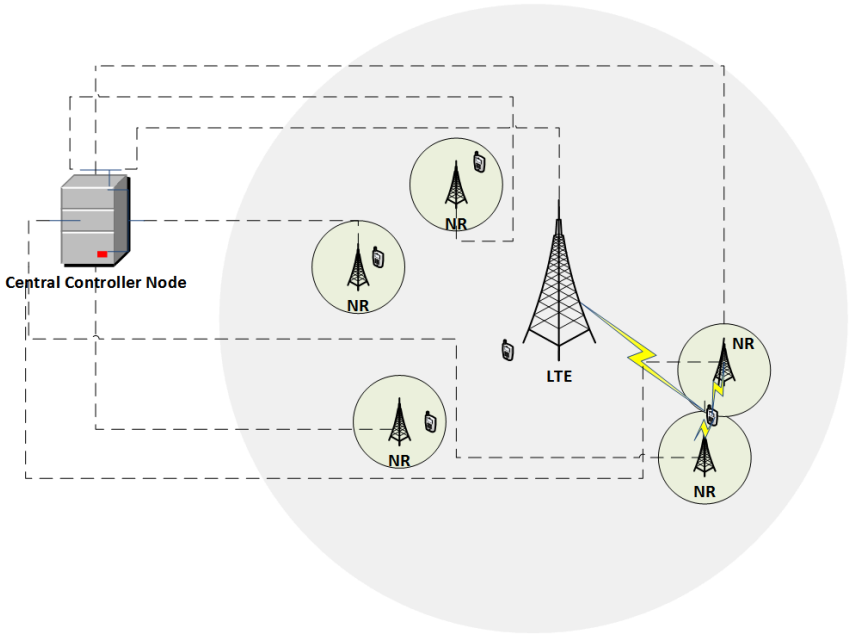

FIGURE 5.1: 5G Heterogeneous wireless access network layout.

The time-averaged transmit rate of the system is given by

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{R_{tot}(t)\}. \tag{5.11}$$

The transmit power allocation for LTE BS and 5G NR BS i at time slot t is given as

$$p^l(t) = \sum_{k\in\mathcal{K}} \sum_{n\in M_l} x_{nk}(t) p_{nk}(t), \tag{5.12}$$

$$p_i^{5G}(t) = \sum_{k\in\mathcal{K}} \sum_{m\in M_{5G}} x_{mk}(t) p_{imk}(t). \tag{5.13}$$

In our research, we do not incorporate power consumption related to circuit power and static power in 5G HWAN BSs, i.e. LTE BS and 5G NR BSs.

5.2.2 Transmission Buffer Dynamics and Stability

5G HWAN maintains a transmission buffer for each mobile user k. The time dynamics of the transmission buffer for user k is given by

$$Q_k(t+1) = \max[Q_k(t) - R_k(t)t_s, 0] + D_k(t), \tag{5.14}$$

where $R_k(t)t_s$ is the channel transmission rate that corresponds to the dequeuing process, i.e. the number of bits transmitted to user k, whereas $D_k(t)$ is the amount of bits placed in the queue and corresponds to the enqueue process of buffer for user k at time slot t. Let $A_k(t)$ be the random traffic arrival rate (Poisson distribution) for user k at time slot t. It is assumed that $A_k(t)$ is i.i.d. over time slot t and independent with respect to k. Furthermore, $A_k(t) \leq A_k^{max}(t)$, where $A_k^{max}(t)$ is the maximum traffic arrival rate. For stable queues and congestion free system it is crucial to maintain the range $0 \leq D_k(t) \leq A_k(t)$, for enqueue process of transmission queue of each user k at time slot t, since $D_k(t)$ is the amount of data obtained from the possibly extensive traffic arrival $A_k(t)$. The individual buffer $Q_k(t)$ is stable if it does not grow infinitely large over time such that

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{Q_k(t)\} < \infty. \tag{5.15}$$

The individual queue is stable if the following condition is true, i.e. the time-averaged dequeue process is equal to or greater than the en queue process.

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{R_k(t)\} \geq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{D_k(t)\}. \tag{5.16}$$

If all the individual queues in the network are stable, only then is a network considered as stable [86]. Furthermore, from the above discussion, we can conclude that the average throughput achieved by user k is the time-averaged enqueue rate, and is given as $\overline{d}_k = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (D_k(t))$. The average delay of user k for a given traffic arrival rate is related to its queue length since the average delay is the ratio of average queue length and average throughput.

5.3 Problem Formulation

To demonstrate our hybrid congestion control and radio resource allocation (HCCRRA) approach, we formulate the average throughput by a concave decreasing utility function, which is used to characterize the joint problem for congestion control and radio resource allocation and is given as follows:

$$U(\overline{\mathrm{d}}) = \sum_{k\in\mathcal{K}} h_Q(\overline{d_k}), \tag{5.17}$$

where $\overline{\mathrm{d}} = [\overline{d_k} : k \in \mathcal{K}]$ is the vector describing the average throughput of all users in the system, and $h_Q(.)$ describes the nondecreasing concave function with α fairness where $\alpha \in (0, \infty)$. For $\alpha = 1$, $h_Q(x) = \log(x)$ whereas for $\alpha \neq 1$ and $\alpha > 0$, $h_Q(x) = (1-\alpha)^{-1}x^{1-\alpha}$.

A higher value of α corresponds to a more fair throughput adaptation, i.e. $\alpha \to \infty$ produces max-min fairness, whereas $\alpha = 0$ corresponds to no fairness, and $\alpha = 1$ corresponds to proportional fairness in throughput adaptation [88].

The HCCRRA stochastic optimization problem P1 is given as follows:

$$\begin{aligned}
\text{P1}: \max_{X,P,d} \quad & U(\overline{\mathrm{d}}), && \text{(5.18a)}\\
\text{s.t.} \quad & \text{C1}: p^{l}(t) \leq p^{l}_{max}, \quad \forall t && \text{(5.18b)}\\
& \text{C2}: p^{5G}_{i}(t) \leq p^{5G}_{i,max}, \quad \forall i,t && \text{(5.18c)}\\
& \text{C3}: \sum_{k\in\mathcal{K}} x_{nk}(t) \leq 1, \quad \forall n,k,t && \text{(5.18d)}\\
& \text{C4}: \sum_{k\in\mathcal{K}} x_{mk}(t) \leq 1, \quad \forall m,k,t && \text{(5.18e)}\\
& \text{C5}: x_{nk}(t), x_{mk}(t) \in \{0,1\}, \ \forall n,m,k,t && \text{(5.18f)}\\
& \text{C6}: \ D_{k}(t) \leq A_{k}(t), \quad \forall k,t && \text{(5.18g)}\\
& \text{C7}: \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{Q_{k}(t)\} < \infty, \ \forall k, && \text{(5.18h)}
\end{aligned}$$

where $X = [x_{nk}(t), x_{mk}(t) : k \in \mathcal{K}]$, and $P = [p_{nk}(t), p_{imk}(t) : n \in M_{l}, m \in M_{5G}, i \in \mathcal{S}, k \in \mathcal{K}]$ are the vectors representing the radio resource allocation policy of the 5G HWAN, whereas the vector $d = [D_{k}(t) : k \in \mathcal{K}]$ represents traffic arrival policy. Constraints C1 and C2 represent the total power constraint from 5G HWAN for LTE BS and 5G NR BSs, respectively. C3 and C4 describe the RB allocation constraint for LTE BS and 5G NR BSs. The physical definition of $x_{nk}(t)$ and $x_{mk}(t)$ is approximated by constraint C5. Constraint C6 is related to the enqueue process. This enqueue process is less than the arrival rate. C7 shows the buffer stability of the 5G HWAN.

The proposed optimization problem P1 is an offline problem, and its theoretical solution is possible if information related to CSI is known. The direct solution of problem P1 using the Markov decision process (MDP) has very high computational complexity, and the system state space grows exponentially to the number of users. Furthermore, it is impractical to know the CSI in advance in a real wireless access network. From a practical perspective, it is crucial to overcome these challenges. Therefore, we formulate an online algorithm by borrowing the concept of Lyapunov optimization, which translates P1 into an online optimization problem and enables real-time decisions that require only current information related to buffer states and channel states. It decomposes problem P1 into optimization policies related to radio rate allocation and admission control. CCN performs the rate allocation, and congestion control is achieved by mobile users adapting its throughput rate according to channel conditions, thus transforming a joint rate allocation and congestion problem into the HCCRRA problem.

5.3.1 Problem Transformation

To solve the optimization problem P1 using stochastic optimization techniques [86], we need to transform it into an equivalent optimization problem that maximizes the single time-averaged utility function. Auxiliary variables $\gamma_k(t)$ and virtual queues $\Gamma_k(t)$ are introduced to obtain the transformation of problem P1 to P2. Transformed problem P2 is given as follows:

$$\text{P2}: \max_{X,P,d,\gamma} \quad \overline{U(\gamma)}, \tag{5.19a}$$

$$\text{s.t.} \quad \text{C1} - \text{C7} \tag{5.19b}$$

$$\text{C8}: \overline{\gamma_k} \leq \overline{d_k}, \quad \forall k \tag{5.19c}$$

$$\text{C9}: \gamma_k(t) \leq A_k^{max}(t), \quad \forall k, t \tag{5.19d}$$

where $\gamma = [\gamma_k(t) : k \in \mathcal{K}]$, $U(\gamma(t)) = \sum_{k\in\mathcal{K}} h_Q(\gamma(t))$ with α fairness, and $\overline{\gamma_k} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1}(\gamma_k(t))$. The following lemma shows that optimization problem P1 is equivalent to the transformed optimization problem P2.

Lemma 4.1: Problem P1 is equivalent to P2.

Proof: See Appendix A.1.

The virtual queues stability is presented by constraint C8. C9 is related to the enqueue process of the virtual queue. Virtual queues evolution dynamics are given as follows:

$$\Gamma_k(t+1) = \max[\Gamma_k(t) - D_k(t), 0] + \gamma_k(t), \tag{5.20}$$

where $\gamma_k(t)$ and $D_k(t)$ are related to the enqueue process and dequeue process of the virtual queue, respectively. Constraint C8 is satisfied if the virtual queue is stable. We introduced the auxiliary variable $\gamma_k(t)$ to represent the enqueue rate of the virtual queues. We are using drift minus reward method to formulate our offline problem into an online problem. Drift-minus-reward method introduces virtual queues $\Gamma_k(t)$ to enforce the desired time average conditions.

To solve problem P2, let $\Theta(t) = [Q_k(t), \Gamma_k(t)] : k \in \mathcal{K}]$ represent the vector of transmission queues and virtual queues, respectively.The Lyapunov drift and Lyapunov optimization are powerful tools for optimizing time-averages in stochastic queuing networks subject to stability. We consider the following quadratic Lyapunov optimization function

$$L(\Theta(t)) = \frac{1}{2}\left[\sum_{k\in\mathcal{K}} Q_k^2(t) + \sum_{k\in\mathcal{K}} \Gamma_k^2(t)\right]. \tag{5.21}$$

The Lyapunov drift function $\Delta(\Theta(t))$ quantifies the difference in the Lyapunov function at time interval $[t, t+1]$.

$$\Delta(\Theta(t))=\mathbb{E}\{L(\Theta(t+1))-L(\Theta(t))|\Theta(t)\}. \tag{5.22}$$

Using Eq. (5.14) and Eq. (5.20), we get the following expression for the Lyapunov drift function

$$\begin{aligned}\Delta\Theta(t) \leq\ & Z-\sum_{k\in\mathcal{K}}\mathbb{E}\Big[Q_k(t)\{R_k(t)t_s-D_k(t)\}|\Theta(t)\Big]\\ & -\sum_{k\in\mathcal{K}}\mathbb{E}\Big[\Gamma_k(t)\{D_k(t)-\gamma_k(t)\}|\Theta(t)\Big],\end{aligned} \tag{5.23}$$

where $Z=\frac{1}{2}\mathbb{E}\Big[\sum_{k\in\mathcal{K}}\Big(t_s^2R_k^2(t)+2D_k^2(t)+\gamma_k^2(t)\Big)|\Theta(t)\Big]$. To obtain separate subproblems for rate allocation and rate adaptation optimization policies, we need to further manipulate the above Eq. (5.23). To get the drift-minus-reward function, we subtract the term $\Lambda = V\mathbb{E}\{U(\gamma(t))|\Theta(t)\}$ from both sides of Eq. (5.23), we have

$$\begin{aligned}\Delta\Theta(t)-\Lambda \leq Z &-\mathbb{E}\left[\sum_{k\in\mathcal{K}}\{\Gamma_k(t)-Q_k(t)\}D_k(t)|\Theta(t)\right]\\ &-\mathbb{E}\left[\sum_{k\in\mathcal{K}}\{Vh_Q(\gamma_k(t))-\Gamma_k(t)\gamma_k(t)\}|\Theta(t)\right]\\ &-\mathbb{E}\left[\sum_{k\in\mathcal{K}}Q_k(t)R_k(t)ts|\Theta(t)\right],\end{aligned} \tag{5.24}$$

where V is a non-negative policy control parameter. To optimize the policies related to congestion control and radio resource allocation, minimize the non-constant part of the right-hand side of Eq. (5.24) at each time slot t. Having online information related to $Q_k(t)$, $\Gamma_k(t)$ and CSI, we can decompose Eq. (5.24) into multiple subproblems. The second term on the right-hand side of Eq. (5.24) is related to the congestion control policy for the incoming traffic. The third term on the right-hand side of Eq. (5.24) is associated with the network utility maximization of each user based on auxiliary variables. The fourth term on the right-hand side of Eq. (5.24) is a function of the scheduling policy, i.e. radio resource allocation. Using the Lyapunov optimization technique, we present our HCCRRA such that the congestion control policy is carried out at each user, and the radio resource allocation process takes place at CCN for the available BSs of LTE and 5G NR as follows. Fig. 5.2 depicts hybrid congestion control and radio resource allocation architecture. The centralized radio resource allocation policy takes CSI from the physical layer and observes the QSI of all users. In contrast, the

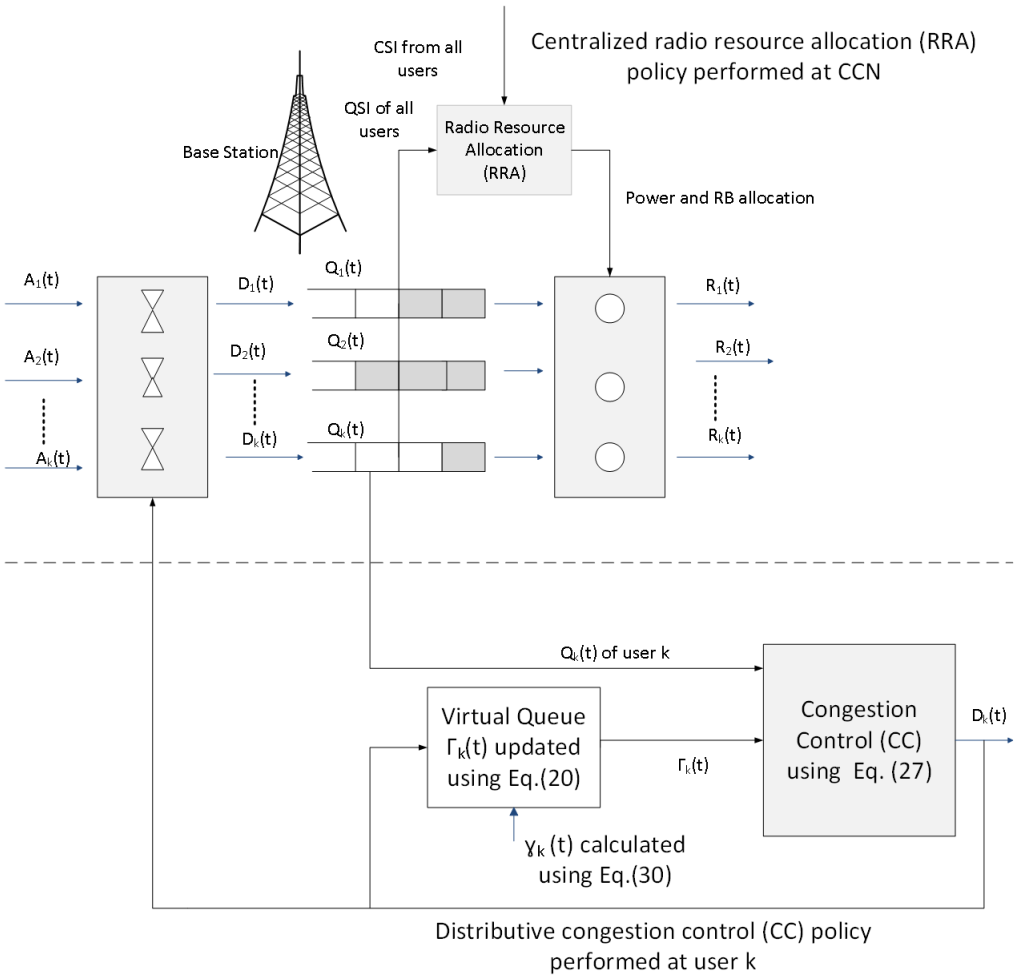

FIGURE 5.2: Hybrid Congestion Control and Radio Resource Allocation architecture.

distributive congestion control policy, implemented at each user end, receives its QSI from the BS.

5.3.1.1 Congestion Control Optimization Policy Derivation

The second term on the right-hand side of Eq. (5.24) is related to the congestion control decision and is given as

$$\max_{d} \quad \sum_{k \in \mathcal{K}} \{\Gamma_k(t) - Q_k(t)\} D_k(t), \tag{5.25}$$

where $d = [D_k(t) : k \in \mathcal{K}]$. In our proposed work, the congestion control process is a distributed process that is carried out at each user k independently. Therefore, we can decouple

Eq. (5.25) and compute it at each user (say k) given as

$$\begin{aligned} \max_{D_k} \quad & \{\Gamma_k(t) - Q_k(t)\} D_k(t), \\ \text{s.t.} \quad & D_k(t) \leq A_k(t). \end{aligned} \tag{5.26}$$

Let $\eta = \Gamma_k(t) - Q_k(t)$. In order to maintain stable queues and avoid congestion in the network it is desirable to maintain the transmission queue $Q_k(t)$ of user k to be smaller than $\Gamma_k(t)$ such that $\eta > 0$.

$$\begin{cases} D_k(t) = A_k(t), \text{if } \eta > 0 \\ D_k(t) = 0, \text{otherwise.} \end{cases} \tag{5.27}$$

Our congestion control policy is based on queue state information related to $Q_k(t)$ and $\Gamma_k(t)$ as each user chooses the throughput at which data is requested by considering the queue state of its transmission queue and virtual queue. Mobile users learn $Q_k(t)$ information locally from its associated BS. Specifically, each user learns its queue length information in the data packet it receives from its connected BS per time slot t. Each user k uses Eq. (5.20) to update its virtual queue $\Gamma_k(t)$ locally.

The auxiliary variables are derived by considering the third term on the right-hand side of Eq. (5.24) given as

$$\max \quad \sum_{k \in \mathcal{K}} \{V h_Q(\gamma_k(t)) - \Gamma_k(t)\gamma_k(t)\}. \tag{5.28}$$

We can further decouple Eq. (5.28) as the auxiliary variables are independent w.r.t. k. Therefore, the auxiliary variable $\gamma_k(t)$ for each user k is computed by maximizing the decomposed problem, given as

$$\begin{aligned} \max_{\gamma_k(t)} \quad & \{V h_Q(\gamma_k(t)) - \Gamma_k(t)\gamma_k(t)\}, \\ \text{s.t.} \quad & \gamma_k(t) \leq A_k(t). \end{aligned} \tag{5.29}$$

The above problem is a convex optimization problem. We consider a logarithmic utility function with $\alpha = 1$ as logarithmic utility function is concave and nondecreasing. We can solve Eq. (5.29) by differentiating it w.r.t. $\gamma_k(t)$, and equating the result to zero, we get

$$\gamma_k(t) = \min \left[\frac{V}{\Gamma_k(t)}, A_k^{max} \right]. \tag{5.30}$$

With $\eta > 0$, transmission queues are maintained stable by allowing new traffic arrival in the transmission queue of user k. The utility improves by accepting larger data, i.e. increase the throughput $D_k(t)$ at which data is requested. Furthermore, it decreases $\Gamma_k(t)$ such that $\gamma_k(t)$ approaches $D_k(t)$. However, in a situation where channel conditions are poor, $Q_k(t)$ accumulates and approaches $\Gamma_k(t)$, and thus high-quality data acceptance can cause congestion. Therefore, to avoid congestion, $D_k(t)$ is reduced.

5.3.1.2 Radio Resource Allocation Policy Optimization at CCN

The decision of radio resource allocation made by CCN at every scheduling time slot is based on the last term on the right-hand side of Eq. (5.24) and is given as

$$\begin{aligned} \max_{X(t),P(t)} \quad & \sum_{k\in\mathcal{K}} Q_k(t) t_s R_k(t), \\ \text{s.t.} \quad & \text{C1- C5.} \end{aligned} \tag{5.31}$$

The transmission rate $R_k(t)$ is a function of resource block allocation $x_{nk}(t)$ and $x_{mk}(t)$, as well as transmit power $p_{nk}(t)$ and $p_{imk}(t)$ allocation. Our problem is nonconvex MINLP problem as it has binary variables $x_{nk}(t)$ and $x_{mk}(t)$, and continuous variables $p_{nk}(t)$ and $p_{imk}(t)$. The optimal solution using the brute force exhaustive search method is infeasible and impractical as it has a high exponential complexity of the order $O(|\mathcal{K}|^{|M_l|})$ and $O(|\mathcal{K}|^{|M_{5G}|})$, for LTE and 5G NR RAT, respectively. Therefore, to solve the problem we reformulate it in the next section.

5.4 Optimal Radio Resource Allocation

To solve the problem given in Eq. (5.31), we need to relax the binary variables and then apply the Lagrange dual decomposition technique to get the primal optimal solution. Eq. (5.31) can be re-written as

$$\begin{aligned} \max_{X(t),P(t)} \quad & \sum_{k\in\mathcal{K}} Q_k(t) t_s \rho_k^l(t) R_k^l(t) \\ & + \sum_{k\in\mathcal{K}} Q_k(t) t_s \rho_k^{5G}(t) R_k^{5G}(t), \\ \text{s.t.} \quad & \text{C1- C5.} \end{aligned} \tag{5.32}$$

We relax the binary variables $x_{nk}(t)$ and $x_{mk}(t)$ by assigning them continuous values from $[0, 1]$. Furthermore, we introduce auxiliary variables $w_{mk}(t)$ and $w_{imk}(t)$ for each user k and RB n and m such that $w_{nk}(t) = x_{nk}(t)p_{nk}(t)$ and $w_{imk}(t) = x_{mk}(t)p_{imk}(t)$, where $n \in M_l$ and $m \in M_{5G}$. The problem given in Eq. (5.32) can be re-written as

$$
\begin{aligned}
&\max_{X(t),P(t)} \sum_{k\in\mathcal{K}} Q_k(t) t_s\{E+F\},\\
&\text{s.t.} \quad \sum_{k\in\mathcal{K}} x_{nk}(t) \le 1, \quad \forall n \in M_l, t\\
&\qquad \sum_{k\in\mathcal{K}} x_{mk}(t) \le 1, \quad \forall m \in M_{5G}, t\\
&\qquad \sum_{k\in\mathcal{K}} \sum_{n\in M_l} w_{nk}(t) \le p^l_{max}, \quad \forall t\\
&\qquad \sum_{k\in\mathcal{K}} \sum_{n\in M_{5G}} w_{imk}(t) \le p^{5G}_{i,max}, \quad \forall i, t\\
&\qquad w_{nk}(t) \ge 0, \quad \forall n \in M_l, k \in \mathcal{K}, t\\
&\qquad w_{imk}(t) \ge 0, \quad \forall m \in M_{5G}, k \in \mathcal{K}, t\\
&\qquad 0 \le x_{nk}(t) \le 1, \quad \forall n \in M_l, k \in \mathcal{K}, t\\
&\qquad 0 \le x_{mk}(t) \le 1, \quad \forall m \in M_{5G}, k \in \mathcal{K}, t,
\end{aligned}
\tag{5.33}
$$

where E and F are given as

$$
E = \rho^l_k(t) \sum_{n\in M_l} x_{nk}(t)\, W \log_2\left(1 + \frac{w_{nk}(t) g_{nk}(t)}{x_{nk}(t)}\right), \tag{5.34}
$$

$$
\begin{aligned}
F = \rho^{5G}_k(t) &\sum_{m\in M_{5G}} x_{mk}(t)\, W \times\\
&\log_2\left(1 + \sum_{i\in\mathcal{S}} \frac{w_{imk}(t) g_{imk}(t)}{x_{mk}(t)}\right).
\end{aligned}
\tag{5.35}
$$

Both E and F are concave functions, and all the constraints are linear and affine. Therefore, our optimization problem is a convex optimization problem. Using the Lagrange duality decomposition technique, we can solve our problem as follow:

5.4.1 Lagrange Dual Decomposition

The Lagrangian of problem given in Eq. (5.33) is given as

$$\begin{aligned} L(\Phi,\Phi_i) = \max_{X,P} \sum_{k\in\mathcal{K}} Q_k(t) t_s \Bigg\{ & \rho_k^l(t) \sum_{n\in M_l} x_{nk}(t)\, W \\ & \times \log_2\left(1+\frac{w_{nk}(t) g_{nk}(t)}{x_{nk}(t)}\right) \\ & + \rho_k^{5G}(t) \sum_{m\in M_{5G}} x_{mk}(t)\, W \\ & \times \log_2\left(1+\sum_{i\in\mathcal{S}} \frac{w_{imk}(t) g_{imk}(t)}{x_{mk}(t)}\right)\Bigg\} \\ & -\Phi \sum_{k\in\mathcal{K}} \sum_{n\in M_l} w_{mk}(t) + \Phi p_{max}^l \\ & - \sum_{i\in\mathcal{S}} \Phi_i \sum_{k\in\mathcal{K}} \sum_{m\in M_{5G}} w_{imk}(t) \\ & + \sum_{i\in\mathcal{S}} \Phi_i\, p_{i,max}^{5G}, \end{aligned} \tag{5.36}$$

where Φ and $\Phi_i = [\Phi_1, \Phi_2, ..., \Phi_S]$ are vectors representing Lagrange multipliers [60]. Since the radio resource allocation is performed per scheduling time slot, we ignore the index t for simplification. The dual function is given by

$$\begin{aligned} D(\Phi,\Phi_i) = & \max_{X,P} L(\Phi,\Phi_i), \\ \text{s.t.} \quad & \sum_{k\in\mathcal{K}} x_{nk} \le 1, \quad \forall n \in M_l \\ & \sum_{k\in\mathcal{K}} x_{mk} \le 1, \quad \forall m \in M_{5G} \\ & 0 \le x_{nk} \le 1, \quad \forall n \in M_l, k \in \mathcal{K} \\ & 0 \le x_{mk} \le 1, \quad \forall m \in M_{5G}, k \in \mathcal{K}. \end{aligned} \tag{5.37}$$

The dual optimization problem is given as

$$\min_{\Phi,\Phi_i} \; D(\Phi,\Phi_i), \tag{5.38}$$

where Φ and Φ_i are non-negative. Applying Karush-Kuhn-Tucker (KKT) conditions, the relationship between w_{nk} and x_{nk} as well as w_{imk} and x_{mk} can be derived by differentiating Eq. (5.36) w.r.t. w_{nk} and w_{imk}, respectively, and is given as

$$w_{nk}^* = \left[\frac{W\, Q_k\, t_s\, \rho_k^l}{ln(2)\Phi} - \frac{1}{g_{nk}}\right]^+ x_{nk}, \tag{5.39}$$

$$w^*_{imk} = \left[\frac{WQ_k\, t_s \rho_k^{5G}}{ln(2)\Phi_i} - \frac{\sum_{i' \neq i} w^*_{i'mk}\, g_{i'mk} + 1}{g_{imk}} \right]^+ x_{mk}, \tag{5.40}$$

where $[x]^+ = \max\{x, 0\}$. Eq. (5.39) and Eq. (5.40) are based on water-filling algorithm [49], [90]. With CSI and users queue information the CCN allocates transmit power to RB ($n \in M_l$ and $m \in M_{5G}$) from LTE BS and 5G NR BS i, respectively. In LTE, the CSI report back procedure is well standardized, i.e. each user first observes its signal strength and then updates its connected BS. This feature is expected to be available in 5G. Substituting the optimal values w^*_{nk} and w^*_{imk} in Eq. (5.36) results in the following:

$$\begin{aligned} L(\Phi, \Phi_i) = & \max_X \sum_{k \in \mathcal{K}} \sum_{n \in M_l} \Psi_{nk}\, x_{nk} \\ & + \sum_{k \in \mathcal{K}} \sum_{m \in M_{5G}} \Pi_{mk}\, x_{mk}, \\ \text{s.t.} \quad & \sum_{k \in \mathcal{K}} x_{nk} \leq 1, \quad \forall n \in M_l \\ & \sum_{k \in \mathcal{K}} x_{mk} \leq 1, \quad \forall m \in M_{5G} \\ & 0 \leq x_{nk} \leq 1, \quad \forall n \in M_l, k \in \mathcal{K} \\ & 0 \leq x_{mk} \leq 1, \quad \forall m \in M_{5G}, k \in \mathcal{K}, \end{aligned} \tag{5.41}$$

where $\Psi_{nk} = Q_k \rho_k^l W \log_2(1 + p_{nk} g_{nk}) - \Phi p_{nk}$ and $\Pi_{mk} = Q_k \rho_k^{5G} W \log_2(1 + \sum_{i \in \mathcal{S}} p_{imk} g_{imk}) - \sum_{i \in \mathcal{S}} \Phi_i p_{imk}$. This is a linear programming problem. Although x_{nk} and x_{mk} can take values from the constraint set $[0, 1]$, but the optimal solution is at the extreme points of the constraint [15]. It indicates that the optimal solution is still binary. The optimal solution for RB allocation from LTE BS is based on the following expression:

$$\begin{cases} x_{nk} = 1, \text{ if } k = \arg\ \max\{\Psi_{nk} :\ k \in \mathcal{K}\} \\ x_{nk} = 0, \text{ otherwise.} \end{cases} \tag{5.42}$$

The optimal solution for RB allocation from 5G NR to user k is given as follows:

$$\begin{cases} x_{mk} = 1, \ \text{ if } k = \arg\ \max\{\Pi_{mk} :\ k \in \mathcal{K}\} \\ x_{mk} = 0, \ \text{ otherwise.} \end{cases} \tag{5.43}$$

Once the optimal values x_{nk} and x_{mk} are determined, then using Eq. (5.39) and Eq. (5.40) we can determine the optimal transmit power for the assigned RBs since $w_{nk} = x_{nk} p_{nk}$ and $w_{imk} = x_{mk} p_{imk}$.

5.4.2 RAT Selection

In Section 5.2.1, we introduced user association index $\rho_k^j(t)$ which is related to RAT efficiency. This user association is performed in a centralized manner, where CCN associates users to the BSs of available RATs. To satisfy each user without overloading the BS of the selected RAT, CCN makes its decision using the CSI of each user k (where $k \in \mathcal{K}$) and the loading information of each RAT j per time slot t. It increases the cost of signalling overhead and is computationally complex. To cope with these two challenges, we transform this user association problem (centralized problem) into the RAT selection problem (distributive problem), where each user selects the BS of the available RATs. Since $\rho_k^j(t)$ is related to the loading status of BS. Therefore, it is impractical from a network perspective to broadcast BS loading status to the users as it may affect its revenue. However, it is important to avoid RAT overloading. We, therefore, propose that the BSs of the available RATs broadcast the minimum throughput of the user k' (where $k' \neq k$) among its connected users at time slot $t-1$, i.e. $D_{k'}(t-1)$ to the users in its coverage zone. Each user measures the received signal strength RSS from the BS of available RATs in its service zone. This feature has already been standardized in LTE [89] and is expected to be available in 5G. If $RSS(t) \geq RSS_T$, where RSS_T is the threshold value of received signal strength, the user then generates its candidate set of BSs. Upon receiving the lowest throughput rate of a user connected to the candidate BS of RAT j at time slot $t-1$, it determines the selected set of RATs, i.e. single RAT or multi-RAT based on Eq. (5.7). The following equation gives the RAT selection index ρ_k^j:

$$\rho_k^j(t) = U_{RSS(t)} \times U_{D_{k'}(t-1)}, \quad k \neq k', \tag{5.44}$$

where $U_{RSS(t)}$ is the utility of received signal strength of user k at time slot t, whereas $U_{D'_k(t-1)}$ is the utility of the throughput of user k' at time slot $t-1$ [14], and can be determined by using the following equations:

$$U_{RSS(t)} = \begin{cases} 0, \text{ if } RSS(t) \leq RSS_T \\ \mathcal{M}, \text{ if } RSS_T < RSS(t) \leq RSS_{max} \\ 1, \text{ if } RSS(t) > RSS_{max}, \end{cases} \tag{5.45}$$

$$U_{D_{k'}(t-1)} = \begin{cases} 0, \text{ if } D_{k'}(t-1) \leq D_k^{min}(t) \\ e^{-\partial\zeta} \text{ if } D_k^{min}(t) < D_{k'}(t-1) \leq D_k^{max}(t) \\ 1, \text{if } D_{k'}(t-1) > D_k^{max}(t)\,, \end{cases} \tag{5.46}$$

where $\mathcal{M} = \frac{RSS(t)-RSS_T}{RSS_{max}-RSS_T}$, $\zeta = (D_{k'}(t-1) - D_k(t))$. RSS_{max}, $D_k^{min}(t)$, and $D_k^{max}(t)$ represent the maximum received signal strength, minimum throughput and maximum throughput for user k at time slot t, respectively, whereas ∂ is the shape parameter for the utility function given in Eq. (5.46). Adjusting ∂ parameter results in changing the shape of the utility

function. As a result, it precisely captures the sensitivity of the utility to the variation of the criterion (i.e. throughput) [14]. Algorithm 2 describes the process of RAT selection and is given below.

Algorithm 2 RAT Selection

1: At each time slot t, observe the RSS of each BS of candidate RATs such that $RSS(t) \geq RSS_T$ and $U_{RSS(t)} \leftarrow RSS(t)$;
2: At each time slot t, sense the broadcasted lowest transmit rate $D_{k'}(t-1)$ of the user connected to candidate RAT, $U_{D_{k'}(t-1)} \leftarrow D_{k'}(t-1)$;
3: Calculate $\rho_k^j(t)$ using Eq. (5.44);
4: The decision of using multi-RAT and single RAT per time slot t is made using Eq. (5.7).
5: Repeat.

5.4.3 Multiplier Updates

Using the gradient descent method, we can formulate the expression for optimal values of Φ^* and Φ_i^* by differentiating the dual function, and is given as follow:

$$\Phi(v+1) = \left[\Phi(v) + \sigma\left(\sum_{k\in\mathcal{K}}\sum_{n\in M_l} w_{nk} - p_{max}^l\right)\right]^+, \tag{5.47}$$

$$\Phi_i(v+1) = \left[\Phi_i(v) + \beta\left(\sum_{k\in\mathcal{K}}\sum_{m\in M_{5G}} w_{imk} - p_{i,max}^{5G}\right)\right]^+, \tag{5.48}$$

where σ and β are the gradient step size, and v is the iteration index. The convergence is guaranteed as the gradient of Eq. (5.37) satisfies the Lipschitz continuity condition, as it is differentiable and continuous.

Algorithm 3 depicts our overall HCCRRA algorithm that includes the process of radio resource allocation and congestion control, and is also shown in Fig. 5.3. $Q_k(t)$ and $\Gamma_k(t)$ are initialized as $Q_k(0) = 0, \forall k$ and $\Gamma_k(0) > 0, \forall k$. The backlogs of the transmission $Q_k(t+1)$ and virtual queues $\Gamma_k(t+1)$ are updated iteratively at every time slot using Eq. (5.14) and Eq. (5.20), respectively, such that at $t = t+1$, $Q_k(t) = Q_k(t+1)$ and $\Gamma_k(t) = \Gamma_k(t+1)$. These updated $Q_k(t)$ and $\Gamma_k(t)$ values are iteratively used in the algorithm to allocate radio resources and maintain stability of a 5G HWAN.

5.4.4 Complexity Analysis

Our proposed HCCRRA scheme has low computational complexity than the joint radio resource allocation and congestion control scheme, which further emphasizes its feasibility. Our proposed approach implements the radio resource allocation policy at the CCN and congestion control policy at each user end in a distributive fashion. The computation complexity

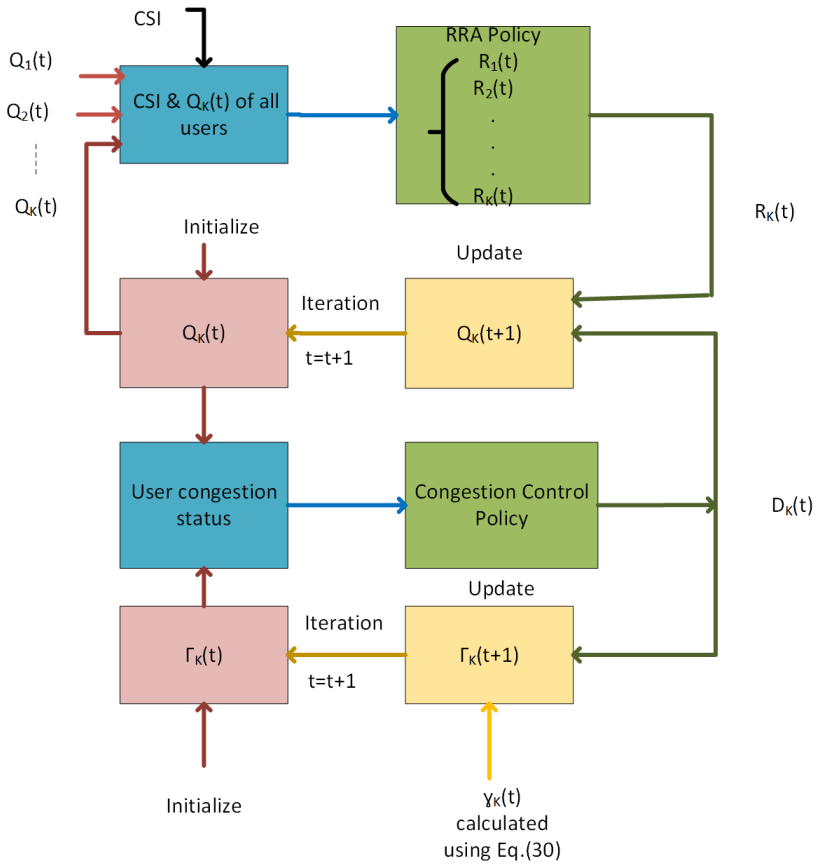

FIGURE 5.3: Hybrid Congestion Control and Radio Resource Allocation Process

Algorithm 3 HCCRRA Algorithm

Step 1: Congestion Control Policy at a user device

Initialize $t \leftarrow 0$, $Q_k(0) = 0$ and $\Gamma_k(0) > 0$;
For each time slot t, observe the transmit queue $Q_k(t)$, CSI $\{g_{nk}(t), g_{imk}(t)\}$, and virtual queue $\Gamma_k(t)$;
Determine the auxiliary variable $\gamma_k(t)$ using Eq. (5.30);
Determine $D_k(t)$ using Eq. (5.27);
Update CCN with the throughput requirement.

Step 2: Radio resource allocation policy at CCN

Initialize $\Phi(0) > 0$ and $\Phi_i(0) > 0$ and iteration index $v \leftarrow 0$, $\epsilon = 0.01$;
Calculate $p_{nk(t)}$ and $p_{imk}(t)$ using Eq. (5.39) and Eq. (5.40), respectively;
Determine $x_{nk}(t)$ and $x_{mk}(t)$ using Eq. (5.42) and Eq. (5.43), respectively;
Update the Lagrange multipliers Φ and Φ_i using Eq. (5.47) and Eq. (5.48), respectively;
If $|\Phi(v+1) - \Phi(v)| \leq \epsilon$ and $|\Phi_i(v+1) - \Phi_i(v)| \leq \epsilon$;
Then $[x_{nk}(t), x_{mk}(t)]$ and $[p_{nk}(t), p_{imk}(t)]$ are optimal;
Else go back to step 6;
$t \leftarrow t + 1$;
Update $Q_k(t)$, and $\Gamma_k(t)$ using Eq. (5.14) and Eq. (5.20), respectively.
Go back to 2 in Step 1

of radio resource allocation policy is $O((|M_l||\mathcal{K}| + |M_{5G}||\mathcal{K}|)(1/\epsilon^2))$, where the number of iterations that achieve convergence accuracy ϵ of Eq. (5.41) is in the order of $1/\epsilon^2$ [60]. The computational complexity of congestion control is $O(1)$ as it is carried out at each user end in a distributive manner. Furthermore, in our proposed HCCRRA algorithm, the CCN requires both CSI and QSI per time slot t to perform radio resource allocation. For the congestion control policy, each user needs to know its corresponding QSI. For CSI updates from users to the BSs, the total signalling overhead on the air interface is $O(|\mathcal{K}|)$ per time slot t, whereas, for QSI updates from the network to the users, the total signalling overhead on the air interface is $O(|\mathcal{K}|)$ per time slot t.

The computation complexity of the RAT selection algorithm is related to the number of users and available RATs. Each user has a different set of candidate RATs depending on its location in the service area. Let J be the available set of candidate RATs for each user k, then $O(|J|)$ is the complexity of Algorithm 2 as each user performs RAT selection in a distributive manner.

5.5 Performance Analysis

In this section, we analyze mathematically the performance bounds (i.e. time-averaged queue backlogs stability of a real-time network and time-averaged optimal utility performance) of our proposed algorithm based on drift minus reward Lyapunov optimization. In practical scenarios, the network state's ergodicity is unlikely to hold due to CSI variations caused by user motion. We, therefore, consider the optimality of the algorithm for an *arbitrary sample path* related to network state $\Omega(t)$ [87]. Let $\Upsilon_{\Omega(t)}$ describe the set of throughput adaptation and resource allocation policies under a given arrival rate and CSI. Theorem 5.1 gives the network stability and optimal utility performance of our proposed algorithm.

Theorem 5.1: In the drift minus reward technique, a non-negative control parameter V is chosen to control the tradeoff between average queue backlog and average network utility. For non-negative V i.e. $V \geq 0$, i.i.d. CSI and arbitrary arrival rate, the network congestion or time-averaged bounded backlog of queues is given as

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Big[\sum_{k\in K} Q_k + \sum_{k\in K} \Gamma_k \Big] \\ \leq \frac{Z + V(U^{max} - U^{min})}{\varrho}, \end{aligned} \tag{5.49}$$

and the utility performance of our proposed algorithm for 5G HWAN is given as

$$U(\bar{d}) \;\geq\; U^* - \frac{Z}{V}, \tag{5.50}$$

where U^* is the optimal utility of the optimization problem P1 shown in Eq. (5.18a-5.18h).
Proof: See Appendix A.2.

Remark 1: Eq. (5.49) of Theorem 5.1, controls congestion by stabilizing the queues, whereas in Eq. (5.50) of Theorem 5.1, we have $U(\bar{d}) \geq U^* - \frac{Z}{V}$. Higher value of V results in very small value of $\frac{Z}{V}$, which leads to the conclusion that average utility is very close to the optimal utility i.e. $U^* \geq U(\bar{d}) \geq U^* - \frac{Z}{V}$. This has been verified in Fig. 5.4. However, Eq. (5.49) shows an increase in delay with an increase in V. Fig. 5.5 verifies this increase in congestion bounds.

Remark 2: In the simulation results, we consider different traffic arrival rates (constant and i.i.d) to analyze the HCCRRA scheme's performance. Our HCCRRA scheme is based on queue lengths only. It does not require knowledge about the traffic arrival rate. Eq. (5.14) shows that queue length is related to the enqueue rate $D_k(t)$, which is the amount of traffic admitted into the system based on the decision made at each user end. Furthermore, to keep the queues stable we have $0 \leq D_k(t) \leq A_k(t)$. We described this in detail in Section 5.2.2. Thus, our proposed approach is independent of the traffic arrival models and is valid for any traffic model.

5.6 Simulation Results

In this section, the proposed hybrid algorithm related to congestion control and radio resource allocation for 5G HWAN is evaluated using Matlab simulations.

5.6.1 Parameter Setting

We consider 5G HWAN with multi-RAT and multihoming features. Our 5G HWAN consists of one macro LTE BS and 5G NR BSs, as shown in Fig. 5.1. The LTE BS is installed at the center of the cell, whereas five 5G NR BSs are randomly overlaid in macro BS coverage. LTE BS provides higher coverage with low bit rates, whereas 5G NR BSs have a low range and higher speeds. We consider 20 MHz as the system bandwidth for LTE with a carrier frequency $f_c = 3.5$ GHz and transmit power of 49 dBm, whereas 5G NR has $f_c = 28$ GHz with system bandwidth of 100 MHz and transmit power of 35 dBm. Since both LTE and 5G NR uses OFDM, their physical layer features remain the same. Both have RBs as the minimum resource allocation unit, composed of 12 subcarriers and 14 OFDM symbols. For LTE the subcarrier spacing is 15 kHz and TTI $= 1$ ms, whereas for 5G NR we assume the subcarrier spacing of 60 kHz and TTI $= 0.25$ ms [91], [92]. Users are assumed to be uniformly distributed in the coverage area. HCCRRA is simulated for different values of control parameter V. Each point of the curve is averaged over 1000 time slots, and the results are obtained with a 95% confidence interval.

5.6.2 Impact of control parameter on network performance

This section highlights the impact of the control parameter V on network performance, such as throughput and delay. Figs. 5.4 and 5.5 show the impact of control parameter on throughput and network congestion with different traffic arrival rates λ. The throughput utility increases to optimal value at a rate of $O(1/V)$ for any arrival rate, whereas the related congestion, approximated by the delay performance, increases linearly with the control parameter, i.e. $O(V)$. Figs. 5.4 and 5.5 indicate the throughput-delay tradeoff, $[O(1/V) - O(V)]$, that validate the theoretical results of Theorem 5.1 given in Eq. (5.49) and Eq. (5.50) respectively. Initially, an increase in throughput can be observed. However, this throughput improvement rate starts to reduce as the value of the control parameter V increases. Furthermore, this increase in V causes congestion as delay linearly increases with V. Thus, the controlling parameter V balances delay and throughput. Larger V results in an increase in throughput but can also increase the delay. Choosing a suitable control parameter V allows 5G HWAN to operate in its ideal state, which refers to a state where a tradeoff between delay and throughput is achieved.

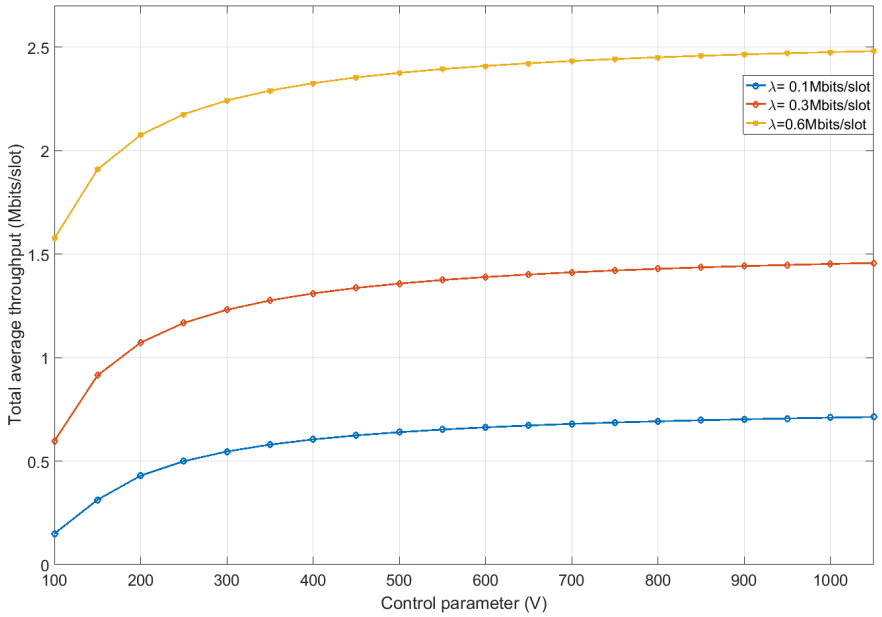

FIGURE 5.4: Total average throughput versus control parameter.

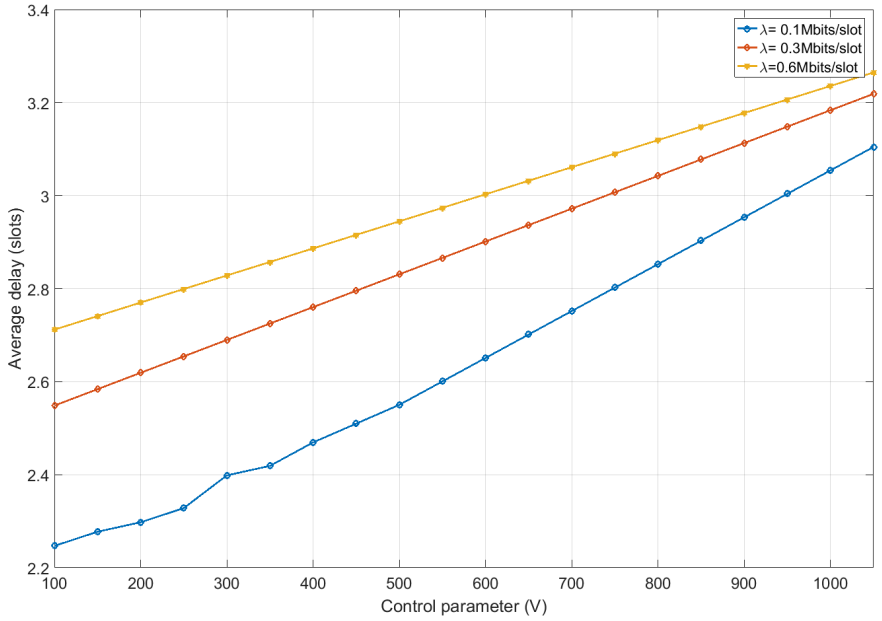

FIGURE 5.5: Average delay versus control parameter.

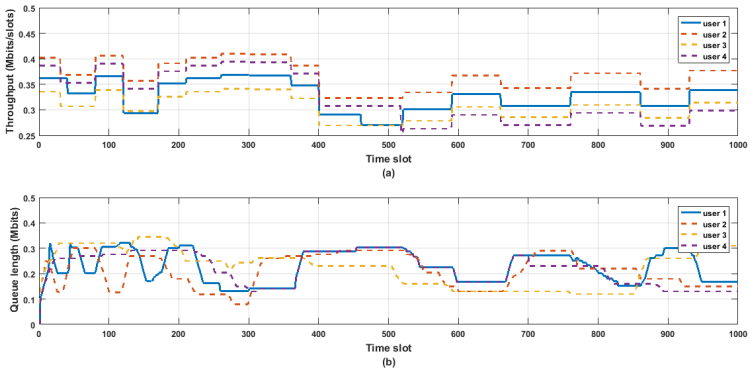

FIGURE 5.6: Performance analysis of individual users using HCCRRA algorithm: (a) Throughput adapatation, (b) Transmission queue length.

5.6.3 Performance analysis of individual users

In Fig. 5.6, we consider the dynamics of throughput adaptation and transmission queues evolution for individual users. We consider 12 users with control parameter $V = 1000$. To show a clear plot, we have shown the dynamics of four users only. The remaining users also enjoy stable queues lengths and throughputs, and they are excluded from the figure. The decision of throughput adaptation is made per time slot, as shown in Fig. 5.6(a). Fig. 5.6(b) shows the dynamics of transmission queues evolution. Fig. 5.6(b) shows that the system is stable as the individual queues are bounded. Thus, Figs. 5.6(a)-(b) show that the algorithm works in real-time since individual users' throughput is optimal and queues are stable for every iteration of a time slot t.

5.6.4 Performance comparison and impact of scenarios on network performance

To validate our proposed HCCRRA scheme's performance, we apply different scenarios on our HCCRRA scheme, 1) 5G HWAN, 2) only 5G NR BSs, and 3) only LTE BS. The system of 5G HWAN is composed of 5G NR BSs and LTE BS, as shown in Fig. 5.1. The scenarios of only 5G NR BSs and only LTE are applied to evaluate the performance comparison. The only 5G NR means that the available users only access the 5G NR BSs, whereas the only LTE BS strategy implies the presence of exclusive LTE BS. Furthermore, we compare our HCCRRA scheme with the maximum sum-rate (MSR) scheme that involves both 5G NR BSs and LTE

BS with no congestion control mechanism. The MSR scheme is modelled as

$$\max_{X(t),P(t)} \quad \sum_{k\in\mathcal{K}} \left[R_k^l(t) + R_k^{5G}(t)\right], \tag{5.51a}$$

$$\text{s.t.} \quad \text{C1}: p^l(t) \leq p^l_{max}, \quad \forall t \tag{5.51b}$$

$$\text{C2}: p_i^{5G}(t) \leq p^{5G}_{i,max}, \quad \forall i,t \tag{5.51c}$$

$$\text{C3}: \sum_{k\in\mathcal{K}} x_{nk}(t) \leq 1, \quad \forall n,t \tag{5.51d}$$

$$\text{C4}: \sum_{k\in\mathcal{K}} x_{mk}(t) \leq 1, \quad \forall m,t \tag{5.51e}$$

$$\text{C5}: x_{nk}(t), x_{mk}(t) \in \{0,1\}, \ \forall n,m,k,t \tag{5.51f}$$

We set the control parameter for the HCCRRA scheme V equal to 1000. Fig. 5.7 shows that initially, the transmit rate is low for all three scenarios, but it increases as the arrival rate increases to maintain stable queues. The transmit rate for 5G HWAN is more than the cases where we implemented only 5G or LTE BS, which is visible and matches with the theoretical background given in Section 5.4. However, the transmit rate for MSR remains stable irrespective of the traffic arrival rate. The reason is clear from the fact that MSR does not consider any traffic arrival rate.

Fig. 5.8 shows the average delay vs. mean traffic arrival rate. Our proposed HCCRRA scheme avoids infinite queue length and maintains network stability by enhancing the transmit rate. Therefore, the proposed HCCRRA scheme does not show a significant increase in delay with an increase in the arrival rate. However, the MSR with no congestion control mechanism has poor delay performance as it does not consider traffic arrival rate and network stability.

In Fig. 5.9, we further investigate the impact of traffic arrival rate on total average power consumption. The MSR scheme has a high power consumption that remains constant with the traffic arrival rate. Since the MSR scheme does not adapt to the traffic arrival rate, as a result, it has high power consumption even at a low arrival rate. From Fig. 5.9, we can visualize that all HCCRRA scenarios, i.e. 5G HWAN, LTE only and 5G NR only, show low power consumption at a low arrival rate. However, the power consumption increases with a higher arrival rate for all scenarios of HCCRRA. HCCRRA maintains stable queues by increasing the transmits rate, which increases power consumption.

5.6.5 Impact of power allocation strategies on network performance

We compare the network performance of our proposed HCCRRA scheme using two power allocation strategies; 1) adaptive power allocation strategy given in Eqs. (5.39) and (5.40), and 2) the equal power allocation strategy. Using suboptimal equal power allocation scheme,

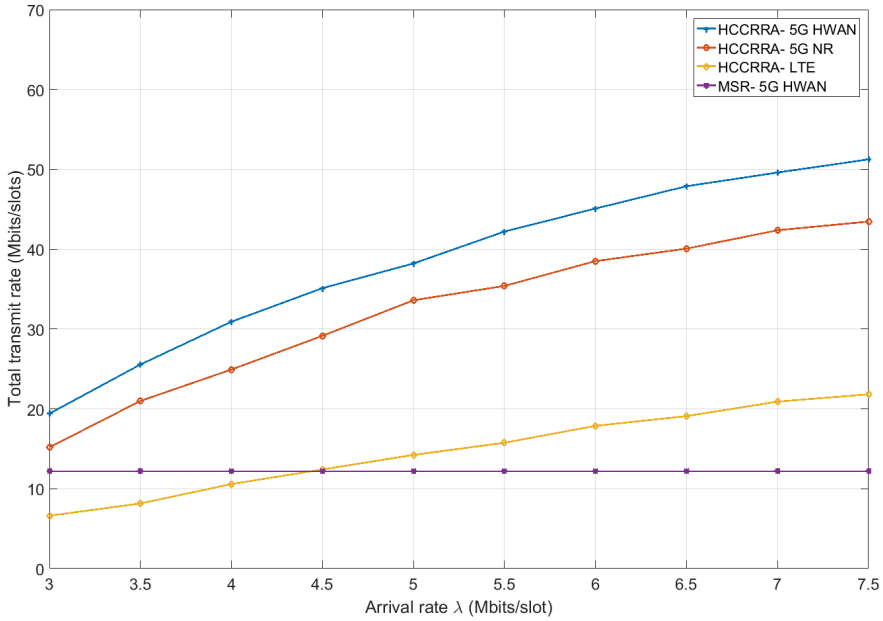

FIGURE 5.7: Total average transmit rate versus traffic arrival rate.

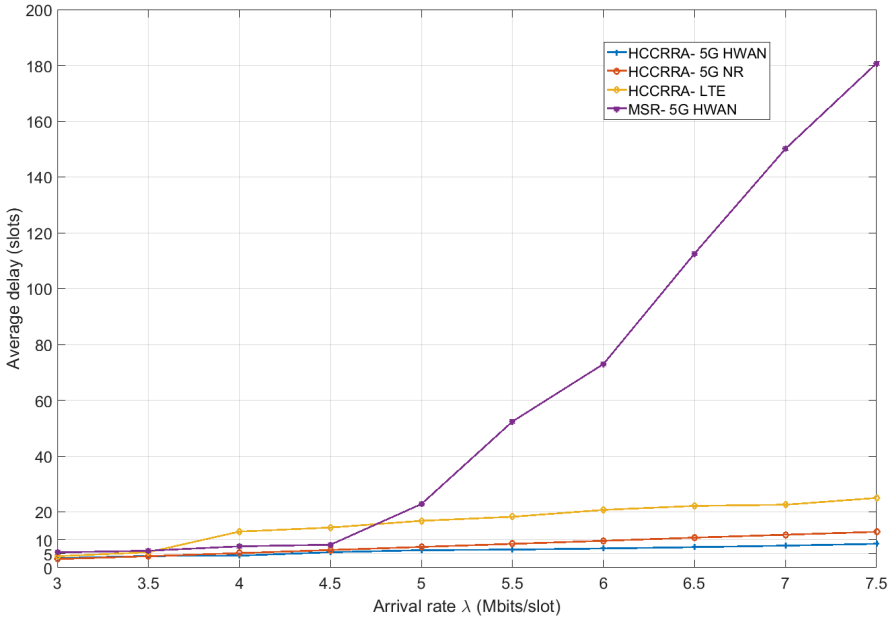

FIGURE 5.8: Delay versus traffic arrival rate.

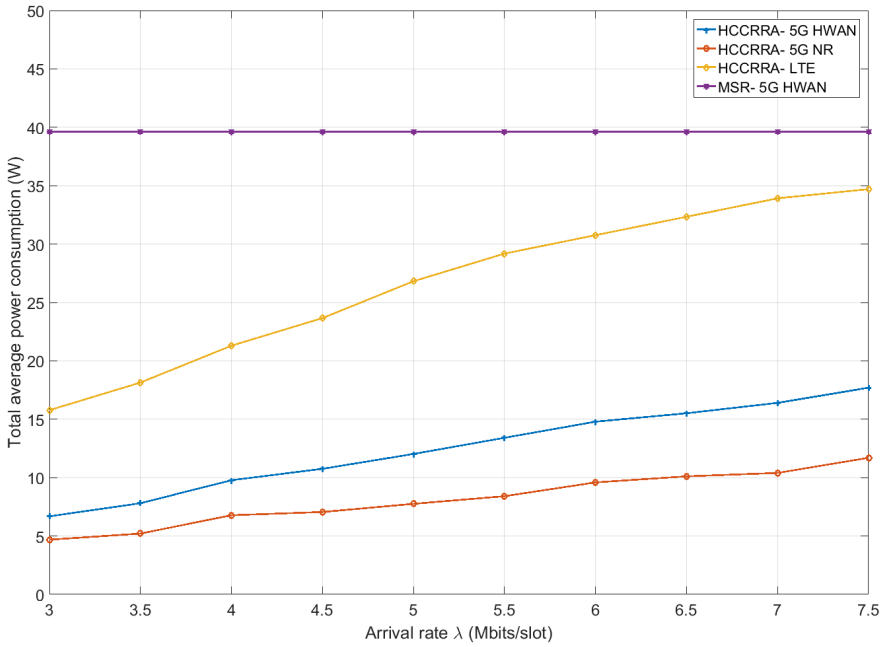

FIGURE 5.9: Average power consumption versus traffic arrival rate.

equal power is allocated per RB, i.e. $p_{nk}(t) = \frac{p^l_{max}}{|M_l|}$ and $p_{imk}(t) = \frac{p^{5G}_{i,max}}{|M_{5G}|}$. RBs allocation is based on maximizing $Q_k(t)\rho^l_k(t)r^l_{nk}(t)$ and $Q_k(t)\rho^{5G}_k(t)r^{5G}_{mk}(t)$ for LTE and 5G NR, respectively. RB selection is related to larger queue size and good channel conditions. We evaluate the network performance, i.e. throughput and delay versus different values of control parameter V. We set the traffic arrival rate as 0.6 Mbits/slot. From Figs. 5.10 and 5.11, we can see that our proposed approach, HCCRRA with dynamic power allocation, outperforms HCCRRA with an equal power allocation scheme. Our proposed optimal power allocation strategy allocates higher power to higher quality channels and results in higher throughput and lower delay.

5.6.6 Fairness analysis

Our proposed congestion control policy is designed with a finite queue length. It maintains stable queues by using the congestion control threshold η, thereby ensuring the Quality-of-Service (QoS) requirements of each user. Therefore, we evaluate the fairness of our proposed HCCRRA scheme using the three scenarios, i.e. 1) 5G HWAN, 2) only 5G NR BSs, and 3) only LTE BS. Fig. 5.12 shows the throughput fairness evaluation of our proposed HCCRRA schemes, where Jain's fairness index JFI $= \frac{\left(\sum_{k\in K}\ \bar{d}_k\right)^2}{|\mathcal{K}|\sum_{k\in K}(\bar{d}_k)^2}$. Furthermore, the control parameter $V = 1000$ and $\alpha = 1$, since larger value of V and $\alpha-$fair utility function gives throughput fairness between the users. From Fig. 5.12, we can see that the 5G NR scenario has lower

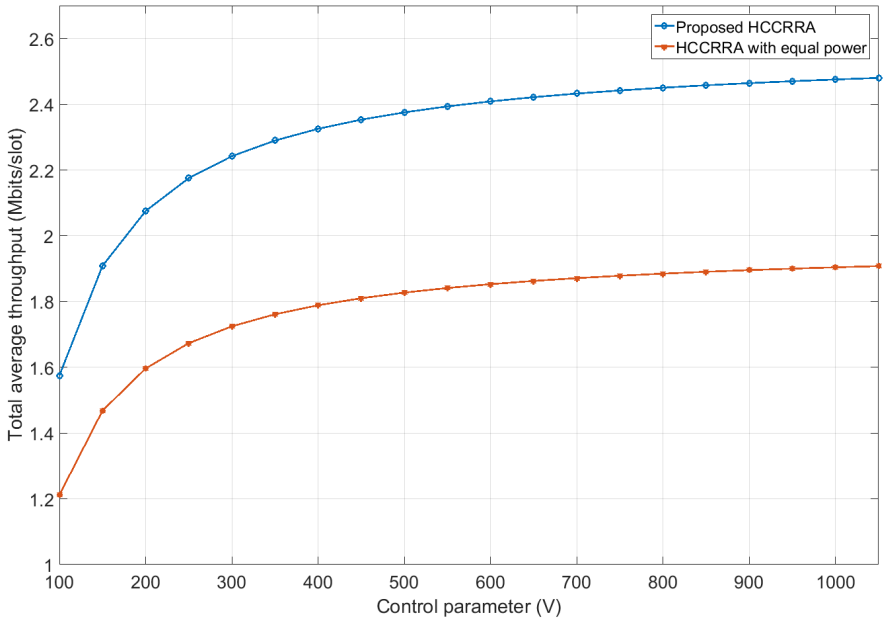

FIGURE 5.10: Average throughput versus control parameter.

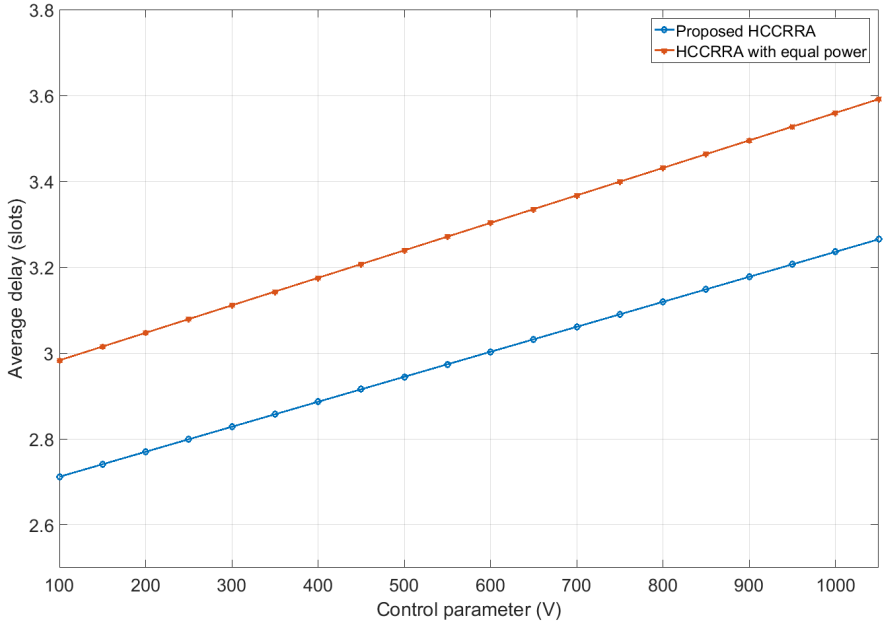

FIGURE 5.11: Average delay versus control parameter.

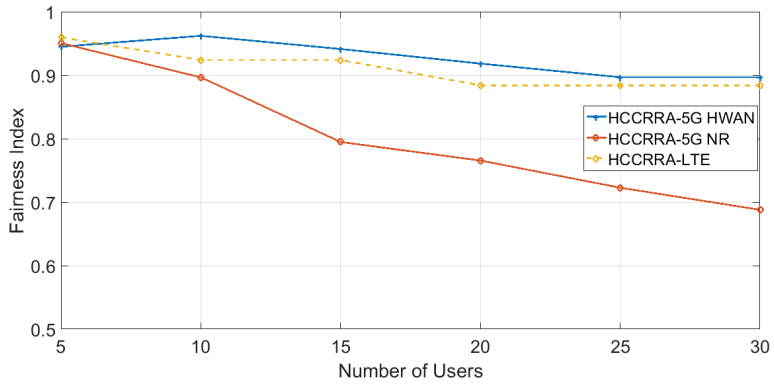

FIGURE 5.12: Performance evaluation in terms of fairness index.

fairness than the 5G HWAN and LTE scenarios since it is related to the fact that 5G NR provides higher throughput to the users in its coverage only. LTE maintains higher fairness than the 5G NR as it provides coverage in the entire macrocell. Moreover, we can visualize that 5G HWAN has higher fairness since almost all the macrocell coverage users connect to the LTE BS and 5G NR BSs, and maintain their queues stable.

5.6.7 Comparison of the proposed RAT selection approach with the traditional approaches

We compare our proposed RAT selection approach (user-centric with network assistance) with the traditional centralized and distributive methods. In the conventional centralized process, the central controller associates users with the BSs of the available RATs. In contrast, network selection's distributive mechanism allows users to select the appropriate BSs among the list of available RATs. Furthermore, we compare our algorithm with the algorithm illustrated in [83]. The authors in [83] consider radio resource allocation subject to each user's minimum data rate (QoS) requirement. Furthermore, they consider an equal power allocation scheme, and there is no congestion control mechanism. We, therefore, consider only the RAT selection scheme of [83] for comparison. We consider the percentage of users that select both RATs (multi-homed users) in the HWAN. Fig. 5.13 shows that initially, at low load, the rate of users that select both RATs is the same for all the three approaches. However, for the traditional centralized approach and our proposed scheme, a decrease in multi-homed users' percentage can be observed with an increase in network load. The centralized approach considers network load, whereas our proposed approach considers the user's minimum throughput connected to the available RAT. Therefore, to maintain optimal network performance, less number of users select multi-RATs. Furthermore, the centralized approach decreases the percentage of multi-RAT users more than required, whereas our proposed policy maintains a fair number of multi-RAT users without deteriorating network performance. Our results are comparable with the RAT selection approach of [83]. On the other

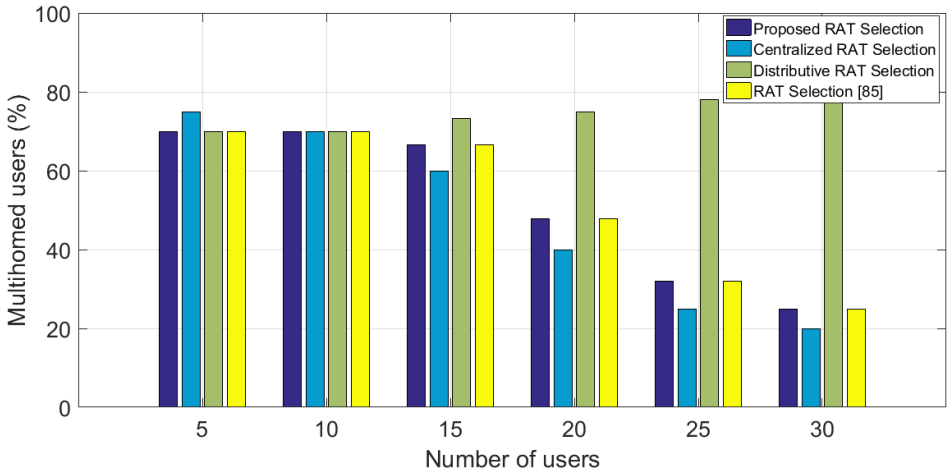

FIGURE 5.13: Percentage of multihomed users.

hand, in the traditional distributive process, the percentage of users who select multi-RATs remains the same as it does not consider network load. It may lead to an increase in the number of blocking and poor network performance and user experience.

5.7 Summary

In this chapter, we focused on hybrid radio resource management for a time-varying 5G HWAN. We have considered a CCN to provide interworking among different RATs. However, proposing a centralized approach for radio resource management can be expensive in computation and signalling overhead. Therefore, we proposed a hybrid radio resource management such that RAT selection is performed at the user end with network assistance. In contrast, the joint problem of radio resource allocation and congestion control, formulated as a stochastic optimization problem, is decomposed into two subproblems using the Lyapunov optimization technique. The CCN performs radio resource allocation, i.e. radio resource block, and transmits power allocation using Lagrange dual technique. Delay-throughput $[O(V) - O(1/V)]$ trade-off is obtained from our proposed HCCRRA scheme by satisfying network stability and power consumption constraints. This has been validated through both simulation results and mathematical analysis. Furthermore, our simulation results evaluate that our proposed algorithm outperforms the traditional MSR approach and the homogeneous scenarios (i.e. using only LTE BS or 5G NR BSs). Our proposed RAT selection approach performs better than the conventional centralized and distributive methods as it considers user throughput and the RAT load.

Chapter 6

Hybrid Radio Resource Management for Streaming over Time-Varying Heterogeneous Wireless Access Network

In this chapter, we investigate hybrid radio resource management for video streaming over time-varying heterogeneous wireless access networks composed of LTE BS and WLAN APs. This chapter extends the research work presented in chapter 3 as here we consider video streaming and video quality adaptation. Our main objective is to perform optimal radio resource allocation and optimal video quality adaptation. In addition, we investigate maximizing user quality of experience (QoE). Like chapter 5, we again used the Lyapunov optimization technique to decompose our two-time scale stochastic optimization scheme related to user quality of experience (QoE) problem into two main sub-problems. One of the sub-problems is related to radio resource allocation that operates at a scheduling time interval. The radio resource allocation policy is implemented at a centralized control node responsible for allocating radio resources from the available wireless networks using Lagrange dual method. The other sub-problem is related to the quality rate adaptation policy that works at a chunk time scale. Our main objective is that each user may select the appropriate quality level of the video chunks adaptively in a distributive way based on buffer state and channel state information. Furthermore, it is crucial to analyze and compare the QoE of our proposed approach over an arbitrary sample path of channel state information with an optimal T-slot algorithm. Moreover, it is desirable to evaluate the performance analysis of our proposed scheme for video streaming over a time-varying heterogeneous wireless access network.

6.1 Background and Introduction

Heterogeneous wireless access networks (HWANs), one of the prospective solutions, integrate different wireless networks such as WLAN, 3GPP Long Term Evolution (LTE), and 5G

new radio (NR). It accommodates a large number of connections with high data rates by enabling multihoming features. Moreover, it has overlapping coverage zones and unprecedented throughput gains compared to the traditional homogeneous wireless access network. Quality of Experience (QoE), described by the international telecommunication union (IUT) as an index of acceptability for an application or service perceived by mobile users, has recently gained attention over the last decade due to the high user quality of experience [93]. Furthermore, practical wireless radio networks support mobile users that operate at a time-varying channel condition. It is, therefore, essential to provide smooth video services over varying radio channel conditions.

Adaptive bitrate streaming has emerged as a critical technology that copes with the radio channel's time-varying nature by allowing bitrate adaptation over time [93]. In adaptive bitrate streaming, the content provider stores and encodes video content at different quality levels. Each encoded video is divided into video segments of fixed duration. Mobile user dynamically adapts its quality level according to its current channel condition, i.e. during a streaming duration, a mobile user with better channel conditions selects a high bit rate. In contrast, a lower bitrate is chosen due to worse channel conditions. Better quality selection improves the QoE of end-users. It is possible that the available radio resources at HWANs may not support high-quality video content, causing video interruptions and freezing (stalls), degrading end-user visual experience. Freezing time is related to playback buffer underrun, indicating an empty playback buffer. It is, therefore, desirable to select optimal quality to maintain playback video continuity. In [85], the authors analyzed network-centric bandwidth and adaptive quality adaptation for a streaming application over a time-varying channel for OFDMA based networks. Their proposed solution considers video quality and freezing time as the QoE metrics. However, this solution lacks heterogeneity since it only considers OFDMA based networks. The authors in [94] proposed QoE based algorithm for radio resource allocation in an LTE-based cellular network for adaptive video streaming. In [95], the authors explored adaptive video streaming for a dense wireless network. Video streaming adaptation is performed through a control-theoretic approach where conflicting QoE metrics such as video quality, freezing time and start-up delay are used as performance metrics. Their results are more promising than the available traditional solutions. One of the attributes affecting mobile user QoE is mobile battery consumption, as mobile users have high expectations from their battery life. The recent literature survey reveals that an increase in user device energy consumption is observed during a streaming application. Multimedia applications are power-hungry and drain the available battery of the device. Furthermore, an increase in video content quality leads to a rise in mobile device energy consumption since there is a tradeoff between quality and energy consumption. The authors in [96] investigated the mobile device energy-oriented adaptive multimedia streaming scheme for wireless channels. They optimized the tradeoff between users QoE for multimedia delivery and

users device energy management. An energy-aware multipath transport protocol for minimizing the mobile device energy consumption while obtaining the targeted video quality is proposed in [97]. In [84], the authors investigated cognitive multihoming, which maintains acceptable video quality and low energy consumption while minimizing cost and maximizing capacity. The authors in [98] proposed an adaptive bitrate quality selection scheme that extends device battery life during a playback session without compromising user QoE. Their energy saving is 10-30% with a slight tolerable drop in QoE.

We propose the video streaming problem as an optimization problem based on maximizing the QoE of end-users and minimizing mobile devices energy consumption subject to network-related constraints. Furthermore, we propose a hybrid solution for this problem by decomposing it into two components; the centralized part that performs radio resource allocation and the distributive part related to quality adaptation. In our proposed approach, the central controller node (CCN) allocates radio resources (bandwidth and power) to the mobile users, whereas each user individually performs quality adaptation based on video and device-related parameters. The factors affecting a mobile user QoE are identified as 1) video quality, 2) freezing time, and 3) battery power consumption. The main contributions are listed below:

1. Maximize long-term QoE related to high-quality video experience and minimum energy consumption using stochastic optimization problem formulation.

2. Introduce the Lyapunov optimization technique [86], [87] to develop an online and straightforward solution that decomposes the joint problem of rate allocation and quality adaptation into sub-policies, such that quality rate adaptation takes place at the user end and transmission rate allocation at CCN that requires CSI and QSI.

3. Develop a hybrid algorithm where the CCN performs radio resource allocation, and quality adaptation occurs at the user end. The CCN schedules the radio resources (subcarrier, power allocation and timeshare) at transmission time scale, and mobile users perform quality rate adaptation at chunk time scale.

4. Perform "cross-layer" optimization, i.e. radio resource allocation at the physical layer and quality adaptation at the application layer.

This chapter is organized as follows. The system model is given in Section 6.2. The problem formulation is given in Section 6.3, where Lyapunov optimization and problem decomposition are presented. Performance analysis of the proposed algorithm is presented in Section 6.4. Simulation results are given in Section 6.5. Finally, the conclusion of the chapter is presented in Section 6.6.

6.2 System Model

The HWAN is composed of a single macro base station (BS) of a cellular network (LTE) located at the cell's center overlaid by L WLAN access points (APs), where these WLAN APs are grouped as a set of $\mathcal{L} = \{1, 2, ..., L\}$, as shown in Fig. 6.1 [10]. These radio networks are orthogonal with no inference as they operate at different frequency bands. CCN allocates and maintains radio resources since all the APs and BS are connected to the controller. We consider K mobile users in the coverage of HWAN, and are grouped as a set $\mathcal{K} = \{1, 2, ..., K\}$. Mobile users have multimodal devices that can connect to more than one radio network simultaneously. The video data requested by the mobile user from the video server over the internet is buffered temporarily at HWAN. The video data is transmitted to users per scheduling interval from the BS/APs. Rate adaptation and rate allocation are based on channel state information (CSI) and queue state information (QSI). The BS/APs are responsible for sending the updated value of QSI to the CCN and mobile user devices. The CCN is updated with the CSI periodically using a feedback link from the receiver. In the individual RANs of HWAN, such as in LTE, the scheduling duration is ten milliseconds, whereas, in IEEE 802.11 WLAN, the channel assignment is for 0.5 milliseconds for the complete transfer of data frame. The CCN allocates radio resources at a global level. The CCN schedules the radio resources for a duration of 10 milliseconds so that mobile users and BS/APs use the scheduling decision of CCN. This scheduling interval of 10 milliseconds is enough to cope with signalling overhead, and propagation delay [46]. The rate adaptation decision occurs at the user device, where the bitrate for the next video chunk is determined once a video chunk/segment is placed in its respective buffer. The video chunk /segment duration is in the range of 1-10 seconds, and we assume rate adaptation occurs at a chunk time [96]. Our propose system scheduling policy operates at a physical transmission time scale t. In contrast, the chunks are requested at integer multiples of segment time, i.e., $t = mt_c$. Here t represents the physical transmission slot, m describes the video chunk, and t_c depicts the physical frames per video chunk time, which is assumed to be integer [85].

6.2.1 Radio Resource Allocation in HWAN

For time-variant HWAN with a cellular BS and WLAN APs, it is essential to describe the downlink resource allocation from OFDMA based system and WLAN in detail. In an OFDMA based system (LTE), we consider N subcarriers of bandwidth G given by $G = B/N$, where B is the total bandwidth. We assume that multiple users could share one subcarrier in a time-sharing manner by using $x_{nk} \geq 0$ as the time-sharing fraction for the allocation of subcarrier n to user k. The transmit power allocated on subcarrier n to user k during the time-sharing slot is $\frac{p_{nk}}{x_{nk}}$. The maximum data rate of mobile user k on subcarrier n is approximated by

Shannon theorem given as follows:

$$r_{nk} = \begin{cases} x_{nk} G log_2(1 + \frac{p_{nk} g_{nk}}{x_{nk} G N_0}), & x_{nk} > 0, \\ 0, & x_{nk} = 0 \end{cases} \tag{6.1}$$

where $p_{nk}(t)$ describes the transmit power for the link between user k and subcarrier n at time slot t. $g_{nk}(t)$ is the channel gain of the link between user k and subcarrier n at time slot t. N_0 is the noise power spectral density. The total transmission rate obtained by user k from the BS's cellular c is given by

$$R_k^c\Big(X(t), P(t)\Big) = \sum_{n=1}^{N} r_{nk}(t), \tag{6.2}$$

where $X(t) = [x_{nk}]_{N \times K}$ and $P(t) = [p_{nk}(t)]_{N \times K}$.

In IEEE 802.11, WLAN users can share the whole bandwidth without any collision as it possesses an enhanced feature of distributive coordination function (DCF). In WLAN, the user k receives data rate r_{lk} from AP l is approximated by determining the instantaneous signal-to-noise-ratio (SNR) and is given as

$$r_{lk}(t) = f\left(\frac{P_{out}^w \, g_{lk}(t)}{N_o}\right). \tag{6.3}$$

The transmitted power from WLAN w's AP is presented as P_{out}^w. The WLAN channel's noise variance is denoted by N_o and $g_{lk}(t)$ describes the link gain between user k and AP l. SNR threshold is used to evaluate the achievable data rate and is approximated by $f(.)$. It is assumed that mobile users connect to a single AP, which provides the highest data rate among the available APs. WLAN uses time division multiple access (TDMA), where users occupy the entire bandwidth for its allocated time slot $t_{lk}(t)$. The transmission rate received by user k from WLAN w's AP is given by

$$R_k^w\big(T(t)\big) = \sum_{l=1}^{L} t_{lk}(t) r_{lk}(t), \tag{6.4}$$

where $T(t) = [t_{lk}(t)]_{L \times K}$. The total transmission rate received by user k in HWAN at time slot t is given by

$$R_k(t) = R_k^c(t) + R_k^w(t). \tag{6.5}$$

The time average transmission rate of the HWAN is given by

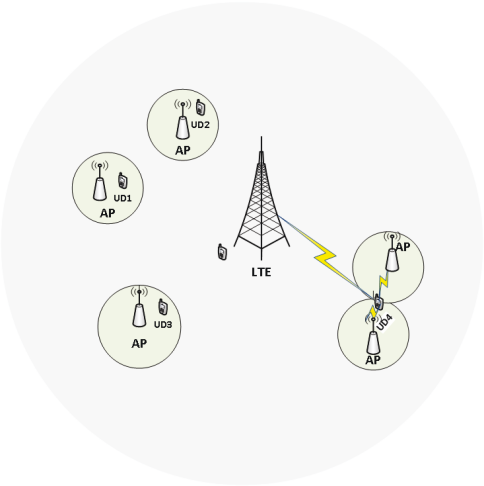

FIGURE 6.1: Heterogeneous wireless access network layout.

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{(R_k(t)\}. \tag{6.6}$$

6.2.2 Video Model

The video content S_k requested by each user $k \in K$ is modelled as a set of chunks. Each video chunk is available in a finite set of quality levels, i.e. $A_k \in [1, 2, ...A_{S_k,max}]$ where $A_{S_k,max}$ is the maximum quality level for a video chunk [99], [100]. Each quality level of a chunk corresponds to a different bitrate in a variable bit rate (VBR) of video encoding, i.e. $D_{S_k}(A_k, m)$ is the number of bits for m-th chunk of quality level A of a video S_k for user k [101], [102]. The quality measure of an m-th video chunk perceived by user k at quality level A_k is $U_{S_k}(A_k, m)$.

6.2.3 Transmission Buffer Dynamics and Stability

Each user $k \in K$ in HWAN requests the quality level of the mth chunk at the beggining of mth chunk time, i.e. at each time slot $t \in \{0, t_c, 2t_c, ...\}$ each user $k \in K$ in the system determines the quality level $A_k(t) \in \{1, 2, ...A_{S_k,max}\}$ of its requested chunk. Based on this decision, at time slot t, the number of bits $D_{S_k}(A_k(t), t)$ and the corresponding video quality measure $U_{S_k}(A_k(t), t)$ of the requested chunk is specified. The decision of quality level is made at time t which is an integer multiple of t_c, therefore, we have $U_{S_k}(A_k(t), t) = 0$ and $D_{S_k}(A_k(t), t)$ at $t \notin \{0, t_c, 2t_c, ...\}$ [100].

HWAN maintains transmission buffer for each mobile user k, which evolves at transmission slot $t \in \{0, 1, 2, 3, ...\}$, and is given by

$$Q_k(t+1) = \max[Q_k(t) - R_k(t), 0] + D_{S_k}(A_k(t), t), \tag{6.7}$$

where $R_k(t)$ is the channel transmission rate and corresponds to dequeuing process i.e. the number of bits transmitted to user k at each transmission time slot t. The enqueue process which corresponds to the placement of bits $D_{S_k}(A_k(t), t)$ in queue $Q_k(t)$ occurs at time slot t. $Q_k(t)$ maintains the bits of chunks requested by user k that are not yet transmitted to the user. An individual buffer $Q_k(t)$ is stable if it holds the following condition, given as

$$\overline{Q_k} = \sup_{\lim_{T\to\infty}} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{(Q_k(t)\} < \infty. \tag{6.8}$$

In a practical system, individual queue is considered as stable if the time-averaged dequeue process is greater than or equal to the time averaged enqueue process.

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{(R_k(t)\} \geq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{D_{S_k}(A_k(t), t)\} \tag{6.9}$$

A network with stable individual queues is considered as stable [86].

6.2.4 Buffer Model

At each user device, the playback buffers store the downloaded unplayed video content. Nowadays, mobile devices such as smartphones are provided with enough space to accommodate the whole video chunks [103]. In our proposed analysis, we, therefore, ignore buffer video chunks dropping due to buffer overflow. The unpredictable wireless channel conditions and limited network resources cause buffer underrun, which results in freezing time [104]. In VBR streaming, different chunks have different sizes and bitrates, making it hard to establish a relationship between buffer playback time and buffer video size. Therefore, we focus on the playback buffer video time. The playback buffer time evolves with the download of a chunk and the video being watched [105]. When a video chunk of length C seconds is added to the playback buffer, its level increases by C seconds, whereas watching a video decreases the playback buffer time. Let at time slot $t = mt_c$, the user k starts downloading chunk m. The download time $t_{d,k}(m,t)$ of chunk m for user k depends on the number of bits of the selected video chunk $d_k(m,t)$ and the transmission rate $R_k(t)$ at time slot t i.e. $t_{d,k}(m,t) = \frac{d_k(A_k(m,t)}{R_k(t)}$ [85]. Furthermore, we assumed that one chunk is downloaded per transmission time slot t. We can therefore simplify the expression for download time as $t_{d,k}(t) = \frac{d_k(A_k(t))}{R_k(t)}$. We assume that the duration of the transmission time slot is greater than or equal to the download time, i.e. $t_p \geq t_{d,k}$ Let the buffer video length is $B_k(t-1)$ seconds

for user k at time slot $(t-1)$. The playback buffer occupancy of user k at slot t is given as

$$B_k(t) = \max[B_k(t-1) - t_{d,k}(t), 0]^+ + C, \tag{6.10}$$

where $[x]^+ = \max[x, o]$ ensures the term is always positive. If $B_k(t-1) \leq t_{(d,k)}(t)$, the buffer drains out during the time slot $t-1$. Hence, the freezing time $T_{f,k}(t)$ for user k at time slot t is given as $T_{f,k}(t) = t_{d,k}(t) - B_k(t-1)$, otherwise $T_{f,k} = 0$.

6.2.5 Energy Consumption Model

The user device equipped with multi-RAT interfaces can simultaneously receive the video content over the 3GPP interface and WLAN interface. The mobile user device reception procedures depend on the available battery power as these devices are battery operated. The energy consumption model is composed of reception energy during streaming session, ramp energy and tail energy [106]. The energy consumption model of a user k for an interface i during video content S_k reception at time slot t is given by

$$E_{i,S_k}(t) = r_{t,i}(t) * D_{S_k}(A_k(t), t) + r_{p,i}(t) + r_{r,i}(t), \tag{6.11}$$

where the energy consumption $r_{t,i}(t)$ (Joules/Kbytes) of an interface i is related to data transfer energy during the streaming session. $D_k(A_k(t), t)$ is the bitrate of user k using an interface i during a time slot t, whereas $r_{p,i}(t)$ measured in joules is the ramp energy consumption rate related to high power state transition (for 3GPP interface) and scanning and associating (for WLAN interface) [46], [106]. Since the mobile device is at a high energy state during the reception of video content, we do not include ramp energy. $r_{r,i}(t)$ is the tail energy. Thus our energy consumption model is related to the energy consumptions during a streaming session. The energy consumption $E_{S_k}(t)$ of the user device for user k after the download of an associated chunk at time slot t is the linear sum of power consumption of all active links

$$E_{S_k}(t) = \sum_{i \in I} E_{i,S_k}(t), \tag{6.12}$$

where $i \in [1, ..I]$. Battery consumption is high when a high-quality bitrate is selected. Therefore, it must adapt to the bitrate, which avoids energy starvation and prolongs a user device battery life during streaming.

6.2.6 QoE Model

For our QoE model, we have considered three factors influencing the quality of experience of mobile users. Two of them are related to the video content, such as the quality of the video and freezing time, whereas the third feature is related to the device battery consumption, which is further related to the quality level of the video chunk selected by mobile users. The

QoE of user k is given as [107]

$$\overline{QoE}_k = f(\overline{U}_k, \overline{T}_{f,k}, \overline{E}_k) \tag{6.13}$$

6.2.6.1 Chunk Quality

At favourable channel conditions, the user perceives higher quality by selecting a higher quality level and vice versa. The time-averaged perceived chunk quality of user k is given by

$$\overline{U}_k = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{U_{S_k}(A_k(t), t)\}. \tag{6.14}$$

6.2.6.2 Freezing Time

The freezing time which determines the duration of buffer underrun is given as

$$\overline{T}_{f,k} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{T_{f,k}(t)\}. \tag{6.15}$$

6.2.6.3 Energy consumption

The average energy consumption of the device is given by

$$\overline{E}_k = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{E_{S_k}(A_k(t), t)\}. \tag{6.16}$$

The time averaged quality of experience of user k is given as

$$\overline{QoE}_k = w_1\overline{U}_k - w_2\overline{T}_{f,k} - w_3\overline{E}_k, \tag{6.17}$$

where w_1, w_2 and w_3 are the non-zero weights related to video quality, freezing time, and the user device battery consumption, respectively. These weights represent each parameter's relative importance, i.e. a lower weight of a metric indicates the user does not care about that metric, whereas more attention is paid to the metric if it has a higher weight.

6.3 Problem Formulation

The joint optimization problem for quality rate adaptation and radio resource allocation involves the maximization of the time-averaged users QoE and is given as follow:

$$\max_{X(t)P(t),T(t),A(t)} \quad \sum_{k=1}^{K} \overline{QoE_k}, \tag{6.18a}$$

$$\text{s.t.} \quad \text{C1} : \overline{Q_k} < \infty, \quad \forall k, \tag{6.18b}$$

$$\text{C2} : \sum_{k=1}^{K} \sum_{n=1}^{N} p_{nk}(t) \leq P^{c}_{max} \tag{6.18c}$$

$$\text{C3} : \sum_{k=1}^{K} t_{lk}(t) \leq 1, \quad \forall l, t \tag{6.18d}$$

$$\text{C4} : t_{lk}(t) \geq 0, \quad \forall l, k, t \tag{6.18e}$$

$$\text{C5} : \sum_{k=1}^{K} x_{nk}(t), \quad \forall n, t \tag{6.18f}$$

$$\text{C6} : 0 \leq x_{nk}(t) \leq 1, \quad \forall n, k, t \tag{6.18g}$$

$$\text{C7} : A_k(t) \in \{1, 2, ..., A_{S_k,max}\}, \quad \forall k, t \tag{6.18h}$$

$$\text{C8} : T_{f,k}(t) \leq \beta_T, \quad \forall k, t \tag{6.18i}$$

$$\text{C9} : E_{S_k}(t) \leq E_{k,th}(t), \quad \forall k, t \tag{6.18j}$$

where $X(t), P(t), T(t)$ represents the radio resource allocation policy of the HWAN, and $A(t) = \{A_k(t)\}$ is the chunk quality level matrix. C1 shows the buffer stability of the HWAN. C2 and C3 approximate the total power constraint from cellular BS and the time-sharing constraint of WLAN AP among mobile users. The physical definition of t_{lk} is approximated by constraint C4. Constraint C5 is related to the subcarrier allocation from LTE BS. Constraint C6 describes the physical definition of x_{nk}. Constraint C7 is related to the quality adaptation, which indicates that each user k picks up the quality level available in the finite set $\{1, 2, 3...A_{S_k,max}\}$. C8 is related to the playback buffer underrun, where β_T is the minimum threshold value. C9 is the energy consumption constraint, where $E_{k,th}$ is the maximum energy consumption threshold value. $E_{k,th}$ has a lower value at the device's lower residual energy. The device's lower battery may lead to a lower quality level for the next downloaded video chunks. The realization of the offline optimization problem Eq. (6.18a-6.18j) is possible if information related to CSI and $D_k(t)_{1:A_{S_{k,max}}}$ is known, which is impractical. Furthermore, the direct solution of this problem has very high computational complexity since the state space grows exponentially with the number of users. To overcome these challenges, we propose an online optimization problem. The Lyapunov optimization concept translates the

problem given in Eq. 6.18a-6.18j into an online optimization problem. It decomposes problem P1 into two optimization policies: 1) radio rate allocation policy performed by CCN, and 2) quality adaptation policy where mobile users achieve quality adaptation by selecting an appropriate quality level for each video segment. The following section presents our proposed dynamic hybrid radio resource allocation and quality adaptation for video streaming over time-varying HWANs in detail.

6.3.1 Problem Transformation

To transform the joint rate allocation and rate adaptation problem into two separate rate allocation and rate adaptation policies performed at CCN and user device, we consider the following Lyapunov optimization function

$$L(Q(t)) \triangleq \frac{1}{2}\sum_{k=1}^{K} Q_k^2(t). \tag{6.19}$$

The Lyapunov drift function $\Delta Q(t)$ quantifies the difference between the Lyapunov function at time interval [t, t+1].

$$\Delta(Q(t)) \triangleq \mathbb{E}\{L(Q(t+1)) - L(Q(t))|Q(t)\} \tag{6.20}$$

Eq. (6.20) grows large when the system moves to an undesirable state. To minimize $\Delta(Q(t))$, control decisions are made at every time slot t. This procedure lowers backlogs which maintain network stability. To obtain two separate rate allocation and quality adaptation optimization policies, we need to subtract the term $V\mathbb{E}\{\mathrm{QoE(t)}\}$ from Eq. 6.20.

$$\mathbb{E}\{L(Q(t+1)) - L(Q(t)) - V\mathrm{QoE(t)}|\mathrm{Q(t)}\}, \tag{6.21}$$

where $\mathbb{E}\{\mathrm{QoE(t)}\} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{QoE_k(t)}$ at every time slot t and $V(V \geq 0)$ is a control parameter used to control the tradeoff between drift $\Delta(Q(t))$ and reward $\mathbb{E}\{\mathrm{QoE(t)}\}$. The upper bound of the drift-minus-reward is derived as follow:

$$\begin{aligned}
&\Delta(Q(t)) - V\mathbb{E}\{\mathrm{QoE(t)}\} \\
&\qquad \leq A + \mathbb{E}\Bigg\{ \sum_{k=1}^{K} \{Q_k(t)\{D_{S_k}(A_k(t),t) \\
&\qquad\quad -Vw_1 U_{S_k}(A_k(t),t) + Vw_3 E_k(A_k(t),t)\Bigg\} \\
&\qquad +\mathbb{E}\Bigg\{ \sum_{k=1}^{K} \Big(Vw_2 T_{f,k} - Q_k(t)R_k(t) \Big) \Bigg\},
\end{aligned} \tag{6.22}$$

where $A = \frac{1}{2}\mathbb{E}\left\{ \sum_{k=1}^{K} \left(R_k^2(t) + D_{S_k}(t) \right) \right\}$.

The second term on the right-hand side of Eq. 6.22 is a function of the bitrate of video chunk and therefore corresponds to the receiver's rate adaptation optimization policy, whereas the last term of Eq. 6.22 represents the CCN scheduling optimization policy. The Lyapunov optimization technique transforms the joint rate allocation and quality adaptation problem into two subproblems given as follow:

6.3.1.1 Quality Rate Adaptation Optimization Policy Derivation

The adaptation decision regarding the video chunk quality level is made by mobile user at each chunk time t, i.e. at transmission time slot $t = mt_c$. The adaptation policy is given by considering second term of Eq. (6.22) as

$$\begin{aligned} &\underset{A_k(t)}{\text{minimize}}\ \mathbb{E}\left\{ \sum_{k=1}^{K} \left(Q_k(t) D_{S_k}(A_k(t), t) \right) \right\} + \\ &\mathbb{E}\left\{ \sum_{k=1}^{K} \left([w_3 E_k(A_k(t), t) - w_1 U_{S_k}(A_k(t), t) \right) \right\} \\ &\text{s.t C7 \& C9} \end{aligned} \tag{6.23}$$

The quality level $A_k(t)$ is the minimization variable that appears in the separate terms of the sum, and therefore, we can optimize it separately over each user. Each user receives its instantaneous queue state information $Q_k(t)$ from the BS at each time slot using simple protocol signalling with minimum overhead. Each user has the information of bitrate $D_{S_k}(., t)$ and quality measure $U_{S_k}(., t)$ at every time slot t. Hence, the quality level $A_k(t)$ of each video chunk is determined at the user end in a distributive way at each time slot t as follows:

$$\begin{aligned} A_k(t) = &\text{argmin}\Big\{ Q_k(t) D_{S_k}(A, t) + V\big[w_3 E_k(A, t) \\ &- w_1 U_k(A, t) \big] \Big\}, \\ &\text{where } A \in \{1, 2, ..., A_{S_k, max}\}, \\ &E_{S_k}(t) \leq E_{k,th}. \end{aligned} \tag{6.24}$$

It should be noted that at $t \notin \{0, t_c, 2t_c, ...\}$, $U_{S_k}(A_k(t), t) = 0$ and $D_{S_k}(A_k(t), t)$, since no video chunks are requested [100]. Eq. (6.24) shows that quality level is related to video content, the user queue backlog, the device energy consumption, and user preferences for video quality and energy consumption. Thus, selecting optimal quality levels for the requested

chunk avoids network congestion and greater energy consumption at the user device without deteriorating the video content's quality measure.

$$\begin{aligned} &\min \sum_{k=1}^{K} \Big(V w_2 T_{f,k}(t) \Big) \\ &\text{where } T_{f,k}(t) \leq \beta_T, \forall k, t. \end{aligned} \tag{6.25}$$

Every user has its own playback buffer,therefore, the freezing variable $T_{f,k}(t)$ is determined per user k.

$$\begin{aligned} &\min \Big(V w_2 T_{f,k}(t) \Big) \\ &\text{where } T_{f,k}(t) \leq \beta, \forall t \end{aligned} \tag{6.26}$$

$$\begin{cases} \text{Case I:} \quad T_{f,k} = 0, & \text{where } T_{f,k} \leq \beta_T \\ \text{Case II:} \quad T_{f,k}(t) = t_{d,k}(t) - B_k(t-1), & \text{where } T_{f,k} > \beta_T \end{cases} \tag{6.27}$$

When $T_{f,k} \leq \beta_T$, i.e. the playback buffer is stable, then the adaptation policy uses Eq. (6.24) to determine the quality level for the next video chunk. However, when $T_{f,k} > \beta_T$, i.e. the playback buffer is in underrun state and the user may experience stalls. Therefore, to maintain smooth QoE, our proposed quality adapatation policy works in two steps as follow. Step i) Since $T_{f,k} = t_{d,k}(t) - B_k(t-1)$ and $t_{d,k} = f\Big(d_k(A,t), R_k(t)\Big)$, therefore, the quality adaptation policy picks a lower quality level for the next chunk to be downloaded to avoid playback buffer underrun. The chunk quality level adjustment is given as

$$A_{S_k,max} = \max\left(A(A_{S_k-1}, 1)\right) \tag{6.28}$$

Step ii) The constraint for the available quality levels, i.e.$A \in \{1, 2, ...A_{S_k,max}\}$ in Eq. (6.24) is updated with the value obtained in Eq. (6.28) and the optimal value of the quality level of the next video chunk is selected using Eq. (6.24).

Our quality adaptation policy maintains smooth streaming by minimizing events of stalls. The third term of Eq, (6.24) has a significant influence on quality adaptation policy. The overall quality rate adaptation policy is given in the Algorithm 4 as follows.

Algorithm 4 Quality Rate Adaptation Policy

1: Initialize: $t \leftarrow 0$, $Q_k(0) = 0$, $T_{f,k}(0), \forall k$
2: Input: Each user observes its $Q_k(t)$ from the BS.
3: Each usre k gets its $E_{S_k}(t)$, rate-quality measure i.e., $D_{S_k}t$, $U_{S_k}(t)$, and $T_{f,k}(t)$ at each chunk time slot, i.e., at transmission time slot $t = mt_c$, where $t \in \{0, t_c, 2t_c, ...\}$.
4: **If** $T_{f,k} \leq \beta_T$
5: Use Eq. (6.24) to determine the quality level $A_k(t)$
6: **Else** use Eq. (6.28) to update the constraint $A \in \{1, 2, ..., A_{S_k,max}\}$ for Eq. (6.24)
7: Use Eq. (6.24) to determine the quality level $A_k(t)$
8: **end if**
9: $t \leftarrow t + 1$
10: Update the $Q_k(t)$ and $B_k(t)$ $\forall k$ using Eq. (6.7) and Eq. (6.10), respectively.
11: go back to step 2.

6.3.1.2 Radio Resource Allocation Policy Optimization at CCN

The decision of radio resource allocation made by CCN at every transmission time slot t is based on the last term of Eq. (6.22) and is given as

$$\begin{aligned} &\underset{X(t),P(t),T(t)}{\text{minimize}} \sum_{k=1}^{K} -Q_k(t)R_k(t) \\ &\text{s.t. C1- C6.} \end{aligned} \tag{6.29}$$

The objective function related to radio resource allocation is a convex optimization problem. The CCN observes the updated QSI of all users from the BS. It gets updated CSI from users at every transmission time t. Although, the radio resource allocation policy is not directly related to the quality adaptation policy. Eq. (6.27) influence the decision of radio resource allocation since download time $t_{d,k} = f\Big(d_k(A,t), R_k(t)\Big)$. Therefore, we introduce a new variable $b_k(t)$, and modify Eq. (6.29) as follow

$$\begin{aligned} &\max_{X(t),P(t),T(t)} \sum_{k=1}^{K} b_k(t)Q_k(t)R_k(t) \\ &\text{s.t. C1- C6.} \end{aligned} \tag{6.30}$$

where using Eq. (6.26) and Eq. (6.27), the playback buffer underrun variable $b_k(t)$ is given as

$$\begin{cases} \text{Case I: } \; b_k(t) = 1, \text{ where } T_{f,k} \leq \beta_T \\ \text{Case II: } \; b_k(t) > 1, \text{where } T_{f,k} > \beta_T \end{cases} \tag{6.31}$$

BS/AP receives $b_k(t)$ from all of its connected users. It then forwards this information to CCN. The CCN gives higher priority to users with value of $b_k(t) > 1$, while taking the

constraints C1- C4 into consideration. Higher $b_k(t)$ is related to a higher risk of playback buffer underrun at the receiver (mobile user) end. Therefore, to reduce the download time $t_{d,k}(t)$, the transmission rate $R_k(t)$ should be increased. It is accomplished by allocating more radio resources to users with a value of $b_k(t) > 1$.

Mobile users are multihoming, where they can simultaneously connect to both cellular BS and WLAN AP. Our objective function for the radio resource allocation policy is given as

$$\max_{X(t),P(t),T(t)} \sum_{k=1}^{K} \{ Q_k(t) b_k(t) [R_k^c(t) +) R_k^w(t)] \} \quad \text{s.t. C1- C6.} \tag{6.32}$$

We decompose Eq. (6.32) into two sub-problems to allocate optimal radio resources (subcarrier and power) from cellular BS and optimal time-share from WLAN. The problem shown in Eq. 6.32 can be easily decomposed into two parts as neither the objective function nor the constraints are coupled.

A) Optimal Subcarrier and Power Allocation: The first part of Eq. (6.32) is concave in terms of x_{nk} and p_{nk} [15]. All the constraints are linear and affine. Our problem of radio resource allocation from cellular BS is a convex optimization problem. Optimal power allocation is derived using the first part of Eq. (6.32),

$$\max_{X(t),P(t)} \sum_{k=1}^{K} Q_k(t) b_k(t) R_k^c(t) \} \quad \text{s.t. C2, C5, C6.} \tag{6.33}$$

Using Lagrange function and Karush-Kuhn Tucker (KKT) conditions, we obtain the optimal power allocation and subcarrier allocation [17]. The optimal solution of Eq. (6.33) is given as

$$\begin{cases} x^*_{nk^*} = 1, \; p^*_{nk^*} = x_{nk^*} W [\frac{Q_k(t) b_k(t)}{\sigma 1n2} - \frac{No}{g_{nk}(t)}]^+, \text{if } k^* = k \\ x^*_{nk^*} = 0, \; p^*_{nk^*} = 0, \qquad \text{otherwise,} \end{cases} \tag{6.34}$$

where $k^* = \arg\max_k \alpha_{nk}(t)$, and $\alpha_{nk}(t) = Q_k(t) b_k(t) W x_{nk} log_2(1 + \frac{p^*_{nk}(t) g_{nk}(t)}{b_k(t) W x^*_{nk} N_o}) - \sigma p_{nk}(t)$. σ is the Lagrange multiplier for sum power constraint. Optimal power and subcarrier allocation is based on "water filling" and "winner takes all" techniques [49], [90]. However, in our approach the water filling technique takes instantaneous values of CQI and CSI, i.e. $Q_k(t)$ and $g_{nk}(t)$ respectively. Furthermore, it dependes on the variable $b_k(t)$ which is related to the state of playback buffer at user end. Eq. (6.34) shows that users with good channel condition receives more transmit power on the allocated subcarrier. Similarly, larger the queue length and $b_k(t)$, the higher the value of $p^*_{nk}(t)$ to maintain a stable network and provide smooth

streaming at end user. "Winner takes all" strategy allows the allocation of subcarrier to at most one mobile user.

B) Optimal Time-Share Allocation: The time-share allocation is derived by using the second part of Eq. (6.32),

$$\max_{T(t)} \sum_{k=1}^{K} \sum_{l=1}^{L} \phi_{lk}(t) t_{lk}(t) \quad \text{s.t. C3- C4.} \tag{6.35}$$

where $\phi_{lk}(t) = \left[Q_k(t) b_k(t) r_k^w(t) \right]$. As per our assumption, mobile users can connect to only one WLAN AP that provides the highest transmit rate among all the available APs, i.e. $t_{lk} > t_{mk} > 0$, where $l \neq m$. Let WLAN AP l serves the set of mobile users $\omega_l(t)$ in a TDMA manner where $\omega_l(t) = \{k \mid r_l k > 0\}$. Connectivity conditions of mobile users from the set $\omega_l(t)$ with WLAN AP l is given as follow:

$$\begin{cases} t_{lk} = 0, \text{ if } k \in \omega_l^-(t) \text{ and } l \neq m \\ t_{lk} = 1, \text{ if } k \in \omega_l^+(t) \text{ and } k = k^* \\ t_{lk} = 0, \text{ if } k \in \omega_l^+(t) \text{ and } k \neq k^* \end{cases} \tag{6.36}$$

where $k^* = \arg\max_{k \in \omega_l^+(t)} \phi_{lk}(t)$, $\omega_l^+(t) = \{k \mid \phi_{lk} > 0, k \in \omega_l(t)\}$ and $\omega_l^-(t) = \{k \mid \phi_{lk} \leq 0, k \in \omega_l(t)\}$. So Eq. (6.35) is modified as

$$\max_{T(t)} \sum_{l=1}^{L} \sum_{k \in \omega_l(t)} \phi_{lk} t_{lk}^w(t) \quad \text{s.t. C3- C4.} \tag{6.37}$$

Based on the above analysis, we can decompose our objective function into L subproblems as each WLAN AP works independently. The objective function for AP l is given as

$$\max_{t_{lk}(t)} \sum_{k \in \omega_l(t)} \phi_{lk} t_{lk}^w(t) \quad \text{s.t. C3- C4.} \tag{6.38}$$

Constraint C3 shows that $\sum_{k=1}^{K} t_{lk} \leq 1$, for WLAN AP l , then $t_{lk}^* \leq 1 - \sum_{k=1}^{K} t_{lk}$. Let WLAN AP l has two users k_1 and k_2, such that $t_{lk_1} = 1 - t_{lk_2}$.The objective function of Eq. (6.38) is given as

$$\begin{aligned}
&\phi_{lk_1}(t)t_{lk_1}(t) + \phi_{lk_2}(t)t_{lk_2}(t) \\
&\leq (1 - t_{lk_2}(t))\phi_{lk_1}(t) + \phi_{lk_2}(t)t_{lk_2}(t) \\
&\leq (\phi_{lk_2}(t) - \phi_{lk_1}(t))t_{lk_2}(t) + \phi_{lk_1}(t),
\end{aligned} \tag{6.39}$$

where $\phi_{lk_i}(t) = \min\{k_i \mid \phi_{lk_i}(t) < 0, \forall\, i = \{1,2\}\}$. To exclusively allocate the time slot t of AP l to user k_2, let maximize $(\phi_{lk_2}(t) - \phi_{lk_1}(t))t_{lk_2}(t) + \phi_{lk_1}(t)$, we have $t_{lk_2}(t) = 1$, and $t_{lk_1}(t) = 0$ and $\phi_{lk_2}(t) > \phi_{lk_1}(t) > 0$. Otherwise $t_{lk_1}(t) = 1$ and $t_{lk_2}(t) = 0$ when $\phi_{lk_1}(t) > \phi_{lk_2}(t) > 0$. This procedure can be applied to K mobile users that belong to the set $\omega_l^+(t)$ to get the optimal time share described in Eq. (6.36). This optimal time share allocation is related to the product of $Q_k(t)$, $b_k(t)$ and $r_{lk}(t)$, since $\phi_{lk}(t)$ increases at higher values of $Q_k(t)$, $b_k(t)$ and $r_{lk}(t)$. As a result the mobile users occupy the entire-time fraction of the WLAN AP to stabilize the transmission queue at time slot t and maintain a freezing free video streaming. Finally, the overall algorithm for radio resource allocation from cellular BS and WLAN AP is given in Algorithm 5.

Algorithm 5 Radio Resource Allocation Policy

1: Initialize: $t \leftarrow 0$, $Q_k(0) = 0$, $b_k(0) = 1, \forall k$
2: Input: Observe $Q_k(t)$, obtain $G_k(t) = \{g_{mk}(t), g_{nk}(t)\}$, and $b_k(t)$ from each user k at each transmission time slot t, where $t \in \{0, 1, 2, ...\}$.
3: Use Eq. (6.34) to calculate the optimal $x_{nk}^*(t)$, $p_{nk}^*(t)$
4: Use Eq. (6.39) to determine $T^*(t)$ for the users associated with the available WLAN AP, i.e from $l = 1$ to L
5: $t \leftarrow t + 1$
6: Update the $Q_k(t), \forall k$ using Eq. (6.7).
7: Each user updates its CSI and $b_k(t)$.
8: go back to step 2.

In **Algorithm** 5 the CCN performs radio resource allocation, each user device performs its quality adaptation and finally $Q_k(t)$ and $B_k(t)$ are updated at both network end and at user end respectively.

C) Complexity Analysis: The computation complexity of Algorithm 4 is $O(1)$, since quality adaptation is performed at user end distributively. The computation complexity of Algorithm 5 is $O(LK) + O(NK/\mu^2)$ since the CCN allocates radio resources. Here $O(NK/\mu^2)$ is the complexity of subcarrier and power allocation of the cellular BS, where $(1/\mu^2)$ corresponds to the number of iterations that achieves convergence accuracy of μ of Eq. (6.33). $O(LK)$ is related to the complexity of the time fraction allocation of WLAN. Our proposed hybrid approach requires CSI, QSI and b_k at each time slot t. For the realization of radio resource allocation policy, the CCN receives CSI and b_k from each user via BS/AP resulting in $O(K)$ signalling overhead on the air interface per time slot t, whereas the quality adaptation

policy requires QSI over the air interface per time slot t. It has a signalling overhead of $O(K)$.

6.4 Performance Analysis

In this section, following the footsteps of our previous work [17], we analyze the performance bounds of our proposed algorithm. User motion causes CSI variations. As a result, the network ergodicity state is unlikely to hold in most practical scenarios. Furthermore, we assume that these variations of CSI occur at the same time scale of video chunk streaming. To prove the optimality of our proposed algorithm, we compare our QoE with that achieved by an optimal policy with T-slot lookahead (i.e., knowledge of the future CSI over an interval of length T slots). For an arbitrary sample path of the network state $\Omega(t)$, we consider the static optimization problem over the j-th frame $j \in [0, 1, 2, ..., F-1]$

$$\begin{aligned} &\max QoE = \frac{1}{T} \sum_{t=jT}^{jT+T-1} \sum_{k=1}^{K} QoE_k(t) \\ &\text{s.t.} \sum_{t=jT}^{jT+T-1} \sum_{k=1}^{K} (D_k(t) - R_k(t)) \leq 0 \end{aligned} \tag{6.40}$$

In Eq. 6.40, the CSI for $t \in \{jT, ..., jT+T-1\}$ are assumed known for the j-th frame. For a time slot t, let consider one slot- Lyapunov drift $\Delta\Theta(t)$ as follow:

$$\Delta(\Theta(t)) = L(\Theta(t+1)) - L(\Theta(t)). \tag{6.41}$$

Subtracting the term $\mathbb{E}\{QoE(t)\}$ and using $V(V \geq 0)$, the drift-minus-reward is derived as follow:

$$\begin{aligned} &L(\Theta(t+1)) - L(\Theta(t))) - V\mathbb{E}\{QoE(t)\} \\ &\leq A - V \sum_{k=1}^{K} QoE_k(t) + \sum_{k=1}^{K} Q_k(t) \left(D_k(t) - R_k(t) \right) \end{aligned} \tag{6.42}$$

where $A = \frac{1}{2}\mathbb{E}\left\{ \sum_{k=1}^{K} \left(R_k^2(t) + D_k^2(t) \right) \right\}$. For a given CSI, let consider $\varrho \in \Upsilon_{\Omega(t)}$ describes any policy that minimizes the R.H.S. of the Eq. 6.42. Theorem 6.1 gives the network stability and optimal QoE of our proposed algorithm.

Theorem 6.1: For $V \geq 0$ and known CSI, the network congestion or time-averaged bounded backlog of queues is given as

$$\frac{1}{FT}\sum_{t=0}^{FT-1}\sum_{k=1}^{K} Q_k(t) \leq \frac{AT + V(QoE^{max} - QoE^{min})}{\epsilon} + \frac{(T-1)\alpha}{2}, \tag{6.43}$$

and the performance of our proposed hybrid algorithm for HWAN is given as

$$\lim_{F\to\infty}\frac{1}{FT}\sum_{t=0}^{FT-1}\sum_{k=1}^{K} QoE_k(t) \geq \lim_{F\to\infty}\frac{1}{F}\sum_{j=0}^{F-1} QoE_j^* - \frac{AT}{V}, \tag{6.44}$$

where QoE_j^* is the optimal value of the j-th frame of the optimization problem shown in Eq. (6.40).
Proof: See Appendix B.

6.5 Simulation Results

This section evaluates our hybrid radio resource allocation and quality adaptation scheme using Matlab simulations. We consider a downlink multi-RAT heterogeneous simulation topology, as shown in fig. 6.1. The LTE BS covers the macro cell and is installed in the center of a cell of radius 500 m. Whereas 4 WLAN APs are deployed in the coverage of LTE BS, at a distance of 250 m from the macro BS. The LTE BS acts as an umbrella that provides maximum coverage, whereas the WLAN APs hotspot offers high throughput. Furthermore, we consider randomly distributed mobile users within the cell. We consider a scheduling slot of 10 ms and a system bandwidth of 18 MHz specified in LTE standards. In an OFDM-based system, each resource block has 12 subcarriers and 14 OFDM symbols with TTI = 1ms and subcarrier spacing of 15 kHz. Furthermore, for LTE, we select the carrier frequency as 3.5 GHz and transmitted power of 49 dBm. The data rate of the users connected to a WLAN AP is determined by applying the rate adaptation scheme based on the SNR threshold, already defined in [61] and is given in Table 6.1.

We consider four video files for the simulation study, each of length 200 video chunks of 0.5 seconds duration. We construct 800 video segments by using these four video files. Furthermore, the chunks are represented in different quality levels. The first and last 200 video chunks, i.e. from 1 to 200 and 601 to 800, are compressed into eight (8) different quality levels. In contrast, the remaining 400 video chunks from 201-400 and 401-600 are compressed into four (4) different quality representations. The quality measure of each video chunk is determined by using the structural similarity index (SSIM) [108]. To download the video file, each user picks a chunk (index) at random at $t = 0$ and requests to download the rest

TABLE 6.1: SNR and rate mapping [61]

SNR range (dB)	Rate (Mbps)
$>$ 24.56	54
[24.05, 24.56]	48
[18.8, 24.05]	36
[17.04, 18.8]	24
[10.79,17.04]	18
[9.03, 10.79]	12
[7.78, 9.03]	9
[6.02, 7.78]	6
$<$6.02	0

of the video file from that chunk onward. Mobile device energy consumption is a function of a bitrate of a video chunk related to the quality level of the selected chunk. The proposed scheme is simulated for 2000 time slots. Thus, the streaming session has a length of 200s, and the number of chunks for each user is 400. We assigned weights of $w_1 = 1$, $w_2 = 1$, and $w_3 = 1$ to the attributes of video quality, freezing time, and battery consumption. Users have the same preferences for all the three attributes. Figs. 6.2-6.4 shows the empirical CDF of the performance metrics, i.e., video quality and rebuffering state, which is related to freezing time. Fig. 6.2 shows the percentage of time spent in the rebuffering state. We analyze multiple streaming sessions per user and evaluate the percentage of the rebuffering state in the total playback time. Fig. 6.3 shows the video quality (SSIM) averaged over transmitted chunks obtained after analyzing multiple streaming sessions per user. The highest optimal value of the average video quality is 0.86, whereas the maximum value of video quality is 1. Fig. 6.4 shows the average video quality (SSIM) over transmitted chunks per user in a single streaming session, i.e., each user starts streaming at $t = 0$ and ends at requesting the 400th video chunk. Fig. 6.5 shows the playback buffer evolution over time t. The figure shows that the playback size reduces; however, there is no stall event in the streaming session. Fig. 6.6 shows the adaptability of video quality and energy consumption subject to the available battery level. From the graph, we can see that to maintain consistent streaming sessions, and lower quality video levels are requested to consume less energy since it is subject to the available energy level. When the available energy level is high, video chunks are requested at a higher quality level, enhancing user quality of experience.

6.6 Summary

In this chapter, we proposed a hybrid radio resource management scheme that operates at a two-time scale. A hybrid solution for radio resource management is achieved using the Lyapunov optimization technique that depends on the current information of CSI, transmission queue information, and the buffer state information. The radio resources are allocated at the

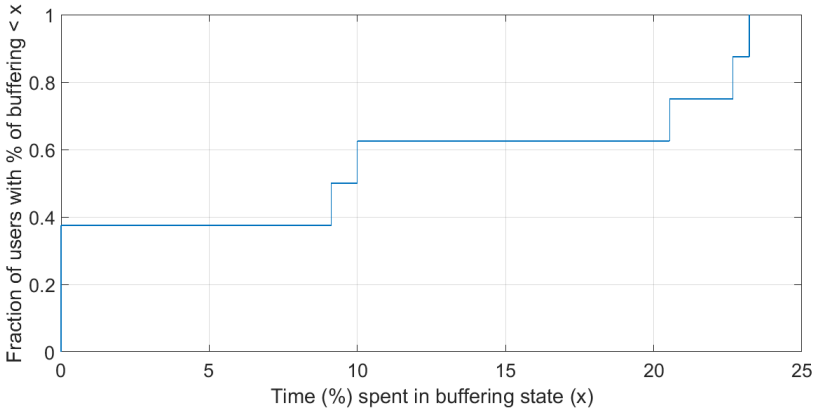

FIGURE 6.2: CDF plot of rebuffering state.

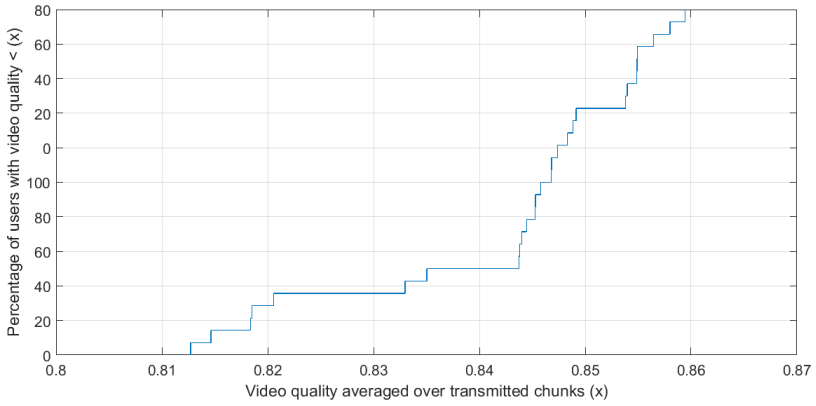

FIGURE 6.3: CDF of video quality averaged over multiple streaming sessions.

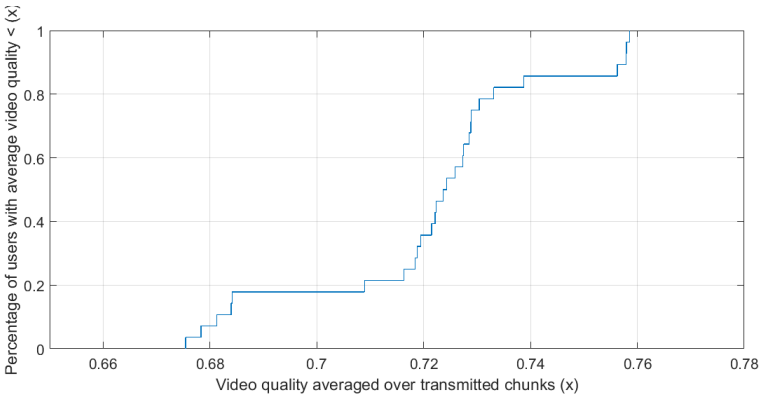

FIGURE 6.4: CDF of video quality averaged over transmitted chunks for a single streaming session.

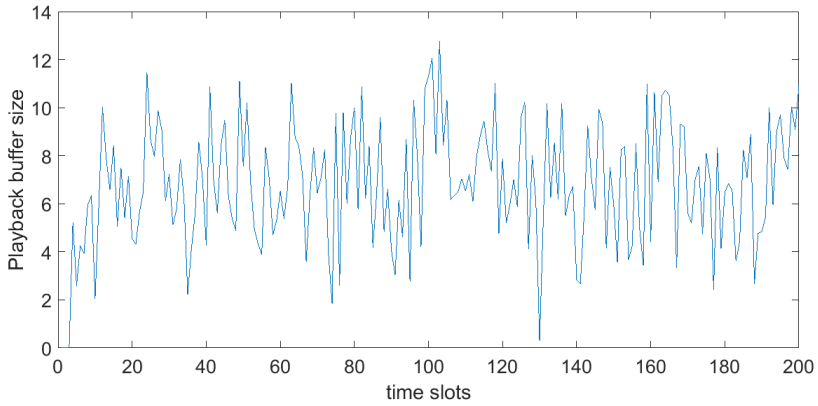

FIGURE 6.5: Playback buffer growth evaluation.

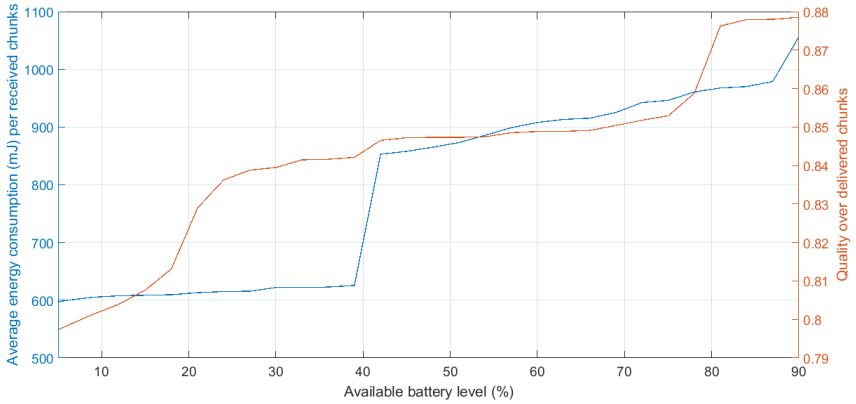

FIGURE 6.6: Energy consumption and video quality adaptation with the available battery level.

physical layer using scheduling time intervals. In contrast, quality adaptation is performed at chunk time intervals in a distributive fashion at each user end. The QoE maximization is formulated as a stochastic optimization problem. End-user QoE is a function of video quality, freezing time and energy consumption of the user device. We compare the QoE analysis of the proposed scheme with the optimal T-slot lookahead algorithm over an arbitrary sample path. Furthermore, the simulation results show the effectiveness of our proposed hybrid radio resource management scheme for video streaming over HWAN.

Chapter 7

Conclusion and Future Work

In this chapter, we conclude our thesis research main ideas and future research directions.

7.1 Conclusion

In this thesis, we have explored hybrid radio resource management for radio resource allocation, RAT selection and congestion management in HWAN. Based on network selection, radio resource allocation and congestion management analysis and discussion highlighted in this thesis, and we conclude the following:

- The HWAN provides multiple connectivity options and opportunities to enhance the perceived QoS of end-users. Therefore, it is crucial to investigate new techniques of RRM which can efficiently allocate radio resources from available networks and maintain the QoS requirements of the user.

- One of the key aspects of RRM is the selection of appropriate networks among the available networks. We proposed a hybrid approach that involves both users and CCN in decision-making. We characterized the performance of our HWAN for single and multihomed connections. The concept of utility is implemented for determining RATs ranking and association. We evaluated the impact of different criteria on each networks' final utility. Our multi-attribute function optimizes conflicting attributes such as cost, throughput and energy consumption. The Matlab-based simulation results show that our multi-criteria network selection scheme gives a more precise decision of network selection compared to the traditional distributive and centralized method of network selection.

- In HWAN, both single connection and multihomed connections coexist. Therefore, we have explored QoS-based radio resource allocation for multi-homing calls in HWAN. The Matlab-based simulation results of our proposed QoS-based radio resource allocation for multi-homing calls in HWAN show significant gains in the overall utility of our multihomed approach than the single RAT approach. Through simulation results,

it is shown that the convergence rate of the proposed scheme is independent of the minimum data rate requirement of each user.

- Load balancing in HWAN is explored in Chapter 4. Our proposed scheme overcomes the imbalance load condition by maintaining an equal load ratio across different RATs. The call blocking probability and bandwidth utilization show the usefulness of our proposed scheme.

- To design RRM, it is important to consider practical time-varying system models. We have proposed a hybrid RRM for a time-varying 5G HWAN to cope with the challenges of signalling overhead and computational complexity in centralized RRM. Our RRM is capable of performing three main tasks: 1) RAT selection, 2) optimal radio resource allocation, and 3) congestion control. Users perform RAT selection with network assistance. The joint problem of radio resource allocation and congestion control is decomposed such that a CCN allocates radio resources to the associated users at each time slot. In contrast, users play a significant role in congestion control by adapting its throughput according to its link-state information. Furthermore, we explored the tradeoff between network utility and delay of our proposed algorithm HCCRRA. We validated this tradeoff with simulation results and analytical calculations. The simulation results show that our proposed scheme outperforms the traditional schemes as it considers user throughput and network load.

- Multi-homing video transmission enhances user quality of experience (QoE). However, an improved video quality results in more power consumption of user device. Therefore, we consider user device energy consumption as a metric of QoE. The other two metrics are related to video, i.e. video quality and freezing time. We have proposed a hybrid RRM for streaming over a time-varying HWAN. Radio resources are allocated by selecting an appropriate video quality that maintains user QoE. Here both users and CCN take an active part in the process of radio resource allocation and quality adaptation. We investigated the performance analysis of the proposed algorithm. The performance of the proposed approach is also evaluated through simulation results.

7.2 Future Research Work

This research focused on hybrid radio resource management for radio resource allocation, rate adaptation, network selection, energy management and QoE-based radio resource allocation for video streaming in the heterogeneous wireless access network. There are open issues that need to be further explored. These issues are summarized as follows:

- In our research, we considered a HWAN cellular layout where a macro-BS acts as an umbrella covering the small BS/AP, which acts as a hot spot. The macro- BS provides

maximum coverage, whereas the hot spot provides a high data rate. We do not incorporate massive deployment of wireless networks. Therefore, exploring our hybrid RRM with dense HWAN cell deployment may lead to more interesting outcomes.

- In Chapter 2, we considered a central controller node where the BS/AP of different networks sends their signalling information. We assumed that the signalling information exchange among the users, the BS/AP, and CNN takes place on time. Therefore, further investigation is required to determine the impact of the accuracy of signalling information exchange and its delivery delay on hybrid RAT selection in HWAN.

- In Chapter 3, we explored radio resource allocation, i.e. subcarrier and power allocation from OFDMA based system and timeshare allocation from WLAN subject to the minimum QoS requirement of each user. We consider the only dynamic power consumption of the BS, which is related to the power consumed by the power amplifier that changes dynamically according to the transmit power. We do not incorporate the impact of static power consumption in our research. This static power consumption is related to baseband processing and power consumption at different circuit blocks such as analog-to-digital conversion, modulation, channel coding and signal detection. Hence, further investigation on the performance related to the energy efficiency of the entire system may give more interesting results.

- Since our proposed work in Chapter 5 has a rate adaptation policy and rate allocation policy; therefore, further exploring is required on the impact of imperfect CSI and QSI on the rate adaptation algorithm. We did not consider the QoS requirements of individual users. Therefore further investigation on the impact of minimum data rate requirement, i.e. QoS constraint on our radio resource allocation policy, may lead to more interesting results.

- The performance comparison of HCCRRA using multi-services scenarios such as variable bit rate (VBR) and constant bit rate (CBR) needs further exploration. Moreover, we assumed slow time-varying scenarios. However, in fast time-varying scenarios, users update BSs about their CSI every 10 ms. Extending our approach by considering fast time scales for radio resource allocation and congestion control policies will provide more realistic and useful results.

- In our HCCRRA algorithm, we did not consider the static power consumption at the BS. Investigating HCCRRA with a BS power model that includes both static power and dynamic power and incorporating the energy efficiency of the entire system will lead to more realistic results.

- In Chapter 6, we considered hybrid radio resource management for video streaming over time-varying HWAN. We consider video quality, freezing time and user device

power consumption as the QoE metrics. We did not consider the impact of video quality variation on user QoE. More complex analysis is required that includes video quality variation as an attribute of user QoE.

Appendix A

Appendix

A.1 Proof of Lemma 5.1

Let ω_1^* and ω_2^* be the optimal solution for optimization problems P1 and P2, respectively. Let U_1^* and U_2^* represent the optimal utility functions of problem P1 and P2, respectively. Applying the Jensen's inequality to the concave function $U(.)$, we have

$$U(\overline{\gamma}) \geq \overline{U(\gamma)} = U_2^*. \tag{A.1}$$

Let the solution ω_2^* satisfy the constraint C8, i.e. Eq. (5.19c) and $U(.)$ is non-decreasing, we have

$$U(\overline{\mathrm{d}}) \geq U(\overline{\gamma}). \tag{A.2}$$

Following the fact that since ω_2^* is feasible for the problem P2, it satisfies the constraints of the problem P1. Therefore, it is concluded that

$$U_1^* \geq U(\overline{\mathrm{d}}). \tag{A.3}$$

This leads us to finally conclude that $U_1^* \geq U_2^*$. The optimal solution ω_1^* of problem P1 is also feasible for problem P2, following the fact that C1-C7 are constraints of both original and transformed problem. We choose $\gamma_k(t) = \overline{d_k}^*$ for all time slots t along with ω_1^* results in feasible policy for problem P2, given as

$$U_2^* \geq \sum_{k \in \mathcal{K}} \overline{U(\gamma)} = \sum_{k \in \mathcal{K}} U(\overline{\mathrm{d}}) = U_1^*. \tag{A.4}$$

Eqs. (A.2 to A.4) conclude that the original and transformed problem are equivalent such that $U_1^* = U_2^*$.

A.2 Proof of Theorem 5.1

Let $\phi \in \mathrm{Y}_{\Omega(t)}$ describe any feasible decisions which minimize the R.H.S. of drift-minus reward function given in Eq. (5.24) and is re-written as

$$\begin{aligned}\Delta\Theta(t) - V\mathbb{E}\{U(\gamma(t))|\Theta(t)\} \leq Z - \sum_{k\in\mathcal{K}} \mathbb{E}\Big[Q_k(t)\{R_k(t)t_s \\ -D_k(t)\}|\Theta(t)\Big] - \sum_{k\in\mathcal{K}} \mathbb{E}\Big[\Gamma_k(t)\{D_k(t) - \gamma_k(t)\}|\Theta(t)\Big] \\ -V\mathbb{E}\{U(\gamma(t))|\Theta(t)\}.\end{aligned} \tag{A.5}$$

Then for any arbitrary $\varrho > 0$, there is a stationary and randomized policy ϕ^* with decision variables independent of queues (both $Q_k(t)$ and $\Gamma_k(t)$) at each time slot t satisfying the following constraints:

$$\mathbb{E}\{R_k^{\phi^*}(t)t_s - D_k^{\phi^*}(t)\} \leq \varrho, \tag{A.6}$$

$$\mathbb{E}\{D_k^{\phi^*}(t) - \gamma_k^{\phi^*}(t)\} \leq \varrho, \tag{A.7}$$

$$\mathbb{E}\{U(\gamma_k^{\phi^*}(t))\} \leq U^* - \varrho, \tag{A.8}$$

where U^* is the theoretical utility. It is further assumed that for any vector γ, we have

$$U^{min} \leq \sum_{k\in\mathcal{K}} U(\gamma_k) \leq U^{max}. \tag{A.9}$$

Putting Eq. (A.6) and Eq. (A.7) in Eq. (A.5) and using the assumption of Eq. (A.9) we get the following:

$$\begin{aligned}\Delta\Theta(t) - V\mathbb{E}\{U(\gamma(t))|\Theta(t)\} \leq Z - V\mathbb{E}\{U(\gamma(t))|\Theta(t)\} \\ -\varrho\sum_{k\in\mathcal{K}} \mathbb{E}[Q_k(t)]|\Theta(t)] \\ -\varrho\sum_{k\in\mathcal{K}} \mathbb{E}[\Gamma_k(t)|\Theta(t)].\end{aligned} \tag{A.10}$$

Summing the inequality over time slots $t \in \{0, 1, ..., T-1\}$, we get

$$\begin{aligned}\mathbb{E}\{L(\Theta(t+1))\} - \mathbb{E}\{L(\Theta(t))\} - \sum_{t=0}^{T-1} V\mathbb{E}\{U(\gamma(t))\} \\ \leq\ T(Z - VU^{min}) - \varrho\mathbb{E}\Big[\sum_{t=0}^{T-1}\sum_{k\in\mathcal{K}} Q_k(t)|\Theta(t)\Big] \\ -\varrho\mathbb{E}\Big[\sum_{t=0}^{T-1}\sum_{k\in\mathcal{K}} \Gamma_k(t)|\Theta(t)\Big].\end{aligned} \tag{A.11}$$

By rearranging Eq. (A.11) and ignoring $\Theta(t)$ as it is a randomized stationary policy, putting $L(\Theta(t+1)) > 0$ and $L(\Theta(0)) = 0$ and taking limits as $T \to \infty$ we get the following expression:

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\Big[\sum_{k\in\mathcal{K}} Q_k(t) + \sum_{k\in\mathcal{K}} \Gamma_k(t)\Big] \\ \leq \frac{Z + V(U^{max} - U^{min})}{\varrho}.\end{aligned} \tag{A.12}$$

Thus Eq. (5.49) is proved.

Now putting Eq. (A.6) and Eq. (A.7) into Eq. (A.5), and equating $\varrho \to 0$, we get

$$\Delta\Theta(t) - V\mathbb{E}\{U(\gamma(t))|\Theta(t)\} \leq\ Z - V\mathbb{E}\{U^*|\Theta(t)\}. \tag{A.13}$$

Ignoring the term $\Theta(t)$ for the randomized stationary policy, and the inequality summation over the time slot $t \in \{0, 1, ..., T-1\}$ gives us the following equation:

$$\begin{aligned}\mathbb{E}\{L(\Theta(t+1))\} - \mathbb{E}\{L(\Theta(0))\} - V\sum_{t=0}^{T-1} \mathbb{E}\{U(\gamma(t))\} \\ \leq Z - VU^*.\end{aligned} \tag{A.14}$$

Using the fact that $L(\Theta(t+1)) \geq 0$ and $L(\Theta(0)) = 0$, and further dividing Eq. (A.14) by T, we have

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\{U(\gamma(t))\} \geq U^* - \frac{Z}{V}. \tag{A.15}$$

We apply Jensen's inequality as the utility function is a non-decreasing concave function, we conclude that

$$U(\overline{d}) \geq \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\{U(\gamma(t))\} \geq U^* - \frac{Z}{V}, \tag{A.16}$$

which proves Eq. (5.50).

Appendix B

Appendix

B.1 Proof of Theorem 6.1

Consider a randomized policy $\varrho^* \in \Upsilon_{\Omega(t)}$ with decision variables $D_k^*(t)$, $R_k^*(t)$, $A_k^*(t)$, and $QoE_k^*(t)$ that may give a larger value of Eq. 6.42

$$\begin{aligned} &L(\Theta(t+1)) - L(\Theta(t))) - V\mathbb{E}\{\mathrm{QoE(t)}\} \\ &\leq A - V\sum_{k=1}^{K} QoE_k^*(t) + \sum_{k=1} Q_k(t)\left(D_k^*(t) - R_k^*(t)\right) \end{aligned} \tag{B.1}$$

Further, the queue evolution from one time slot to the next is bounded by α, and is given as

$$Q_k(t) - Q_k(jT) \leq (t - jT)\alpha, \forall k \in K \tag{B.2}$$

Putting Eq. B.2 in Eq. 6.42, we get the following

$$\begin{aligned} &L(\Theta(t+1)) - L(\Theta(t))) - V\mathbb{E}\{\mathrm{QoE(t)}\} \\ \leq &A + \sum_{k=1}^{K}\left((t-T)\alpha + Q_k(jT)\right)\left(D_k^*(t) - R_k^*(t)\right) \\ &- V\sum_{k=1}^{K} QoE_k^*(t) \end{aligned} \tag{B.3}$$

Summing the inequality over $t \in [jT, ..., (j+1)T - 1]$, and $\sum_{t=jT}^{jT+T-1}(t-T) = \frac{T(T-1)}{2}$, $D_k^*(t) - R_k^*(t) \leq \alpha_1$, and $\alpha\alpha_1 = 2A$, we get

$$\begin{aligned} &L(\Theta(j+1)T) - L(\Theta(jT))) - V \sum_{t=jT}^{jT+T-1} \{\mathrm{QoE(t)}\} \\ \leq & AT^2 + \sum_{k=1}^{K} Q_k(jT) \sum_{t=jT}^{jT+T-1} \left(D_k^*(t) - R_k^*(t) \right) \\ & - V \sum_{t=jT}^{jT+T-1} \sum_{k=1}^{K} \mathrm{QoE_k^*(t)} \end{aligned} \tag{B.4}$$

For any arbitrary $\epsilon > 0$, and a randomized policy ϱ^* at each time slot t satisfy the following constraints:

$$\frac{1}{T} \sum_{t=jT}^{jT+T-1} \left(D_k^*(t) - R_k^*(t) \right) \leq -\epsilon \tag{B.5}$$

It is further assumed that $\mathrm{QoE}^{\max}$ is the maximum value of QoE for frame j, we have

$$\mathrm{QoE}^{min} \leq \sum_{k=1}^{K} \mathrm{QoE}^* \leq \mathrm{QoE}^{max}. \tag{B.6}$$

$$\begin{aligned} &L(\Theta(j+1)T) - L(\Theta(jT))) - V \sum_{t=jT} \{\mathrm{QoE(t)}\} \\ \leq & AT^2 + VT \left(\mathrm{QoE_k^{max}(t)} - \mathrm{QoE_k^{min}(t)} \right) - \epsilon T \sum_{k=1}^{K} (Q_k(jT) \end{aligned} \tag{B.7}$$

Putting Eq. B.2 in Eq. B.7, we get

$$\begin{aligned} &L(\Theta(j+1)T) - L(\Theta(jT))) \\ \leq & AT^2 + VT \left(\mathrm{QoE_k^{max}(t)} - \mathrm{QoE_k^{min}(t)} \right) - \epsilon \sum_{t=jT}^{jT+T-1} \sum_{k=1}^{K} (Q_k(t) \\ & + \frac{T(T-1)\alpha\epsilon}{2} \end{aligned} \tag{B.8}$$

Summing Eq. B.8 over the frames $j \in [0, ..., F-1]$, we get

$$\begin{aligned} L(\Theta(FT) - L(\Theta(0))) \leq & AT^2F + VTF\times \\ & \left(\mathrm{QoE}_{\mathrm{k}}^{\max}(\mathrm{t}) - \mathrm{QoE}_{\mathrm{k}}^{\min}(\mathrm{t})\right) \\ & - \epsilon \sum_{j=0}^{FT} \sum_{k=1}^{K} Q_k(t) + \frac{FT(T-1)\alpha\epsilon}{2} \end{aligned} \tag{B.9}$$

Rearranging and ignoring the appropriate terms and taking the limits as $F \to \infty$, and dividing by ϵFT, we get the following expression

$$\begin{aligned} \frac{1}{FT} \sum_{t=0}^{FT-1} \sum_{k=1}^{K} Q_k(t) \leq & \frac{T(T-1)\alpha}{2} + \frac{AT}{\epsilon} \\ & + \frac{V\left(\mathrm{QoE}_{\mathrm{k}}^{\max}(\mathrm{t}) - \mathrm{QoE}_{\mathrm{k}}^{\min}(\mathrm{t})\right)}{\epsilon} \end{aligned} \tag{B.10}$$

Thus Eq. 6.43 is proved.

Let consider the optimal policy which acheives the optimal solution $\mathrm{QoE}^{\mathrm{opt}}(\mathrm{t})$. Updating Eq. B.4 with the optimal value of QoE, we get the following expression

$$\begin{aligned} & L(\Theta(j+1)T) - L(\Theta(jT))) - V \sum_{t=jT}^{jT+T-1} \{\mathrm{QoE}_{\mathrm{k}}(\mathrm{t})\} \\ \leq & AT^2 - VT\mathrm{QoE}_{\mathrm{j}}^{\mathrm{opt}}(\mathrm{t}) \end{aligned} \tag{B.11}$$

Summing Eq. B.11 over $j \in [0, ..., F]$, we get

$$\begin{aligned} & L(\Theta(FT) - L(\Theta(0))) - V \sum_{t=0}^{FT-1} \sum_{k=1}^{K} \{\mathrm{QoE}_{\mathrm{k}}(\mathrm{t})\} \\ \leq & AT^2F - VT \sum_{j=0}^{F-1} \mathrm{QoE}_{\mathrm{j}}^{\mathrm{opt}}(\mathrm{t}) \end{aligned} \tag{B.12}$$

Dividing both sides of Eq. B.12 by VFT, and setting $L(\Theta(FT)) \geq 0$, we get

$$\begin{aligned} \frac{1}{FT} \sum_{t=0}^{FT-1} \sum_{k=1}^{K} \{\mathrm{QoE}_{\mathrm{k}}(\mathrm{t})\} \geq & \frac{1}{F} \sum_{j=0}^{F-1} \mathrm{QoE}_{\mathrm{j}}^{\mathrm{opt}}(\mathrm{t}) \\ & - \frac{L(\Theta(0)))}{VFT} - \frac{AT}{V} \end{aligned} \tag{B.13}$$

Taking the limits as $F \to \infty$ and putting $L(\Theta(0)) = 0$, proves Eq. 6.44, i.e.,

$$\lim_{F\to\infty}\frac{1}{FT}\sum_{t=0}^{FT-1}\sum_{k=1}^{K}\{\mathrm{QoE_k(t)}\}\geq\lim_{F\to\infty}\frac{1}{F}\sum_{j=0}^{F-1}\mathrm{QoE_j^{opt}(t)}-\frac{AT}{V} \tag{B.14}$$

Bibliography

[1] CISCO, “Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022”, Feb. 2019.

[2] A. Maeder, A. Ali, A. Bedeker, A.F. Cattoni, D. Chandramouli et al., “A scalable and flexible radio access network architecture for fifth generation mobile networks,” IEEE Communications Magazine, vol. 54, no. 11, pp. 16-23, Nov. 2016.

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

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

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

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

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

[8] Y.-W. Chen, I.-H. Peng, and S.-T. Guan, “Dynamic bandwidth management for handoffs with RSVP in 802.16/WLAN environment,” in 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW’07), 2007, vol. 2, pp. 243–248.

[9] A. Wilson, A. Lenaghan, and R. Malyan, “Optimising wireless access network selection to maintain qos in heterogeneous wireless environments,” in Wireless Personal Multimedia Communications, 2005, pp. 18–22.

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

[11] G. Koundourakis, D. I. Axiotis, and M. Theologou, "Network-based access selection in composite radio environments," in 2007 IEEE Wireless Communications and Networking Conference, 2007, pp. 3877–3883.

[12] O. Ormond, P. Perry, and J. Murphy, "Network selection decision in wireless heterogeneous networks," in 2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications, 2005, vol. 4, pp. 2680–2684.

[13] M. Ismail, A. Abdrabou, and W. Zhuang, "Cooperative decentralized resource allocation in heterogeneous wireless access medium," IEEE Transactions on Wireless Communications, vol. 12, no. 2, pp. 714–724, 2012.

[14] N. Zarin and A. Agarwal, "A hybrid network selection scheme for heterogeneous wireless access network," in 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1–6, 2017.

[15] N. Zarin and A. Agarwal, "QoS based joint radio resource allocation for multi-Homing calls in heterogeneous wireless access network," in Proc. of the 16th ACM International Symposium on Mobility Management and Wireless Access, pp. 37-42, Oct. 2018.

[16] N. Zarin and A. Agarwal, "A centralized approach for load balancing in heterogeneous wireless access network,"in 2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE), pp. 1–5, 2018.

[17] N. Zarin and A. Agarwal, "Hybrid Radio Resource Management for Time-Varying 5G Heterogeneous Wireless Access Network," IEEE Transactions on Cognitive Communications and Networking, vol. 7, no. 2, pp. 594–608, 2021.

[18] Chan, Pauline ML, Yim-Fun Hu, and Ray E. Sheriff. "Implementation of fuzzy multiple objective decision making algorithm in a heterogeneous mobile environment." in Wireless Communications and Networking Conference, Vol. 1. IEEE, pp. 332–336, 2002.

[19] J. Hou and D. C. O'Brien, "Vertical handover-decision-making algorithm using fuzzy logic for the integrated Radio-and-OW system," IEEE Transactions on Wireless Communications, vol. 5, no. 1, pp. 176–185, 2006.

[20] P. M. 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, vol. 39, no. 12, pp. 42–51, 2001.

[21] Cesana, Matteo, Nicola Gatti, and Ilaria Malanchini. "Game theoretic analysis of wireless access network selection: models, inefficiency bounds, and algorithms." Proceedings of the 3rd International Conference on Performance Evaluation Methodologies and Tools. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008.

[22] D. Niyato and E. Hossain, "Dynamics of network selection in heterogeneous wireless networks: An evolutionary game approach," IEEE transactions on vehicular technology, vol. 58, no. 4, 2008.

[23] K. Zhu, D. Niyato, and P. Wang, "Network selection in heterogeneous wireless networks: Evolution with incomplete information," in IEEE Wireless Communication and Networking Conference, pp. 1–6, 2010.

[24] C. Sun, E. Stevens-Navarro, V. Shah-Mansouri, and V. W. Wong, "A constrained MDP-based vertical handoff decision algorithm for 4G heterogeneous wireless networks," Wireless Networks, vol. 17, no. 4, pp. 1063–1081, 2011.

[25] Q. Song and A. Jamalipour, "A quality of service negotiation-based vertical handoff decision scheme in heterogeneous wireless systems," European Journal of Operational Research, vol. 191, no. 3, pp. 1059–1074, 2008.

[26] L. Wang and D. Binet, "MADM-based network selection in heterogeneous wireless networks: A simulation study," in 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, pp. 559–564, 2009.

[27] P. Kosmides, A. Rouskas, and M. Anagnostou, "Network selection in heterogeneous wireless environments," in 18th International Conference on Telecommunications, pp. 250–255, 2011.

[28] X. Gelabert, J. Perez-Romero, O. Sallent, and R. Agusti, "A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks," IEEE Transactions on Mobile Computing, vol. 7, no. 10, pp. 1257–1270, 2008.

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

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

[31] C. Chi, X. Cai, R. Hao, and F. Liu, "Modeling and analysis of handover algorithms," in IEEE GLOBECOM 2007-IEEE Global Telecommunications Conference, pp. 4473–4477, 2007.

[32] W. Shen and Q.A. Zeng, "Resource management schemes for multiple traffic in integrated heterogeneous wireless and mobile networks," in Proceedings of 17th International Conference on Computer Communications and Networks, pp. 1–6, 2008.

[33] E. S. Navarro, Y. Lin, and W. S. Wong, "An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks," IEEE Transaction on Vehicular Technology, vol. 57, no. 2, pp. 1243-1254, 2008.

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

[35] A. Abdelhadi, M. Ghorbanzadeh, and C. Clancy, "Optimal radio resource allocation for hybrid traffic in cellular networks: Centralized and distributed architecture," arXiv preprint arXiv, pp.1411.4011, 2014.

[36] A. A. Sabbagh, R. Braun, and M. Abolhasan, "A mobility optimization CRRM approach for Next Generation Wireless Networks," in 2012 International Conference on Computer & Information Science (ICCIS), vol. 2, pp. 609–613, 2012.

[37] M. El Helou, S. Lahoud, M. Ibrahim, K. Khawam, B. Cousin, and D. Mezher, "A hybrid approach for radio access technology selection in heterogeneous wireless networks," Wireless Personal Communications, vol. 86, no. 2, pp. 789–834, 2016.

[38] L. Wang and G.-S. G. Kuo, "Mathematical modeling for network selection in heterogeneous wireless networks—A tutorial," IEEE Communications Surveys & Tutorials, vol. 15, no. 1, pp. 271–292, 2012.

[39] Q.T. Nguyen-Vuong, N. Agoulmine, E. H. Cherkaoui, and L. Toni, "Multicriteria optimization of access selection to improve the quality of experience in heterogeneous wireless access networks," IEEE Transactions on Vehicular Technology, vol. 62, no. 4, pp. 1785–1800, 2012.

[40] I. Tsompanidis, A. H. Zahran, and C. J. Sreenan, "A utility-based resource and network assignment framework for heterogeneous mobile networks," in IEEE Global Communications Conference (GLOBECOM), pp. 1–6, 2015.

[41] R. Trestian, O. Ormond, and G.-M. Muntean, "Power-friendly access network selection strategy for heterogeneous wireless multimedia networks," in 2010 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1–5, 2010.

[42] S. Navaratnarajah, M. Dianati, and M. A. Imran, "Analysis of energy efficiency on the cell range expansion for cellular-WLAN heterogeneous network," in 2015 International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 514–519, 2015.

[43] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11 b," in IEEE INFOCOM. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societie, vol. 2, pp. 836–843, 2003.

[44] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in Proceedings of IEEE INFOCOM, pp. 998–1006, 2013.

[45] M. A. Senouci, S. Hoceini, and A. Mellouk, "Utility function-based TOPSIS for network interface selection in heterogeneous wireless networks," in IEEE international conference on communications (ICC), pp. 1–6, 2016.

[46] R. Amin, J. Martin, J. Deaton, L. A. DaSilva, A. Hussien, and A. Eltawil, "Balancing spectral efficiency, energy consumption, and fairness in future heterogeneous wireless systems with reconfigurable devices," IEEE Journal on Selected Areas in Communications, vol. 31, no. 5, pp. 969–980, 2013.

[47] 3GPP, IP flow mobility and seamless wireless local area network (WLAN) offload. TS 23.261, V10.2.0, Mar. 2012.

[48] K. Chebrolu and R. R. Rao, "Bandwidth aggregation for real-time applications in heterogeneous wireless networks," IEEE Transactions on Mobile Computing, vol. 5, no. 4, pp. 388–403, 2006..

[49] X. Wang and G. B. Giannakis, "Resource allocation for wireless multiuser OFDM networks," IEEE Transactions on Information theory, vol. 57, no. 7, pp. 4359–4372, 2011.

[50] P. Xue, P. Gong, J. H. Park, D. Park, and D. K. Kim, "Radio resource management with proportional rate constraint in the heterogeneous networks," IEEE Transactions on Wireless Communications, vol. 11, no. 3, pp. 1066–1075, 2011.

[51] A. R. Ekti, X. Wang, M. Ismail, E. Serpedin, and K. A. Qaraqe, "Joint user association and data-rate allocation in heterogeneous wireless networks," IEEE Transactions on Vehicular Technology, vol. 65, no. 9, pp. 7403–7414, 2015.

[52] J. Zou, Q. Xi, Q. Zhang, C. He, L. Jiang, and J. Ding, "QoS-aware energy-efficient radio resource allocation in heterogeneous wireless networks," in IEEE International Conference on Communication Workshop (ICCW), pp. 2781–2786, 2015.

[53] A. R. Elsherif, W.-P. Chen, A. Ito, and Z. Ding, "Adaptive small cell access of licensed and unlicensed bands," in IEEE International Conference on Communications (ICC), pp. 6327–6332, 2013.

[54] X. Zhang and F. Yang, "Joint bandwidth and power allocation for energy efficiency optimization over heterogeneous LTE/WiFi multi-homing networks," in IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6, 2017.

[55] Y. Li et al., "Energy-efficient transmission in heterogeneous wireless networks: A delay-aware approach," IEEE Transactions on Vehicular Technology, vol. 65, no. 9, pp. 7488–7500, 2015.

[56] S. Kim, B. G. Lee, and D. Park, "Energy-per-bit minimized radio resource allocation in heterogeneous networks," IEEE Transactions on Wireless Communications, vol. 13, no. 4, pp. 1862–1873, 2014.

[57] U. Toseef, Y. Zaki, C. Görg, and A. Timm-Giel, "Development of simulation environment for multi-homed devices in integrated 3GPP and non-3GPP networks," in Proceedings of the 10th ACM international symposium on Mobility management and wireless access, pp. 29–36, 2012.

[58] J. Marašević, J. Zhou, H. Krishnaswamy, Y. Zhong, and G. Zussman, "Resource allocation and rate gains in practical full-duplex systems," IEEE/ACM Transactions on Networking, vol. 25, no. 1, pp. 292–305, 2016.

[59] Y. Liu, L. Cuthbert, X. Yang, and Y. Wang, "QoS-aware resource allocation for multimedia users in a multi-cell spectrum sharing radio network," in Proceedings of the 7th ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks, pp. 45–52, 2012.

[60] S. Boyd, S. P. Boyd, and L. Vandenberghe, Convex optimization. Cambridge university press, 2004.

[61] J. Yee, and H. P.-Esfahani, "Understanding wireless LAN performance trade-offs, " Communication Systems Design, vol. 11, pp. 32-35, 2002.

[62] J. Andrews, "Seven ways that hetnets are a cellular paradigm shift," IEEE Communication Magazine, vol. 51, no. 3, pp. 136-144, Mar. 2013.

[63] M. Bennis et al., "When cellular meets wifi in wireless small cell networks," IEEE Communication Magazine, vol. 51, no. 6, pp. 44-50, Jun. 2013.

[64] K.K. Yap et al., "Making use of all the networks around us: a case study in android," in Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design, pp. 19–24, 2012.

[65] I. Blau, G. Wunder, I. Karla, and R. Sigle, "Decentralized utility maximization in heterogeneous multicell scenarios with interference limited and orthogonal air interfaces," EURASIP Journal on Wireless Communications and Networking, vol. 2009, pp. 1–12, 2009.

[66] B. Li and D. Yang, "An effective cooperative load balancing scheme for heterogeneous network," in 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), pp. 1–5, 2011.

[67] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," IEEE Transactions on Wireless Communications, vol. 12, no. 5, pp. 2484–2497, 2013.

[68] C. Xue, J. Luo, R. Halfmann, E. Schulz, and C. Hartmann, "Inter gw load balancing for next generation mobile networks with flat architecture," in VTC Spring IEEE 69th Vehicular Technology Conference, pp. 1–5, 2009.

[69] G. H. Carvalho, I. Woungang, A. Anpalagan, and E. Hossain, "QoS-aware energy-efficient joint radio resource management in multi-RAT heterogeneous networks," IEEE Transactions on Vehicular Technology, vol. 65, no. 8, pp. 6343–6365, 2015.

[70] H. Son, S. Lee, S.-C. Kim, and Y.-S. Shin, "Soft load balancing over heterogeneous wireless networks," IEEE Transactions on Vehicular Technology, vol. 57, no. 4, pp. 2632–2638, 2008.

[71] M. Anedda, G.-M. Muntean, and M. Murroni, "Adaptive real-time multi-user access network selection algorithm for load-balancing over heterogeneous wireless networks," in IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1–4, 2016.

[72] I.-P. Belikaidis et al., "Multi-RAT dynamic spectrum access for 5G heterogeneous networks: The SPEED-5G approach," IEEE Wireless Communications, vol. 24, no. 5, pp. 14–22, 2017.

[73] F. Jiang, Y. Liu, B. Wang, and X. Wang, "A relay-aided device-to-device-based load balancing scheme for multitier heterogeneous networks," IEEE Internet of Things Journal, vol. 4, no. 5, pp. 1537–1551, 2017.

[74] ITU-R, "IMT vision–framework and overall objectives of the future development of IMT for 2020 and beyond," International Telecommunication Union, Switzerland, Recommendation ITU-R M.2083-0, Sep. 2015.

[75] A. Maeder, A. Ali, A. Bedeker, A.F. Cattoni, D. Chandramouli et al., "A scalable and flexible radio access network architecture for fifth generation mobile networks," IEEE Communications Magazine, vol. 54, no. 11, pp. 16-23, Nov. 2016.

[76] Access, Evolved Universal Terrestrial Radio, "Study on small cell enhancements for E-UTRA and E-UTRAN— higher layer aspects," 3GPP, v.12.0.0, Dec. 2013.

[77] J. Krause, "Study on scenarios and requirements for next generation access technologies," 3GPP TR 38.913, Sept. 2016.

[78] A. Ghosh et al., "Heterogeneous cellular networks: From theory to practice," IEEE Communications Magazine, vol. 50, no. 6, pp. 54-64, 2012.

[79] Y. Guo, Q. Yang, F. Fu, and K. S. Kwak, "Quality-oriented rate control and resource allocation in dynamic OFDMA networks," in IEEE Global Communications Conference (GLOBECOM), pp. 1-6, Dec. 2015.

[80] Kim, G. Caire, and A. F. Molisch, "Quality-aware streaming and scheduling for device-to-device video delivery," IEEE/ACM Transactions on Networking, vol. 24, no. 4, pp. 2319-2331, Aug. 2016.

[81] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," IEEE Wireless Communications, vol. 22, no. 2, pp. 152-160, Apr. 2015.

[82] Y. Guo, Q. Yang, J. Liu, and K.S. Kwak, "Quality-aware Streaming in heterogeneous wireless networks", IEEE Transactions on Wireless Communications, vol. 16, no.12, pp. 8162-8174, Oct. 2017.

[83] V.F. Monteiro, D. A. Sousa,T. F. Maciel, F. R. P. Cavalcanti, C. F. e Silva, and E.B. Rodrigues, "Distributed RRM for 5G Multi-RAT Multiconnectivity Networks." IEEE Systems Journal, vol.13, no. 1 pp. 192-203, Jun. 2018.

[84] S. Agarwal and S. De, "Cognitive multihoming system for energy and cost aware video transmission," IEEE Transactions on Cognitive Communications and Networking, vol. 2, no. 3, pp. 316-329, Sep. 2016.

[85] Y. Guo, Q. Yang, F.R. Yu and V.C. Leung, "Dynamic quality adaptation and bandwidth allocation for adaptive streaming over time-varying wireless networks," IEEE Transactions on Wireless Communications, vol. 16, no. 12, pp. 8077-8091, Sep. 2017.

[86] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," Synthesis Lectures on Communication Networks, vol. 3, no. 1, pp. 1-211, 2010.

[87] D. Bethanabhotla, G. Caire and M. J. Neely, "Adaptive video streaming for wireless networks with multiple users and helpers," IEEE Transactions on Communications, vol. 63, no. 1, pp. 268-285, Dec. 2014.

[88] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," IEEE/ACM Transactions on networking, vol. 8, no. 5, pp. 556–567, 2000.

[89] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception" TS 36.101, v. 10.9.0, Dec. 2012.

[90] J. Li, M. Peng,Y. Yu and Z. Ding, 2016. "Energy-efficient joint congestion control and resource optimization in heterogeneous cloud radio access networks." IEEE Transactions on Vehicular Technology, vol. 65, no. 12, pp. 9873-9887, Feb. 2016.

[91] 3GPP, "Study on 3D channel model for LTE." RP. 36.873, Dec. 2017.

[92] 3GPP, "Study on channelmodel for frequencies from 0.5 to 100 GHz" Rep. 38.901, v.14.3.0, Dec. 2018.

[93] Y. Sànchez, "iDASH: Improved dynamic adaptive streaming over HTTP using scalable video coding," in Proc. ACM Multimedia System, pp. 257–264, 2011.

[94] N. Eswara, S. Chakraborty, H. P. Sethuram, K. Kuchi, A. Kumar, and S. S. Channappayya, "Perceptual QoE-Optimal Resource Allocation for Adaptive Video Streaming," IEEE Transactions on Broadcasting, vol. 66, no. 2, pp. 346–358, 2019

[95] K. Miller, D. Bethanabhotla, G. Caire, and A. Wolisz, "A control-theoretic approach to adaptive video streaming in dense wireless networks," IEEE Transactions on Multimedia, vol. 17, no. 8, pp. 1309–1322, 2015.

[96] L. Zou, R. Trestian, and G.-M. Muntean, "E3DOAS: Balancing QoE and energy-saving for multi-device adaptation in future mobile wireless video delivery," IEEE Transactions on Broadcasting, vol. 64, no. 1, pp. 26–40, 2017.

[97] Z. Deng, Y. Liu, J. Liu, X. Zhou, and S. Ci, "QoE-oriented rate allocation for multipath high-definition video streaming over heterogeneous wireless access networks," IEEE Systems Journal, vol. 11, no. 4, pp. 2524–2535, 2015.

[98] C. Díaz, A. Fernández, F. Sacristán, and N. García, "Energy-and Quality-Aware Video Request Policy for Wireless Adaptive Streaming Clients," IEEE Transactions on Consumer Electronics, vol. 66, no. 4, pp. 366–375, 2020.

[99] A. Begen, T. Akgul, and M. Baugher, "Watching video over the web: Part 1: Streaming protocols," IEEE Internet Computing, vol. 15, no. 2, pp. 54–63, 2011.

[100] D. Bethanabhotla, G. Caire, and M. J. Neely, "WiFlix: Adaptive video streaming in massive MU-MIMO wireless networks," IEEE Transactions on Wireless Communications, vol. 15, no. 6, pp. 4088–4103, 2016.

[101] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[102] A. Ortega, "Variable bit rate video coding," in Compressed Video over Networks, CRC Press, pp. 353–392, 2018.

[103] C. Zhou, C.-W. Lin, and Z. Guo, "mDASH: A Markov decision-based rate adaptation approach for dynamic HTTP streaming," IEEE Transactions on Multimedia, vol. 18, no. 4, pp. 738–751, 2016.

[104] J. Qiao, X. S. Shen, J. W. Mark, and L. Lei, "Video quality provisioning for millimeter wave 5G cellular networks with link outage," IEEE Transactions on Wireless Communications, vol. 14, no. 10, pp. 5692–5703, 2015.

[105] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over HTTP," in Proceedings of the ACM Conference on Special Interest Group on Data Communication, pp. 325–338, 2015.

[106] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: a measurement study and implications for network applications," in Proc. 9th ACM SIGCOMM Internet Measurement Conference, 2009.

[107] M. Li and C. Y. Lee, "A cost-effective and real-time QoE evaluation method for multimedia streaming services," Telecommun. Syst., vol. 59, no. 3, pp. 317–327, Jul. 2015.

[108] The SSIM Index for Image Quality Assessment. [Online]. Available:http://goo.gl/ngR0UL