LOAD BALANCING IN HETEROGENEOUS WIRELESS COMMUNICATIONS NETWORKS

Optimized load aware vertical handovers in satellite-terrestrial hybrid networks incorporating IEEE 802.21 media independent handover and cognitive algorithms

M. ALI

Submitted for the Degree of

Doctor of philosophy

School of Engineering, Design and Technology
University of Bradford

2012

ABSTRACT

Name: Muhammad Ali

Thesis Title: Load balancing in heterogeneous wireless communications networks: Optimized load aware vertical handovers in satellite-terrestrial hybrid networks incorporating IEEE 802.21 media independent handover and cognitive algorithms.

Keywords: Heterogeneous wireless networks, vertical handover, Media Independent Handover (MIH), load balancing in heterogeneous networks, satellite networks, cognitive algorithms for load balancing, fuzzy load balancing, fuzzy-neural load balancing, UMTS, WiMax.

Heterogeneous wireless networking technologies such as satellite, UMTS, WiMax and WLAN are being used to provide network access for both voice and data services. In big cities, the densely populated areas like town centres, shopping centres and train stations may have coverage of multiple wireless networks. Traditional Radio Access Technology (RAT) selection algorithms are mainly based on the 'Always Best Connected' paradigm whereby the mobile nodes are always directed towards the available network which has the strongest and fastest link. Hence a large number of mobile users may be connected to the more common UMTS while the other networks like WiMax and WLAN would be underutilised, thereby creating an unbalanced load across these different wireless networks. This high variation among the load across different co-located networks may cause congestion on overloaded network leading to high call blocking and call dropping probabilities. This can be alleviated by moving mobile users from heavily loaded networks to least loaded networks.

This thesis presents a novel framework for load balancing in heterogeneous wireless networks incorporating the IEEE 802.21 Media Independent Handover (MIH). The framework comprises of novel load-aware RAT selection techniques and novel network load balancing mechanism. Three new different load balancing algorithms i.e. baseline, fuzzy and neural-fuzzy algorithms have also been presented in this thesis that are used by the framework for efficient load balancing across the different co-located wireless networks. A simulation model developed in NS2 validates the performance of the proposed load balancing framework. Different attributes like load distribution in all wireless networks, handover latencies, packet drops, throughput at mobile nodes and network utilization have been observed to evaluate the effects of load balancing using different scenarios. The simulation results indicate that with load balancing the performance efficiency improves as the overloaded situation is avoided by load balancing.

Acknowledgments

My PhD research study has been a very interesting journey of learning and hardworking, where I received inspirations from various people which supported me in distinct way. I would like to take this opportunity to thank all those people at this very moment.

First of all I would like to extend my deepest gratitude to my supervisors Dr Prashant Pillai and Prof Yim-Fun Hu. I am very grateful to Prof Hu for providing me this opportunity to carry out my research study and for her guidance and support throughput my PhD study. I am also thankful to her for providing me opportunity to work in different projects. Dr Prashant who has been a constant source of encouragement and has always inspired me to work hard. Dr Prashant has been a great advisor and I am thankful to him for detailed discussions on my research and guiding me for improving my research and developing skills. These have helped me in understanding various aspects of heterogeneous wireless networking. I would also like to thank both Prof Hu and Dr Prashant for spending time to understand the technical aspects of thesis and for providing encouraging and constructive comments to improve it. I also want to thank all my colleagues here at Future Ubiquitous Networks (FUN) lab for their encouragement and support over the past years.

Most importantly I would like to thank my parents for their constant love and support. Especially my mom, she has always encouraged me to work hard and my dad who has been working hard all his life to make our future bright. Finally I would like to thank my parents and my (late) grandparents for their sincere and loving prayers in my favour.

To my parents Muhammad Ishfaq and Irshad Begum

& my wife Ayesha Ali

Table of Contents

List of Figures

List of Tables

LIST OF ACRONYMS

Acronym	Full Form
2G	2nd Generation
3G	3rd Generation
AAA	Authentication, Authorisation and Accounting
ABC	Always Best Connected
ANN	Artificial Neural Network
AP	Access Point
APA	Adaptive Power Allocation
AR	Available Resources
ARQ	Automatic Repeat Request
ASN	Access Service Network
ATM	Asynchronous Transfer Mode
B3G	Beyond 3G
BGAN	Broadband Global Area Network
BS	Base Station
BSS	Base Station Subsystem
CA	Coverage Area of Network
CAC	Call Admission Control
CAPEX	Capital Expense
CCK	Complementary Code Keying
CN	Core Network
CSMA	Carrier Sense Multiple Access
CSN	Connectivity Service Network
CS-SAP	Convergence Sublayer Service Access Point
DCF	Distributed Coordination Function
DSSS	Direct Sequence Spread Spectrum
DVB	Digital Video Broadcast
DVB-RCS	Digital Video Broadcast Return Channel Satellite
FEC	Forward Error Correction
FL	Fuzzy Logic
FLC	Fuzzy Logic Controller
GA	Genetic Algorithm
GEO	Geostationary Earth Orbit
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communication
IAI-2	INMARSAT Air Interface 2
IP	Internet Protocol
JCAC	Joint Call Admission Control
LEO	Low Earth Orbit
LL	Least loaded
LOS	Line of Sight

MAC	Medium Access Control
MAHO	Mobile Assisted Handover
MCHO	Mobile Controlled Handover
MEO	Medium Earth Orbit
MICS	Media Independent Command Service
MIES	Media Independent Event Service
MIH	Media Independent Handover
MIHF	Media Independent Handover Function
MIIS	Media Independent Information Service
MLME	MAC Layer Management Entity
MNs	Mobile Nodes
MS	Mobile Station
MSC	Mobile services Switching Centre
MSCs	Message Sequence Charts
NAHO	Network Assisted Handover
NAM	Network Animator
NAS	Non Access Stratum
NCHO	Network Controlled Handover
NCMS	Network Control Management System
NF	Neural Fuzzy
NLOS	Non Line of Sight
NN	Neural Network
OBP	On-Board Processing
ODR	Offered Data Rate
OPEX	Operating Expense
OFDM	Orthogonal Frequency Division Multiple Access
PAN	Personal Area Network
PHY	Physical Layer
PoA	Point of Attachment
PoS	Point of Service
QoS	Quality of Service
RAN	Radio Access Network
RATs	Radio Access Technologies
RDR	Required Data Rate
RNC	Radio Network Controller
RRM	Radio Resource Management
RU	Resource Utilization
SC	Single Carrier
SDR	Software Defined Radio
SGSN	Serving GPRS Support Node
SLA	Service Level Agreement
SMS	Short Messaging Services

SS	Signal Strength
UE	User Equipment
UMTS	Universal Mobile Telecommunication System
UNPs	User's Network Preferences
VLR	Visitor Location Register
WiMax	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network
WWAN	Wireless Wide Area Network

Chapter 1:   INTRODUCTION

1.1 Overview

In the past decade, there has been a remarkable growth in the use of wireless and mobile communications. While on one hand the number of users accessing such services has increased, the amount of data traffic and types of applications have also increased. While traditionally mobile networks were predominantly for voice communications, the advent of 3G technology has seen a rise in the use of data services also. Hence, these wireless and mobile networks are now used for different types of voice and data communications. The users of these networks expect anytime, anyplace good service. To cater to this ever increasing demand for data services has led to the development of various radio access technologies like 3G, 4G and IEEE 802.16 WiMax that support high data rates and long communication ranges. At the same time there has been an increase in the use of satellite networks for data communications, especially in rural areas lacking terrestrial infrastructure, and for aeronautical and maritime communications. It is envisaged that in the future these different networks would need to collaborate in order to meet the ever increasing user demands for seamless broadband services on the move. Such collaborative heterogeneous networks may be managed by the same network service provider. These days, some wireless operators are providing their services over not only 2G and 3G cellular networks, but also Wireless Local Area Network (WLAN) and IEEE 802.16 WiMax. If these future networks are managed by different

service providers, then it is assumed that they would have Service Level Agreements (SLA) in place to support the co-operation.

The last few years have also seen various improvements in the development of the end user terminals especially with the advent of smart phones and tablets. These new devices have many features like reduced size, increased battery time, support for video calls and internet browsing. It has also been seen that these new devices usually support multiple interfaces for different radio access technologies which allow these devices to be connected to different wireless and mobile networks at the same time. The latest trends of research and development show that soon multimode terminals Software Defined Radio (SDR) would be available that will be using to also reconfigure their radio as per requirements and availability [1].

Modern mobile devices like smart phones, PDA's and Tablet PCs already support multiple wireless technologies simultaneously like UMTS, WLAN and Bluetooth and in the very near future could also support WiMax. While most of these devices are able to scan for the different available networks, the user would manually select which network he or she may want to use. So a user may connect to UMTS for voice services and may use WLAN to access the data services. It is envisaged that in the near future these user terminal may be able to apply some complex Radio Access Technology (RAT) selection techniques to find the most suitable network from the available networks. Such a RAT selection technique may need to consider various parameters like the received signal strengths, errors rates, costs, user preferences, QoS requirements, etc. Such a RAT selection technique would not only play an important part when a user turns on power of his or her

mobile device but also when the user moves around between the coverage areas of different wireless networks.

In order to support the mobility of users, most of the today mobile networks already support seamless handovers. However these are restricted to handovers within the same technology, i.e. horizontal handovers. It is envisaged that to efficiently use the network services the future mobile devices shall also support handovers across different radio access technologies. This process of switching mobile devices connectivity from one technology to another type of technology is called vertical handover. These future user terminals would be able to scan the various available networks of different access technologies and then use sophisticated RAT selection algorithms to select the most suitable network. Such an algorithm would generally be user-centric whereby they would consider user preferences in the decision making process for RAT selection. However such user-centric RAT selection algorithm often leads to highly unbalanced load across the different networks as users may tend to want the cheapest link or the link with the best signal strength, etc. Hence a scenario may arise where a large number of users are connected to one network but another network which is also available in that given area has very few users. In such a case, the heavily loaded network may face congestion and eventually result in call blocking and call dropping. This unbalanced load situation in networks with overlapping coverage area also causes the poor radio resource utilization as some networks remain lightly loaded and some get overloaded.

1.2 Problem statement

In wireless communication networks, the increasing number of mobile subscribers and dynamically abrupt changes in number of active mobile users is a real challenge for the network providers as it leads to real time load variations in the network. This dynamic change in load on a network is due to many reasons like peak hours at hot spots or motorways, special events like football match, exhibition or a festival celebration. The network performance gets significantly degraded at the time when network gets heavily loaded.

In urban areas it is common in most places that multiple networks provide coverage over the same geographically located area. For example a busy town market area may possess coverage of WLAN, cellular networks like WiMax and UMTS, and satellite networks. In this context while one of the available networks in particular area gets overloaded, other networks covering the same geographical area may remain lightly loaded. This results in poor utilisation of available wireless resources and poor network performance, thereby poor user experience. While network operators considered users' population density and mobility patterns for planning network deployment, each service provider would be required to have large infrastructure in place to cater to the needs of their users in these densely populated areas. Hence the different networks of heterogeneous wireless networks, whose coverage areas overlap experience imbalance of radio resource utilization and performance degradation of due to the unbalanced load across the different wireless networks.

Traditional RAT selection algorithms are mainly based on the Always Best Connected (ABC) paradigm whereby the mobile nodes are always directed

towards the available network which has the strongest, fastest or cheapest link. This however could create a high variation among the load across the different co-located networks thereby causing congestion on overloaded network and eventually increase in call blocking and call dropping probabilities. The unbalanced load situation in co-located networks also causes the poor radio resource utilisation as some networks remain under loaded and some become overloaded. Hence there is a need for some load balancing strategies to efficiently utilise the available radio resources and avoids these unwanted congestion situations on overloaded wireless networks.

1.3   Target solutions

Heavily loaded networks can accumulate several drawbacks as discussed in the previous. These drawbacks can be overcome by looking at various ways of collaboration between wireless access technologies and to maximise their utilisation. There are two ways to avoid radio networks capacity shortage:

- Increasing the resource capacity/infrastructure
- Balance the load among other underutilised networks to maximize the capacity with the existing infrastructure.

The former approach would require extra costs i.e. Capital Expenses (CAPEX) and Operational Expenses (OPEX) and while able to meet the peak demand requirements it will suffer from underutilisation most of the other times of day. However by developing new load balancing systems, the network resource utilization may be maximised with existing network infrastructure or resources pool, by moving load from heavily loaded to lightly loaded networks.

In this thesis the load balancing approach has been adapted to avoid the overloaded situation in the radio access networks. Novel load balancing algorithms have been designed and developed for the RAT selections for WLAN, WiMax, UMTS and satellite networks. For efficient resource utilization the load balancing algorithms have been implemented in the mobile node as well as in network entity such as base station (BS), Radio Network Controller and Access Point. The proposed solution for load balancing involves the utilization of IEEE 802.21 Media Independent Handover (MIH) [2] for moving load (mobile nodes) between different wireless networks. The MIH framework defines a common interface between different link layer technologies for the support of seamless mobility between heterogeneous IEEE-802 networks and between IEEE-802 and other mobile wireless technologies. This unified interface is presented as an abstraction layer function, the Media Independent Handover Function (MIHF), for handover detection, initiation and decision via Layer 2 triggers. The MIH provides the seamless mobility to mobile nodes between heterogeneous networks using a set of services known as Media Independent Command Service (MICS), Media Independent Event Service (MIES) and Media Independent Information Service (MIIS).

1.4 Contributed work and achievements

In this thesis a new load balancing framework is proposed that achieves efficient and seamless load balancing across different terrestrial wireless and satellite networks. It supports both "RAT selection triggered" based and "network triggered" handover approaches. Novel algorithms are also proposed, implemented and evaluated to find the most suitable approach for load balancing. Extensions to the MIH standard have been proposed in order

to support handovers between satellite and terrestrial networks. New primitives are introduced in the MIH for forwarding the network information like load to the MIIS in the MIH architecture. New simulation modules were implemented in the simulation framework to support load balancing algorithms, and the MIH extensions.

These above mentioned contributions helped in achieving the following goals:

- **A load balancing framework for heterogeneous wireless networks:** This thesis proposes a novel load balancing framework that is necessary to provide efficient load management across different networks. The proposed load balancing framework comprises of a Load-aware RAT selection algorithm on the mobile node, a network load balancing algorithm on the radio access network. The framework supports heterogeneous wireless networks containing satellite and terrestrial wireless networks. It utilizes and extends the MIH protocol to facilitate the load balancing process with the help of seamless vertical handovers.

- **A baseline algorithm for load balancing in heterogeneous wireless networks:** The baseline algorithm is a simple non-cognitive algorithm for balancing the load between co-located wireless networks. This algorithm takes eight parameters namely, signal strength, available resource, coverage area, speed of mobile node, cost of network, user preference, offered data rate and required data rate of user respectively. This algorithm compares different parameters and generates a list of network IDs which are suitable for the mobile node to handover. This list of network IDs is sorted based on the suitability with the most suitable

network on the top and the least suitable at the bottom of list. The most suitable networks are the least loaded networks and least suitable networks are the heavily loaded networks.

- **A fuzzy algorithm for load balancing in heterogeneous wireless networks:** This thesis proposed a fuzzy logic based intelligent algorithm to solve the load balancing problem. A Fuzzy Logic Controller (FLC) is designed and developed in MATLAB and exported to be used with the simulation framework. Eight input parameters are provided to the fuzzy logic controller. The FLC processes the input parameters using fuzzy logic operations and generates a list of networks with handover decision factors. The fuzzy based load balancing algorithm performs better than the baseline least-loaded load balancing algorithm. However, as with any fuzzy system, this system also faces the problem of "curse of dimensionality" whereby the complexity of fuzzy system increases exponentially with the increase in the number of input dimensions. The proposed fuzzy algorithm also faces similar problems as the number of input parameters are eight and each input parameter has three member functions. This phenomenon makes it very hard to tune the fuzzy membership function to provide maximum efficiency.

- **A neural-fuzzy algorithm for load balancing in heterogeneous wireless networks:** To overcome the limitations of fuzzy algorithm described above, a neural-fuzzy based load balancing system has also been proposed and developed. The Fuzzy based control system, with eight inputs each having three membership functions, consisted of 6561 set of rules. On the other hand, the neural- fuzzy system does not rely

only on such a comprehensive rule base for efficient performance. The total number of rules is greatly reduced in this neural-fuzzy system as compared to the standard fuzzy based system as it is trained with the input and output data to tune the weights in order to improve the performance. While the training of the neural-fuzzy system with such a huge number of input/output data requires a comprehensive amount of time, once the training is complete and the weights in the neural network are adjusted then there is no need to use rules. The performance of the proposed neural-fuzzy algorithm is also evaluated and compared to the baseline and standard fuzzy algorithm.

- **A simulation framework for load balancing in heterogeneous wireless networks:** The simulation model for load balancing in heterogeneous wireless networks has been developed using different tools and languages such as C/C++, TCL and MATLAB. The model is developed for NS2 and supports different wireless networks such as WLAN, WiMax, UMTS and satellite networks. The proposed model supports multi-interface mobile nodes moving across different wireless networks for load balancing purpose. Different scenarios with moving users and different networks with overlapping coverage areas can be simulated using this model in NS2. The model is fully compatible with the existing versions of NS2 therefore different utilities of NS2 like events traces and Network Animator (NAM) traces can be generated from the simulation scenarios using the load balancing model. The standard NS2 models were modified to support satellite and terrestrial networks in a

same model. A new module for load balancing algorithms was also implemented.

Following is the list of publications which are achieved during the progress of this PhD programme.

Book Chapter:

➢ K. Xu, P. Pillai, Y.F. Hu and M. Ali, "Interoperability among heterogeneous networks for Future Aeronautical Communications", Future Aeronautical Communications, INTECH publishers, Chapter accepted March 2011, book published August 2011.

Journal papers:

➢ M. Ali, P Pillai and Y.F.Hu, "Load-Aware Radio Access Selection in Future Generation Satellite-Terrestrial Wireless Networks", International Journal of Wireless & Mobile Networks (IJWMN), Feb 2012.

➢ M. Ali, P Pillai and Y.F.Hu, "Load-aware radio access selection in heterogeneous terrestrial wireless networks", International Journal of Computer Networks & Communications (IJCNC), August 2011.

➢ J.Baddoo,P. Gillick, P. Pillai, R. Morrey, A. Smith, K. Xu, M. Ali and Y Cheng, " Integration and Efficient Management of multiple Radios in Satellite-Terrestrial based Aeronautical Communication Networks", ICST Transactions on Ubiquitous Environments, 2012.

Conference papers:

➤ M. Ali, P. Pillai and Y. F. Hu, "Load aware radio access selection in future generation wireless networks", 4th International ICST Conference on Personal Satellite Services, PSATS 2012, March 2012, Bradford, UK.

➤ M. Ali, P. Pillai and Y. F. Hu, "TCP Performance evaluation over heterogeneous wireless networks using MIH", 26th International conference on CADCAM, Robotics & factories of future, Kuala Lumpur, Malaysia, July 2011.

➤ M. Ali, K. Xu, P. Pillai, Y.F.Hu, "Common RRM in satellite-terrestrial based Aeronautical communication networks", PSATS-2011, February 17-18, 2011 - Malaga, Spain.

1.5 Report organization

This thesis is divided into 6 chapters. Following this introduction chapter, chapter 2 presents an overview of different wireless networks, their protocol stack, network architecture and issues and approaches for seamless integration between these different wireless networks for load balancing. At the end of chapter 2, IEEE 802.21 media independent handover is briefly described along with the primitives mapping of different wireless networks.

A literature review of RAT selection techniques and load balancing approaches in heterogeneous wireless networks is presented in Chapter 3. The advantages and limitations of the various RAT selection techniques are discussed and their role in designing an efficient load balancing framework is explained. The chapter also compares the various existing load balancing

approaches and highlights their weaknesses which provides the motivation of this research work.

Chapter 4 presents the detailed description of the proposed load balancing framework for heterogeneous wireless networks. This chapter first explains the target network architecture, protocol stack and the general design for the load balancing framework. It explains the extensions required in the IEEE 802.21 standard for supporting load balancing between heterogeneous satellite-terrestrial wireless networks. The network initiated and mobile node initiated vertical handover procedures required for load balancing across different networks are explained in details with the help of message sequence charts. Finally this chapter presents the three proposed load balancing algorithms i.e. baseline, fuzzy and neural-fuzzy. These algorithms are explained in details with the help of examples and predefined input parameters. The structural components of all three algorithms are presented and their advantages and limitations have been discussed.

The simulation model created in the NS2 for load balancing is described in Chapter 5. This chapter first describes the process of load balancing model development and then goes on to explain the different simulation scenarios. This chapter also presents the results of each scenario which have been analysed in detail. Different performance parameters have been monitored to show the advantages of load balancing in the heterogeneous wireless networks. Finally Chapter 6 presents the conclusion of this thesis which summarise the contributions made by this research and also presents some recommendations for additional research for future development.

Chapter 2: Heterogeneous wireless networks

2.1 Scope

This chapter presents the brief description of the different wireless communication technologies that are considered in this thesis. These are:

- IEEE 802.11 – commonly known as Wireless Local Area Network (WLAN)

- IEEE 802.16 – commonly known as Worldwide Interoperability for Microwave Access (WiMax),

- Universal Mobile Telecommunication System (UMTS), and

- Satellite networks

An efficient load balancing framework would be able to seamlessly move user connections from one technology to another and achieve a more uniform balance across different networks. In order to achieve such a seamless integration amongst the heterogeneous networks it is important that the framework would consider their respective characteristics like coverage areas, costs, data rates, etc. and also look at various techniques for network integration. This chapter first looks into the network architecture of the various target wireless access technologies and then describes how the IEEE 802.21 Media Independent Handover mechanism adopted in this thesis may be used for handovers across different technologies.

2.2 Wireless Networks Overview

Wireless communication technologies are attributed as a platform to establish or extend network communications in a mobile, portable and cost effective way. It provides the capability to connect users living in sparsely populated or/and remote areas where connectivity via existing fixed technologies may not be cost effective and reliable. Some of the commonly used wireless technologies are Global System for Mobile Communication (GSM) [3], General Packet Radio Service (GPRS) [4], WLAN IEEE 802.11 [5, 6, 7, 8], WiMax IEEE802.16 [9, 10] and 3GPP's UMTS [11, 12].
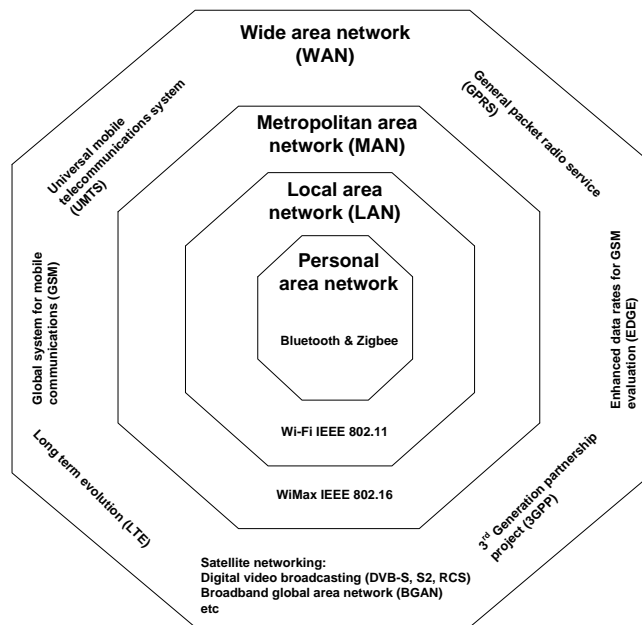
Figure 2-1: Classification of wireless networks

As shown in Figure 2-1, wireless communication systems are classified into four different types according to their range: Wireless Personal Area Network (WPAN), Wireless Local Area Network (WLAN), Wireless Metropolitan Area Network (WMAN) and Wireless Wide Area Network (WWAN). Bluetooth, Zigbee and infrared technologies are from the Personal Area Network (PAN) technologies which constituent short range. The medium range networks

technologies can be further subdivided into lower medium and higher medium range. WLAN belongs to lower medium range and WiMax belongs to higher medium range technologies. 3G UMTS and GSM are both wide area network technologies which are of large range. Finally the satellite networks with largest coverage range which can provide global coverage. The UMTS, WiMax, WLAN and satellite networks harbour the promise of fully distributed mobile communication, anytime, anywhere.

The advancement in wireless communications networks is bringing fundamental changes to telecommunication networking and is making hybrid networks a reality. Different approaches for hybrid wireless networks have been presented over the last decade. The idea of developing advanced wireless communication systems and their interworking is to provide a user with various services at low cost and enhance Quality of Service (QoS), anywhere anytime, with optimum utilization of available radio resources. There are a number of different wireless access technologies in existence; each of these technologies has specific advantages and disadvantages. The hybrid wireless network provides a way of putting together the advantages of all these networks and presents a very flexible wireless network system. Before going into the details of integration of wireless networks the following section gives a brief overview for some of the most commonly used wireless networks and their characteristics.

2.2.1 Universal Mobile Telecommunications System

Universal Mobile Telecommunications Systems (UMTS) is a third generation mobile cellular technology [13]. It offers voice service, Short Messaging

Services (SMS) and IP based packet data services. Both circuit-switched and packet-switched services are offered for point-to-point and point-to-multipoint communications. It supports high bit rate of up, 384 Kbps and 2 Mbps for circuit-switched and packet-switched data communications. The offered services have varied QoS parameters for maximum transfer delay, delay variation and bit error rate. There are four different types of QoS classes for UMTS network services which are:

- o **Conversational class:** This class involves applications like voice, video gaming, and video telephony.

- o **Streaming class:** This class consists of applications like video on demand, multimedia and webcast.

- o **Interactive class**: comprises of application such as network gaming, web browsing and database access.

- o **Background class:** this class is composed of applications like SMS, Email and downloading.

2.2.1.1 UMTS Network Architecture

UMTS is composed of three interacting architectural components namely Core Network (CN), Radio Network Subsystem (RNS) and User Equipment (UE) [11]. Figure 2-2 represents the block diagram of the UMTS architecture and its major components. The CN provides routing, switching and transit function as for the user traffic. It also contains databases and network management functions. The basic CN architecture of UMTS is based on that of the GSM /General Packet Radio Service (GPRS) network. The UMTS

terrestrial radio access network (UTRAN) provides the air interface access method for the UE. The Node-B is the base station which is in turn controlled by the Radio Network Controller (RNC). These different network entities are described in detail in the following sub-sections. The Core Network (CN) is composed of circuit-switched and packet-switched domains. Mobile services Switching Centre (MSC), Visitor Location Register (VLR) and Gateway MSC are circuit switched elements. Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN) both are packet switched elements while EIR, HLR, VLR and AUC are shared by both domains. Asynchronous Transfer Mode (ATM) is defined for UMTS core transmission. Circuit switched connection is handled by ATM Adaptation Layer type 2 (AAL2) and packet connection protocol AAL5 is designed for data delivery.
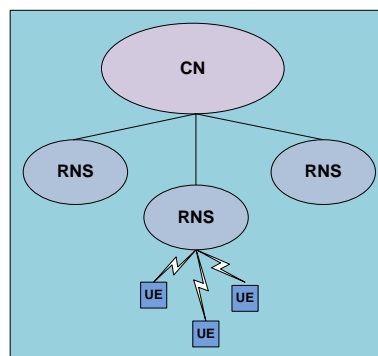
Figure 2-2: UMTS network architecture

2.2.1.2 Radio Network Subsystem

In UMTS the role of RNS is similar to the base station Subsystem (BSS) of GSM. It manages the air interface for the whole network.
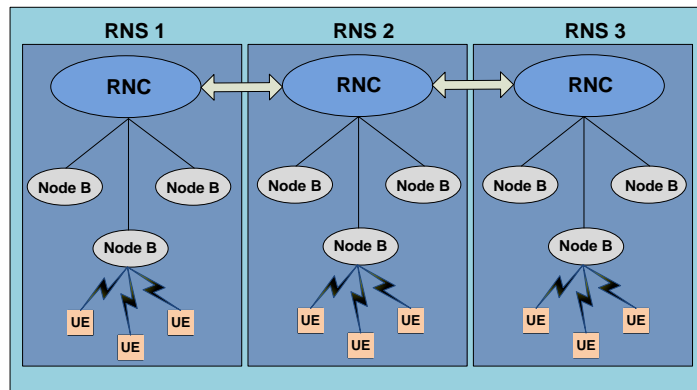
Figure 2-3: Radio network subsystems

The RNS is also known as the UTRAN UMTS Radio Access Network. Figure 2-3 shows the major components of the RNS. The main components of RNS are the Radio Network Controller (RNC) and the Node B.

These are briefly described below:

Radio Network Controller:

The RNC controls Node B's which are connected locally. One RNC can have one or more node B's connected to it. The RNC is responsible for the radio resource management operations and mobility management operations. The data encryption and decryption is also performed at the RNC. The RNC facilitates handover operations of the UEs between different node B's and also interacts with neighbouring RNCs to perform handovers of UEs to the node B's of other RNCs.

Node B:

The base station transceiver in the UMTS is known as the node B. It is composed of a transmitter and a receiver to establish communication with UEs which are located within the range of communication cell.

User Equipment:

The User Equipment (UE) is the mobile terminal used by the end user to access the various services like voice communication, web browsing, video streaming, radio listening, etc.

2.2.2 IEEE 802.16

The IEEE 802.16 family of standards is often referred to as World Wide Interoperability for Microwave Access (WiMax). It is an effective metropolitan area access technology with many encouraging features such as high speed, cost efficiency and flexibility [14]. The coverage area of WiMax spans 30 to 50 km. Data rates of more than 100 Mbps in a 20MHz channels are offered [15]. The Wireless MAN air interface which is the WiMax standard was first approved by IEEE 802.16-2001 standard [10] in 2002.

The 802.16d standard employs three kinds of physical layer technologies, which are: Single Carrier (SC) applied in the frequency range of 10-66 GHz, OFDM 256 points in frequency range of 2-11 GHz fixed wireless access, and Orthogonal Frequency Division Multiple Access (OFDMA) 2048 points with frequency up to 11 GHz for long distance between operator point of presence and Wireless Local Area Network [16]. Mobile WiMax is based on the IEEE 802.16e standard and operates in the spectrum bands of 2.3 GHz, 2.5 GHz 3.3 GHz. The main advantages of WiMax, as compared to other MAN access network technologies is the more sophisticated QoS support. WiMax can interwork with, satellite and terrestrial wireless networks. It also serves to backbone for WLAN hotspots for connecting to the broadband internet services. It offers broadband connections which support multiple scenarios,

including fixed, portable and mobile wireless access and cover range of up to 40 km for Line of Sight (LOS) and up to 10 km for Non Line of Sight (NLOS) operations. WiMax network architecture is more flexible, encourages interworking and roaming and is cost effective as compared to the other MAN access technologies.

Reference [16] describes history of WiMax standards and its advantages such as MAC of IEEE 802.16, which supports different transport technologies including IPv4, IPv6, Asynchronous Transfer Mode (ATM) and Ethernet. WiMax 802.16e standard supports power saving and sleep modes to extend the battery life of mobile devices, hard and soft handoffs are also supported to provide seamless connections to users. A study in reference [14] presents the estimation for WiMax MAC header overhead to reserve sufficient amount of slots for the constant-rate applications. The study presents several simulation scenarios to demonstrate how the scheduling solution allocates resources in various cases. The solution was based on round-robin scheduling. It would have been more significant if the study would have provided comparative results against other wireless technologies. Reference [17] proposed a multi-channel Carrier Sense Multiple Access (CSMA) with collision avoidance. This paper evaluates the performance of receiver based channel selection, comparing with IEEE 802.11 Distributed Coordination Function (DCF) using ns-2 simulator. However, at a given time, only one packet can be transmitted on any channel, but multiple packets can be received at various channels at the same time. A study in reference [18] uses Forward Error Correction (FEC) and Automatic Repeat Request (ARQ) to support streaming services and investigated the problem of real time

streaming media over WiMax. It exploits the flexible features in the MAC layer within the 802.16a standard. The authors proposed the size of MAC packet data units to make adaptive to the instantaneous wireless channel condition. Reference [19] propose an integrated Adaptive Power Allocation (APA) and Call Admission Control (CAC) downlink resource management framework for OFDM-TDD based multiservice network by taking into account the service provider and subscriber.

Reference [20] proposed a pricing model for adaptive bandwidth sharing in an integrated WLAN/WiMax network. Game theory has been used to analyse and obtain pricing for bandwidth sharing between a WiMax base station and WLAN access point routers. Reference [21] discussed the interference issues and proposed an efficient approach for utilization of WiMax mesh through design of a multi-hop routing and scheduling algorithm scheme. This scheme considered both traffic load demand and interference conditions. The simulation results showed that the proposed schemes had effectively improved network throughput performance in IEEE 802.16 mesh networks and high spectral utilization is received. Another investigation in reference [22] is based on Wireless OFDM networks which relates to the FIREWORKS project. The study analyses the characteristics to improve IEEE 806.16 standards. Designing of Radio Resource Management (RRM) algorithms is also recommended in this study. An investigation in reference [23] compares delay performance of two bandwidth request mechanisms, detailed in 802.16 standards, random-access and polling. It has been drawn that the polling mode provides better QoS performance than random access mesh mode. Generic information regarding WiMax frequency and range is provided in

most of the other studies. Some are narrative description of the WiMax's standards and MAC layers. However, a study in reference [23] compares interference issues and testing results on data rate and delay. Therefore, it is important to study interoperability issues and Quality of Service requirements for WiMax in hybrid Wireless networks.

2.2.2.1 WiMax Network Architecture

Figure 2-4 shows the simple representation of an IP based WiMax network architecture. The WiMax network can be divided into three segments namely Mobile Station (MS), Access Service Network (ASN) and Connectivity Service Network (CSN). The MS is used by the end user to gain the network access; ASN consists of one or more base stations (BS) which are connected to one or more ASN gateways which connect the Radio Access Network (ASN) to the CSN. The IP connectivity and other IP core network functions are provided by the CSN.
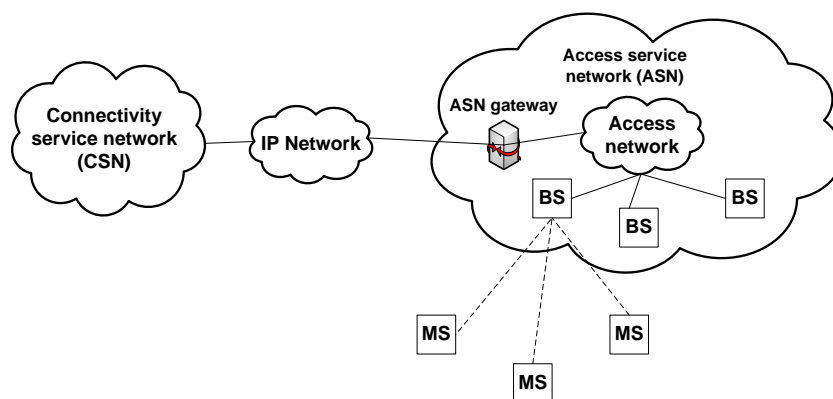
Figure 2-4: WiMax network architecture

These segments are described below:

The Base station provides the air interface to the mobile station and may also provide the micro mobility management functions like DHCP proxy,

traffic classification, multicast group management, session management, key management, QoS policy enforcement, Radio Resource Management (RRM), establishing tunnel and triggering handover.

The Access service network gateway acts as a layer 2 traffic accumulation point in the ASN. The operational responsibilities of ASN gateways include are QoS and policy enforcement, routing to the selected CSN, foreign agent functionalities for mobile IP, establish and management of mobility tunnel with BS, AAA client functionality, caching of subscriber profiles and encryption keys, location management, radio resource management and call admission control.

The Connectivity service network is responsible for providing the connectivity to the internet and other public and corporate networks. It also includes the Authentication, Authorization and Accounting (AAA) servers that provide the authentication for user devices and other specific services. User policy management of QoS, security, IP addressing management and support for roaming between network service providers and location management are also included in the responsibilities of CSN [24].

2.2.3 IEEE 802.11

The IEEE 802.11 standard family provides the specifications for short-range Wireless Local Area Network (WLAN) connectivity. The standard family consists of specifications like IEEE 802.11a, IEEE 802.11b and IEEE 802.11g and IEEE 802.11n, etc. The latest IEEE 802.11n is more reliable, secure and faster than the older standards. The coverage area for an 802.11 based WLAN is around 100-150m. WLAN hotspots are widely used to

provide internet access in restaurants, hotels, offices, airports and school campuses, etc. This is due to ease of availability of the equipment and ease of use, its low maintenance and servicing cost.

2.2.3.1 Evolution

The first set of specifications released for WLAN was operating at 2.4GHz. IEEE 802.11 group created several task forces which includes a, b, g, f, e, h, i, n. IEEE 802.11b was the most common and popular WLAN standard; it uses frequency range of 2.4 GHz-2.4835 GHz. Maximum data rates of 1, 2, 5.5 and 11 Mbps are supported by this standard using Direct Sequence Spread Spectrum (DSSS) [25, 26]. IEEE 802.11a is a high speed WLAN offering speeds up to 54 Mbps [27] in the 5 GHz band [28]. It uses modulation technique known as OFDM which reduces the multipath interferences. The IEEE 802.11g standard is an extension of the 802.11b standard operating in 2.4 GHz band [29]. It supports up to 54 Mbps due to the combination of OFDM and Complementary Code Keying (CCK). OFDM advantages include increased spectral efficiency and multipath effects.

2.2.3.2 WLAN Network architecture

The basic network architecture for the WLAN is shown in Figure 2-5. It is composed of two basic components i.e. the Access Point (AP) and the wireless clients. The AP is connected to the internet using wired link and it provides internet service to the wireless clients connected to it.
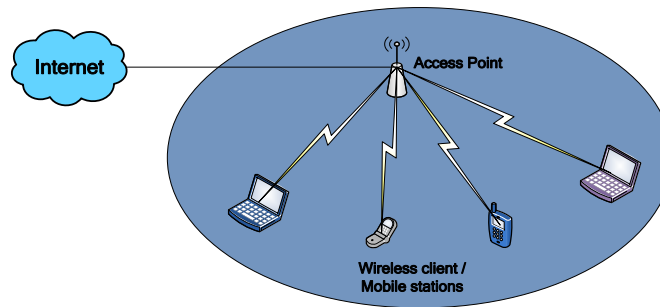
Figure 2-5: Basic WLAN network architecture

2.2.4 Satellite Networks

Satellite networks have a growing importance due to their vast geographic coverage, fast deployment and intrinsic multicast/broadcast abilities as compared to terrestrial networks. Satellite systems can also be used to provide broadband and multimedia services to the end-users [30]. Satellites are multiple access systems with limited transmission capacity compared to terrestrial networks. Therefore it is challenging to develop an efficient resource management technique that delivers acceptable QoS to users while ensuring the provision of adequate efficiency. Satellites can be classified according to their processing capability as bent pipe or non-regenerative satellites and On-Board Processing (OBP) or regenerative satellites. Bent pipe satellites are physical layer devices which are simply signal repeaters in the sky. The signals received on the satellite uplink are amplified and broadcast at a different frequency on the downlink. However the advanced OBP satellites may accommodate baseband digital processing, uplink bandwidth controller and fast packet switching on board. OBP satellites are link layer devices and they form a mesh network topology rather than a star topology as in the bent pope satellite networks.

2.2.4.1 Classification of satellites

Satellites can also be categorized according to their orbits. The general categories are Low Earth Orbit (LEO), Geostationary Earth Orbit (GEO), Medium Earth Orbit (MEO) and High Elliptical Orbit (HEO) as shown in Figure 2-6 below.
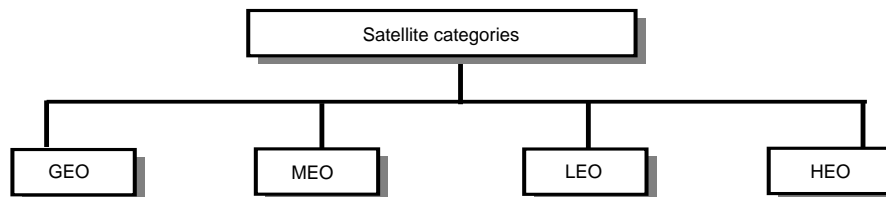
Figure 2-6: General categories of satellite orbits

2.2.4.1.1 GEO

One of the most common types of satellite in use today is the GEO satellite or geostationary satellite. The Geostationary satellite was first proposed in the year 1945 by Arthur C. Clarke; a science fiction author. The reason these satellites are called geostationary satellite is because they follow a geostationary orbit which is approximately 35,863 km above the earth's surface. The speed at which these satellites move is exactly the same as the speed of the earth so that the geostationary satellites appear to be always at the same spot above the earth; this is the reason why they are called geostationary. Figure 2-7 shows the distance of a geostationary satellite from the earth surface.
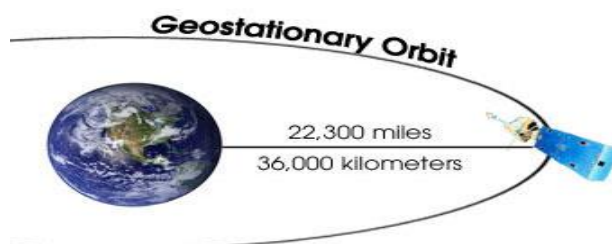
Figure 2-7: Geostationary Orbit

One of the main functions of a geostationary satellite is to provide constant communication while covering almost one third of the Earth's surface. To ensure LOS (Line of Sight) propagation; it is obvious that the receiving and the sending antennas must be fixed in relation with each other's location. For this reason, all geostationary satellites in a global beam; three in number (as shown in Figure 2-8) , are locked with respect to each other's location and they are located 35,863 km above the earth, covering almost 99% of the population.

Figure 2-8: GEO Satellites coverage areas

GEO satellites provide fixed services like the FSS or Fixed Satellite Service and mostly use the Ku and the C band of the radio spectrum. Some of the main reasons or advantages of using a geostationary satellite include:

- Earth stations can easily track the satellite.

- Three GEO satellites, 35,863 above the earth surface, can cover almost all inhabited areas on earth, with exception of few areas close to the poles.

- GEO satellites are stationary relative to the antennas; hence they are not prone to problems related with frequency changes.

- Suitable for providing communication access to remote areas

- Can be used for security and various environmental monitoring purposes.

With all its advantages, however, geostationary satellites also bring some problems or limitations like:

- As the distance of a GEO satellite is approximately 35,863km from the earth; it makes the received signal weak as it travels that much distance and also suffers from high propagation delay.

- Geostationary satellites don't give adequate coverage to the areas near the south and north poles.

- Due to the broadcasting nature of satellites, the data sent through a GEO satellite is publicly available.

- While GEO satellites incur low maintenance cost, they do incur very high hardware and deployment costs.

In this research, non-regenerative satellite systems are considered. Some of the most commonly used satellite systems have been considered such as different flavours of Digital Video Broadcast (DVB) over satellite like DVB-S [31,32], DVB-S2 [33, 34], DVB-SH [35,36], DVB-RCS [37,38] and Broadband Global Area Network (BGAN) [30].

2.2.4.1.2 MEO

MEO or Medium Earth Orbit satellites are positioned approximately 8,000 to 18,000 km above the earth's surface. MEO satellites are located between the two VAN Allen Belts and takes 6-8 hours in order to complete one orbit of the earth. While some MEO satellites have almost ideal orbits, hence maintaining a fixed altitude from the earth and they travel at constant speed, other MEO

satellites follow elongated orbits. MEO satellites are closer to Earth in comparison with GEO satellites, therefore the transmitters located at Earth needs less power to communicate with these satellites. On the other hand, a MEO satellite is higher in altitude as compared to a LEO satellite; therefore it has a greater footprint (the area being covered on the earth's surface). Figure 2-9 shows the MEO satellites in their orbits.
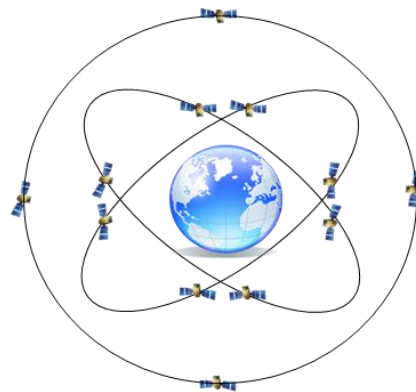
Figure 2-9: Medium Earth Orbits

One of the advantages of using a MEO satellite over a GEO one is that the overall round-trip signal propagation delay time in case of MEO satellite is approximately 50ms which is quite less as compared to GEO satellite (250ms). However, in order to cover the entire Earth, more MEO satellites are required as compared to GEO satellites.

2.2.4.1.3 LEO

Low Earth Orbit satellites follow a polar orbit having an altitude ranging between 500 and 1500 km. LEO satellites have got very high orbital velocity (20,000 to 25,000 km/h). Figure 2-10 shows the satellites in the LEO orbit.
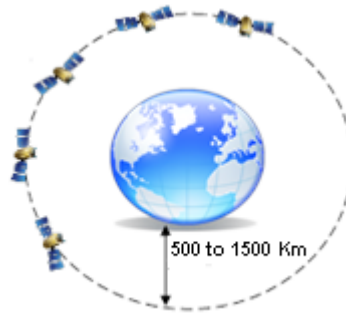
Figure 2-10: LEO Satellite System

LEO satellites are relatively small satellites and they are quite easy to launch, repair and modify. One of the advantages of using a LEO satellite is that additional satellite instruments can easily be attached to it since it is not that far from earth and can be easily reached using a space shuttle. LEO satellites are generally used as a network of satellites capable of handling e-mail and broadcasting data at greater speed as compared to GEO or MEO satellites. Since they are quite close to the earth, they provide higher data transfer rates.

In brief, LEO satellites have the following properties:

- LEO satellites have altitudes ranging between 500 and 1500 km.

- They follow a slightly elliptical or circular orbit.

- The diameter of the coverage provided by a LEO satellite is approximately 8000km.

- In order to complete an orbit, LEO satellite takes approximately 2 hours.

- A LEO satellite is prone to orbital deterioration because of the atmospheric drag which is the result of being too close to the earth's atmosphere.

- A LEO satellite has the least round-trip delay of 20ms as compared to MEO (50ms) and GEO satellite (250ms).

- A LEO satellite remains visible from a point on earth for only 20 minutes.

A typical LEO system consists of multiple satellites, referred to as constellation of satellites, that work together to form a satellite network where each satellite can be considered as a switch. The satellites that are located quite close to one another are connected through special links called as the Inter-satellite Links or ISLs. The satellite communicates with the user on earth using a User Mobile Link or UML and it communicates with the earth station using a Gateway Link or GWL.

2.2.4.1.4 HEO

The High Elliptical Orbit (HEO) satellites follow an elliptic orbit around the earth. The approximate distance of HEO satellite from the earth ranges from 18,000 km to 35,000 km. The main objective behind the development of HEO satellites is to cover countries where large population is situated in high southern or northern latitudes. Today, various systems are in use that provide an arrangement for the apogee that provides non-stop coverage to a particular area with higher latitude. An apogee is "the highest altitude point of the orbit". This is the point where the satellite reaches its farthest point from the earth. HEO satellites follow a special orbital plane having inclination ranging between 50 degree and 7 degree. In order to complete one orbital period, HEO satellite takes approximately 12 hours.
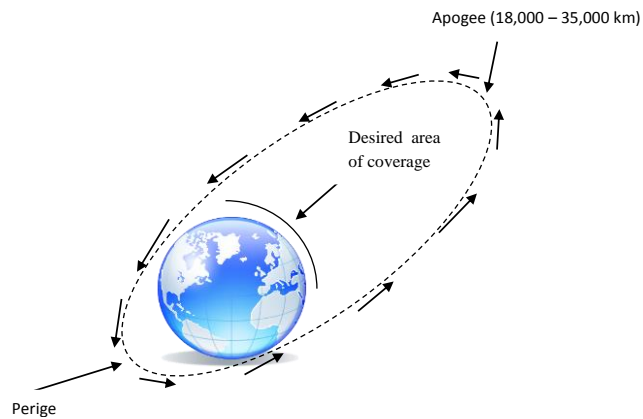
Figure 2-11: A Highly Elliptical Orbit

During its orbit, there is a point where the satellite comes very close to the Earth and hence its velocity increase. After then, it goes far from the earth with a decreased orbital velocity. If an orbit has a very elliptical shape; then the satellite spends most of its time at the Apogee. Since the orbital speed at high altitude is low, a HEO satellite spends the bulk of its time in the higher altitude (at apogee). This helps in giving the desired area maximum coverage for a longer period of time. As you can see in Figure 2-11 of a HEO; at apogee, the satellite has a longer dwell time (the time a satellite remains over a particular part of the earth) as it is at high altitude and hence it gives maximum coverage to the desired area. However, when it is at the point in the orbit where the orbital velocity increases (closer to the perigee) then no coverage can be given to the desired area. In order to sort out this problem, two HEO satellites are used that are exactly located opposite to each other, so when one HEO satellite is at apogee, the other one is at perigee and when the first one comes at perigee, the desired area is still covered by the second HEO satellite which is now at the apogee. In this way, there is always one HEO satellite giving coverage to the desired area.

2.2.4.2 Target Satellite network (BGAN)

The satellite network considered in this research is the geostationary based satellite network called Broadband Global Aeronautical Network (BGAN).The development of fourth generation of satellites for the BGAN project was initiated in December 1999 by INMARSAT. The INMARSAT-4 satellites comprise three GEO satellites to provide the communication services to the mobile terminal. The first two GEO satellites are located over the Indian Ocean Region (64ºE) and Atlantic Ocean Region (53ºW) respectively providing coverage to the target areas. INMARSAT launched a third satellite in 2008 to improve the coverage. This third satellite is positioned in geostationary orbit at 98ºW. Each INMARSAT-4 satellite weighs 3 tons and supports approximately 200 spot beams. It provides transparent amplification for the BGAN communications (user plane and control plane). Transmission between the RNC and satellite is via the C band, whereas transmission between the satellite and MTs is via the L band [30].

The INMARSAT BGAN is intended to form part of the satellite component of the Third Generation (3G) IMT-2000/Universal Mobile Telecommunications System. Among the design objectives is the interoperability with an industry standard 3G Core Network and the re-use of the UMTS Non Access Stratum (NAS) layers. The BGAN system adopted the same UMTS architecture in the core network but uses an INMARSAT proprietary air interface - INMARSAT Air Interface 2 (IAI-2), which provides a complete Access Stratum Protocol Stack and Physical Layer optimised for the geo-stationary satellite environment. The air interface is based on TDM and TDMA/FDM schemes in forward direction and return direction respectively [30].

BGAN is the first system to provide guaranteed data rates on demand. It is also the first satellite communication system to provide both voice and broadband mobile communication services on a global area, where three GEO satellites of BGAN system are covering almost every part of the earth's surface. The system network architecture has the capability to provide UMTS compatible services and both circuit switched and packet switched services. The BGAN MT is a light weight portable satellite terminal, which is easy to carry, simple to setup for use, can deliver data rates of up to half a megabit. In order to achieve high transmission efficiency and flexibility, it is possible to adapt the bandwidth, coding rate according to the MTs class and channel conditions. In the BGAN baseline system, 3 classes of MTs are supported with different maximum transmission rates of 492 Kbps, 432 Kbps, and 216 Kbps respectively when receiving data and maximum transmission rates of 492 Kbps, 144 Kbps and 72 Kbps respectively when transmitting [30, 39].

2.2.5 Comparison of wireless networks

This section provides a comparison between the different wireless access technologies which may form part of future heterogeneous wireless networks. Major candidates of the heterogeneous wireless access networks are WLAN, UMTS, WiMax and satellite networks. Table 2-1 represents the generic comparison of the wireless access technologies, such as their standards, bandwidth, frequencies, and modulation techniques and data rates.

Table 2- 1: Comparison of Wireless Access Technologies

Access Technology	Standards	Modulation	Data rate	Bandwidth	Frequency
WLAN	802.11	BPSK, QPSK, 16QAM, 64QAM, DSSS, MIMO	54 to 150 Mbps	20 MHz	2.412 to 2.484, 5.15 to 5.25 GHz
WiMax	802.16	QPSK, 16QAM, 64QAM	75 Mbps (Max.)	5 to 20 MHz	2 to 66 GHz
UMTS	3G	W-CDMA /OFDM/OF DMA	2 Mbps	1 to 2 MHz	850/1900/21 00 MHz
BGAN	S-UMTS	MFTDMA 16QAM	492 Kbps	200 KHz	L band, 1.5 to 1.6 GHz

2.3 Interworking of heterogeneous wireless networks

The coexistence of diverse but complementary architectures and wireless access technologies is a major trend in heterogeneous wireless networks. An appropriate integration and interworking of existing wireless systems are vital in this context. 3GPP and 3GPP2 both have proposed interworking architectures for 3G cellular networks and wireless local area networks (WLAN). However, the proposed interworking architectures are delayed due to some drawbacks; the most significant being the seamless roaming and absence of guaranteed quality of service (QoS).

In modern days the coverage areas of different wireless networks overlap or coexist and this can be utilized in numerous ways to provide users anytime anywhere connectivity to mobile users, by providing seamless mobility, resource sharing or load balancing between heterogeneous wireless networks. An example of the target topology is shown in Figure 2-12 where coverage areas of WLAN, WiMax, UMTS, and satellite networks are overlapping each other.
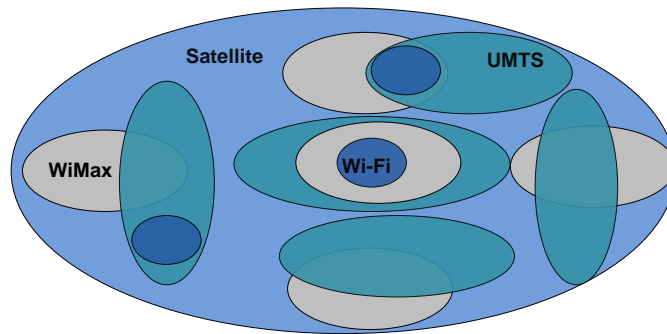
Figure 2-12: Heterogeneous wireless networks coverage areas coexisting

In the future heterogeneous wireless networks it is expected to exhibit heterogeneity in terms of wireless access technologies and services. The advantage of 3G cellular networks like UMTS is the global coverage while their weaknesses lie in their operational cost and bandwidth capacity. While on the other hand, WLAN technologies like 802.11 offers higher bandwidth and low operational cost but covers relatively short range. However, the evolution of portable devices has made it possible to support different Radio Access Technologies (RATs) on a single terminal. Hence, instead of putting efforts into developing new radio interfaces and technologies for the future needs of users, integration of these systems provide an alternative and beneficial option. This paves the path for future heterogeneous wireless networks. The integration of different networks will unify the advantages of these systems and will minimise the disadvantages, allowing a great market opportunity.

The heterogeneity in terms of network protocols and RATs in heterogeneous wireless access networks demands for common interconnection element. Since Internet Protocol (IP) technology enables the support of applications in a scalable and cost-effective way; it is expected to become the core backbone of future heterogeneous wireless access networks [40]. Hence,

current trends in communication networks evolution, in order to hide heterogeneities and to achieve convergence of different access networks are directed towards an all IP paradigm.

The integration of WLAN and 3G cellular networks may be done in phases. Loose and tight coupling are the two major architectures for 3G/WLAN interworking that have been proposed by both 3G wireless initiatives i.e. 3GPP and 3GPP2, for their respective system [40, 41]. However, the new integration brings new challenges such as security issues, QoS guarantees, interworking, mobility management and integration point. These issues are key challenges in order to support global roaming and service continuity of mobile nodes (MNs) across various networks in an efficient way. A study in reference [41] provides a generic overview of possible techniques and levels of the interworking for the heterogeneous wireless networks. It also included the interworking of heterogeneous networks using IEEE 802.21 media independent handover technique [2].

2.3.1 Basic constraints for integration

An integrated and interworking architecture for future generation heterogeneous wireless access network should address specific requirements and possess the following characteristics:

o **Economical:** The architecture should minimize the use of new infrastructures and use the existing infrastructures as much as possible.

o **Scalable and Reliable:** The integrated systems should be supported by the target network architecture and provide fault tolerance.

o **Seamless Mobility:** During inter-system or intra-system roaming/handovers, the architecture should support seamless mobility to eliminate connection interruptions and QoS degradation.

o **Security:** The level of security and privacy provided by this architecture should be equivalent or better than the existing wired and wireless networks.

o **Billing:** The cross network billing can be decided with the help of Service Level Agreement (SLA) in case of multiple operators. In case of the same operator providing services over multiple access technologies the subscribers can be informed about the different charges for different types of networks in the terms and conditions.

The forecast that interworking hybrid architecture will dominate the market is difficult because the selection of model is not primarily based only on performance, but on its cost and profitability. Hence the new hybrid architecture can be achieved by a certain trade-off of the above mentioned constraints.

2.3.2 Needs for Interworking of wireless networks

The increasing drive towards the convergence is result of various developments in the past decade. At the beginning there has been growth and evolution of several access technologies that are diverse in the nature of supported services, coverage ranges and provisioned data rates. Fixed line technologies e.g. LANs, Cable modems, DSL etc. have continued to provide internet access in a variety of settings which includes residential users as well as commercial users. Whereas cellular technologies traditionally catered

for voice services, and have also been providing data, multimedia services, and access to internet via existing and evolving standards and technologies. While WLANs have become a popular means of Internet access with high-speed but with coverage over small ranges. Whereas emerging wireless metropolitan area access technologies such as WiMax facilitates last mile broadband wireless access. These heterogeneous wireless access technologies have produced a roam or need for a framework to be developed where these existing wireless access technologies can be integrated and resources in the wireless networks can be utilized efficiently.

A major issue in interworking architecture is that different applications have varying bandwidth requirements and different access networks provide different quality of access and bit rates at different sites, therefore different networks may be of best utility at different times. As a result multimode or multi-interface devices have been introduced. Smart phones are gaining popularity because of providing connectivity with cellular and WLAN. Smart phones have a built wireless card. The need of seamless operations or sessions in mobile wireless devices, across heterogeneous wireless networks demands for the emergence of such multi-interface devices.

The internet infrastructures and cellular networks are expanding to deliver multimedia, voice and data services. However, data and voice services have been provisioned by cellular networks and internet infrastructure respectively. Data and multimedia service provisioning is an integral part of cellular services due to the need for internet access on the move and services like messaging and multimedia. However, on the other hand due to the low tariffs over the cellular counterpart with increased reliability, multimedia and voice

services over the internet in cellular system has become popular. The advancement of these two segments of infrastructures addressing the same needs for the users permits interoperability and convergence of user devices and services. For carriers with mobile and fixed-line users, convergence entails allowing their users to utilize multimedia, data and voice services seamlessly on multiple devices in a way that the user quality of experience is enhanced, while reducing the threat posed by small operators and carriers offering data and voice services at reduced prices. The drive towards convergence of heterogeneous wireless networks has been gaining momentum due to these factors and due to the fact that heterogeneous access technologies will continue to co-exist [2, 42, 43, 44]. Multimedia conferencing, multimedia services, IP and peer-to-peer TV, HDV availability at homes, constitute only a part of the whole scope of services driving the trend towards convergence.

2.3.3 Interworking of Terrestrial and Satellite networks

The phenomenon of future telecommunication service provisioning is moving towards unified service architecture and global ubiquitous networking. The global ubiquitous networking is not possible without efficient interworking between different access network technologies. This trend demands for defining, implementing and deploying common services control architecture, capable of supporting wide variety of services for users. To bridge the communication between the densely populated urban areas and sparsely populated remote areas the ideal candidate is the technology of satellite communications as it can be used as an alternate in areas where there is no terrestrial alternative. The interworking of satellite networks with existing

terrestrial networks, whether they are fixed or mobile wireless networks can be exploited in a wide variety of ways. Most important of which is the efficient resource utilization as the satellite resources are scarce and expensive whereas terrestrial networks resources are comparatively inexpensive and a better alternative. Therefore, mobile users using satellite networks can be moved to the terrestrial wireless networks when they move to the terrestrial coverage areas.

The satellite networks, performing in isolation, cannot compete with terrestrial systems in urban areas as fixed and mobile technologies e.g. ADSL, GSM, UMTS/3G, etc. are well advanced in urban/sub-urban areas. The main market for satellite networks is the areas where these terrestrial technologies are inaccessible. These areas are small and bring poor revenue for satellite operators. The future for next generation satellite networks is in an integrated architecture with terrestrial networks. The success also depends on the ability to provide, in full compatibility with terrestrial systems, broadband data rate applications, as in today's internet. On the other hand it is a good trade-off for terrestrial networks as it will provide them with the opportunity to increase the capacity of their systems, satisfy the ever growing community of internet users and support large-scale deployment of different emerging bandwidth-intensive services. There are two critical issues that arise when considering interworking of satellite systems with terrestrial systems. First, there are challenges in integrating satellite and terrestrial networks, mainly when terminal mobility is essential and secondly satellite systems are costly in general [41].

2.3.4 Interworking of Target Networks

Figure 2-13 represents the target network architecture for the interworking of wireless networks like WLAN, WiMax, UMTS and satellite networks. In the architecture, it is assumed that the mobile node supports multiple interfaces and can use any network when available. The target network model assumes that either a single operator controls all the access networks (for example, Vodafone provides 3G, WiMax and WLAN services) or Service Level Agreements (SLA) are in places when the networks and operated by different service providers. This is important for a mobile node roaming into other networks to have seamless mobility across different networks.
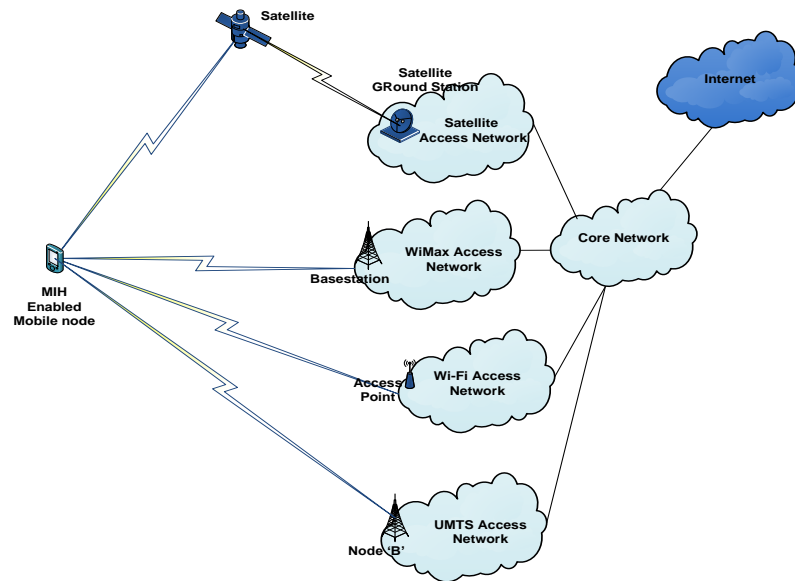
Figure 2-13: Target network architecture

The IEEE 802.21 standard for Media Independent Handovers (MIH) has been adopted in this research for providing a unified framework for interworking of the heterogeneous networks. The MIH requires additions on the layer 2 and between layer 2 and layer 3 of each access technology to support seamless mobility. The following section briefly describes the IEEE

802.21 media independent handover which has been employed to achieve the interworking between target wireless network technologies.

2.3.4.1 IEEE 802.21 Media Independent Handover

The IEEE 802.21 Media Independent Handover (MIH) framework [2] defines a unified interface between different link layer technologies for the support of seamless mobility between heterogeneous IEEE 802 networks and between IEEE 802 and other mobile wireless technologies. This unified interface is presented as an abstraction layer function, the Media Independent Handover Function (MIHF), for handover detection, initiation and decision via Layer 2 triggers. Figure 2-14 shows the IEEE802.21 MIHF reference model and SAPs.
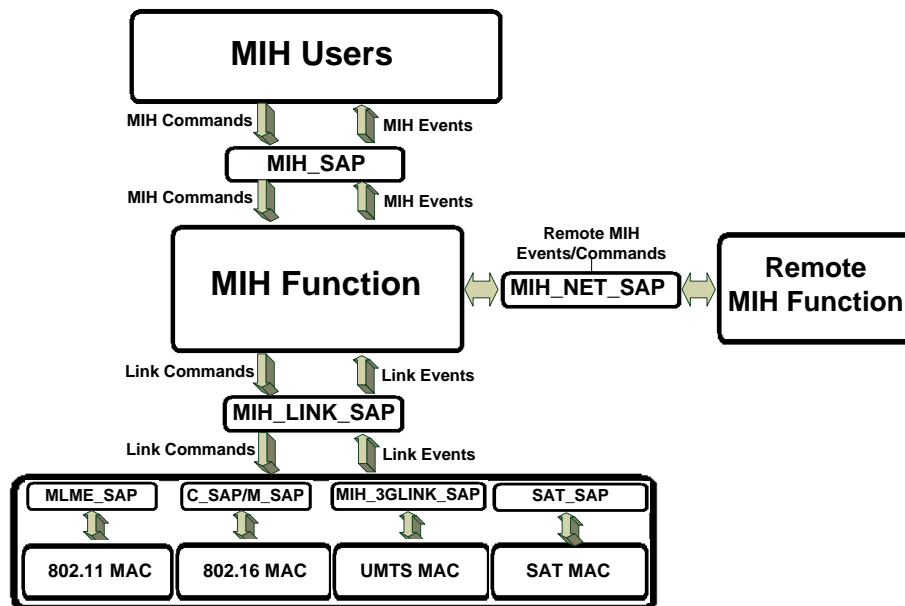
Figure 2-14: IEEE 802.21 reference model extended for satellite support

Entities that use the services provided by the MIHF are called MIH users. An MIHF in a network entity that communicates directly with an MIHF in a mobile node acts as a Point of Service (PoS) of that Mobile Node (MN). The MIHF receives media independent commands from higher layers and translates

43

them to media specific commands for link layer and similarly receives events from different link layer technologies and maps them to corresponding media independent events. The MN exchanges MIH information with its MIH PoS using L3 transport if the PoS is not located in the same network entity as its network Point of Attachment (PoA). The layer 2 at the network side is termed as PoA for the MN.

To facilitate media independent handover, the MIHF provides the following three services:

- **Media Independent Event Service (MIES):** The MIES reports events on dynamic changes in link characteristics, links status and link quality to upper layers through the MIHF.

- **Media Independent Command Services (MICS):** The MICS is used to gather information about the status of the connected links. Upon reception of event notification, MIH users make use of the MICS to pass link commands to the lower layers via the MIHF to manage and control the link layer behaviour for handover decision.

- **Media Independent Information Services (MIIS):** MIIS provides the capability for obtaining the necessary information for handovers, including neighbouring networks, link layer information and service availability. This information will be used to assist network discovery and selection to enable more effective handover.

MIH users access MIHF services through a variety of SAPs. Each SAP consists of a set of service primitives that specify the interactions between

the service user and provider. Three SAPs are currently defined within the MIH framework:

- **MIH_SAP:** The MIH_SAP is present for the upper layer access to the lower layers via the MIHF. A media independent interface provides the interface between the MIHF and the upper layers of the mobility management protocol stack. In order to receive MIHF generated events and link layer events that are forwarded by the MIHF, the upper layers need to subscribe with the MIHF as MIHF users. MIHF users can directly send commands to the local MIHF using the service primitives of the MIH_SAP.

- **MIH_LINK_SAP:** The MIH_LINK_SAP connects the MIHF and the underlying link layers. It is an abstract media dependent interface between the MIHF and media specific link layer to allow MIHF to use services from the lower layers of the protocol stack. For each link layer technology, the MIH_LINK_SAP maps to the media specific SAPs.

- **MIH_NET_SAP:** MIH_NET_SAP for service transport between the local and the remote MIHFs. It is as interface of the MIHF that provides transport services over the data plane on the local node to support the exchange of MIH information and messages with remote MIHFs. Transport services provided by the MIH_NET_SAP can use either L2 or L3 signalling.

The MIH_NMS_SAP was originally proposed to be included in the MIH generic reference model [45] and a set of primitives for MIH_NMS_SAP has also been defined in reference [46]. However since MIHF makes use of

existing management SAPs defined for specific link layer technologies, the MIH_NMS_SAP was not included in the current version of 802.21 standard documents [2]. For example, the MLME_SAP defines the interface between the MIHF and the management plane of an IEEE 802.11 network. In IEEE 802.16 the M_SAP is defined in to provide the interface between the MIHF and the IEEE 802.16 management plane functions.

2.3.4.2 MIH Mapping

This section presents the mapping of MAC layer signalling of all considered access technologies in this research with the *MIH_Link_SAP* primitives. The *MIH_Link_SAP* primitives like *MIH_Link_Detected*, *MIH_Link_Up*, *MIH_Link_Going_Down*, *MIH_Link_Down*, *MIH_Handover_Immenent*, and *MIH_Handover_Complete* have been utilized from the MIH event services. The MIH command service primitives like *Link_Configure_Thresholds*, *Link_Capability_Discover*, *Link_EventSubscribe*, *Link_Event_Unsubscribe*, *Link_Get_Parameters* and *Link_Action* have been used in mapping the link specific primitives. Finally the MIH information service primitive like *MIH_Get_Information* are used. The following subsections briefly describe the mapping for different access technologies with the help of MIHF addition in the protocol stack of each radio access technology.

2.3.4.2.1 Protocol stack and primitives for IEEE 802.11

Figure 2-15 shows the MIHF and the SAPs for the protocol stack of IEEE 802.11. The *L_SAP* can encapsulate the MIH messages in the data frames and provides the interface between MIH function and data plane of IEEE

802.11. The MIH messages can only be transported over the data plane once the mobile node has established its association with the AP.
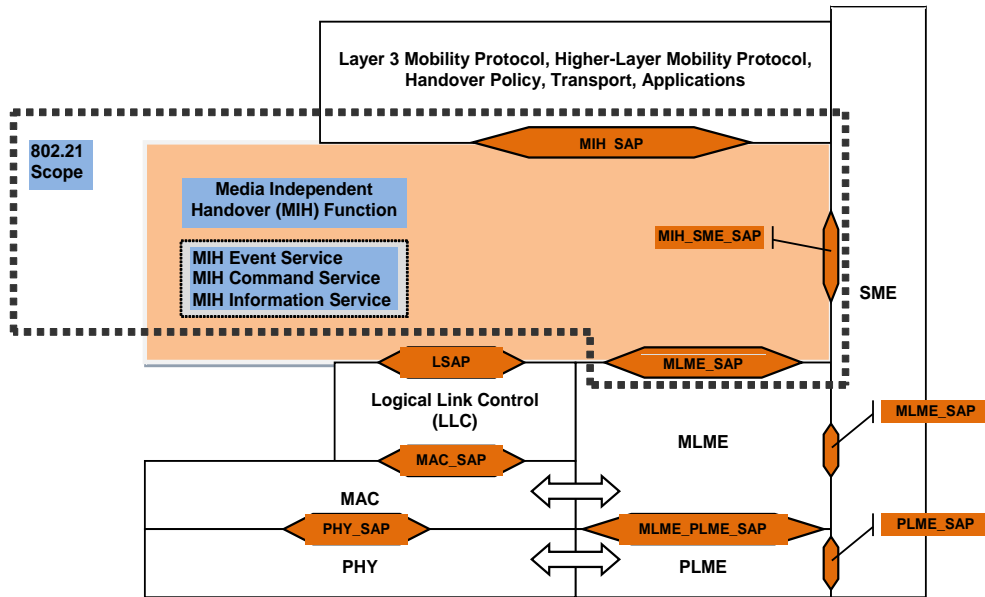
Figure 2-15: IEEE 802.11 protocol stack with respect to MIH reference model [2]

Table 2- 2: MIH primitives mapping for IEEE 802.11 [2]

MIH Link SAP Primitives	Service category	IEEE 802.11 primitives
Link_Detected	Event	N/A
Link_Up	Event	MLME-LinkUp.indication
Link_Down	Event	MLME-LinkDown.indication
Link_Parameters_Report	Event	MLME-MEASURE.confirm, MLME-MREPORT.indication
Link_Going_Down	Event	MLME-LinkGoingDown.indication
Link_Handover_Imminent	Event	MLME-LinkHandoverImminent. Indication
Link_Handover_Complete	Event	MLME-LinkHandoverComplete. Indication
Link_PDU_Transmit_Status	Event	MA-UNIDATA-STATUS.indication
Link_Configure_Thresholds	Command	MLME-MEASURE.request MLME-MREQUEST.request

The *MIH_MLME_SAP* provides the interface between MLME and MIHF and is used before the mobile device establishes an association with the AP. The immediate MIH counterpart in the lower layers of IEEE 802.11 is MAC Layer Management Entity (MLME) therefore it instantiates *MIH_LINK_SAP* in

reference model of MIH for IEEE 802.11. For the transport of MIH messages over L2 before and after the mobile node association with access point, the *L_SAP* and *MIH_MLME_SAP* instantiate the link layer part of the generic *MIH_NET_SAP* as shown in Figure 2-15. Table 2-2 presents the mappings of MIH Link SAP primitives to the IEEE 802.11 primitives [2].

2.3.4.2.2 Protocol stack and primitives mapping of IEEE 802.16

Figure 2-16 represents the MIHF position in the IEEE 802.16 protocol stack. The Network Control Management System (NCMS) and the MIHF share the *M_SAP* and the *C_SAP* for access of mobility management services of mobility control and management entity in the IEEE 802.16 protocol stack. The *C_SAP* provides the interface between MIHF and the control plane.
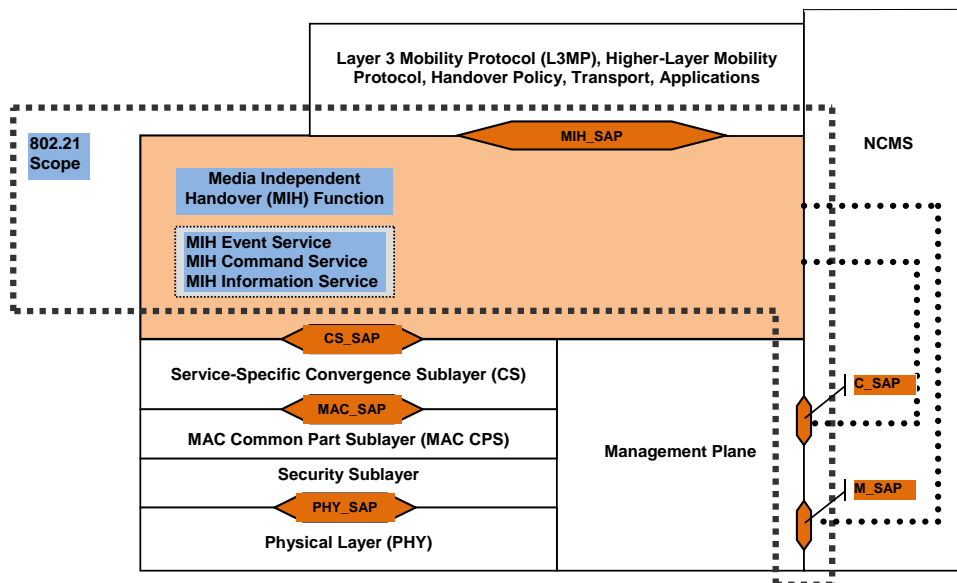
Figure 2-16: IEEE 802.16 protocol stack with respect to MIH reference model [2]

The *C_SAP* and *M_SAP* also transport MIH messages to peer MIHF entities. The Convergence Sublayer Service Access Point (*CS_SAP*) in the IEEE 802.16 provides the interface used to transfer packets from layer 3 to layer 2, once the connections have been established within the network entity. The

M_SAP provides the communication interface between MIHF and the management plane which allows encapsulation of MIHF payload in the management messages.

Table 2- 3: Primitives mapping for IEEE 802.16 [2]

MIH Link SAP Primitives	Service category	Mapping IEEE 802.16 primitives
Link_Detected	Event	*C-NEM-RSP (Ranging)*
Link_Up	Event	*C-NEM-RSP (Registration)*
Link_Down	Event	*C-NEM-RSP (Deregistration)*
Link_Parameters_Report	Event	*C-HO-IND (HO-Scan)*
		C-HO-RSP (HO-Scan)
		C-RRM-RSP
		C-SFM-RSP
Link_Going_Down	Event	
Link_Handover_Imminent	Event	*C-HO-RSP (HO-Mobile)*
Link_Handover_Complete	Event	*C-NEM-RSP (Ranging)*
Link_PDU_Transmit_Status	Event	
Link_Capability_Discover	Command	
Link_Event_Subscribe	Command	
Link_Event_Unsubscribe	Command	
Link_Get_Parameters	Command	*C-SFM-REQ/RSP*
		C-HO-REQ/RSP/IND (HO-Scan)
		C-RRM-REQ/RSP
Link_Configure_Thresholds	Command	*C-HO-REQ/RSP (HO-Scan)*
Link_Action	Command	*C-NEM-REQ/RSP*
		C-SFM-REQ/RSP
		C-IMM-REQ/RSP (Idel_Mobile_Initiation)
		M-SSM-REQ/RSP

The primitives specified by the M_SAP are used by the mobile node to transfer packets to the BS before and after the mobile node has completed the entry procedure in the network. Table 2-3 shows the MIH SAP primitives mappings to the IEEE 802.16 primitives [2].

2.3.4.2.3 Protocol stack and primitives mapping of UMTS

Similar to the other access technologies IEEE 802.21 also provides specification of the SAPs for UMTS support. The media dependent SAP of IEEE 802.21 for UMTS which is called *MIH_3G_Link_SAP*, which is the

interface between layer 2 of UMTS protocol stack layer and the MIHF. There are no new primitives defined in the specification of UMTS as the pre-existing service primitives in the standard have been directly mapped to the MIHF services. Figure 2-17 presents the MIHF and the scope of IEEE 802.21 in UMTS protocol stack.
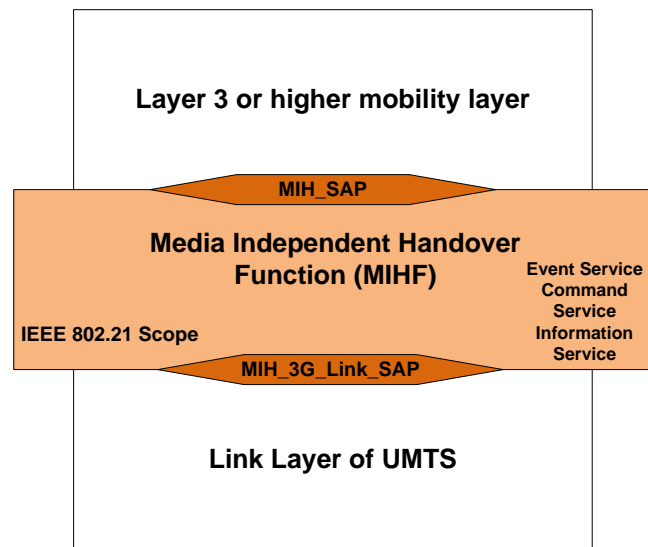
Figure 2-17: UMTS protocol stack with respect to MIH reference model [2]

The mappings of IEEE 802.21 MIH Link SAP with the UMTS MAC layer primitives are shown in Table 2-4.

Table 2- 4: MIH primitives mapping for UMTS [2]

MIH Link SAP Primitives	Service category	UMTS primitives
Link_Detected	Event	*System_Information_Block*
Link_Up	Event	*SMSM-ACTIVE, RABMSM-ACTIVATE*
Link_Down	Event	*SMSM-DEACTIVEATE, SMSM-STATUS RABMSM-DEACTIVATE, RABMSM-STATUS, RABMAS-RAB-RELEASE*
Link_Parameters_Report	Event	*SMSM-MODIFY, RABMSM-MODIFY*
Link_Handover_Complete	Event	*RABMAS-RAB-ESTABLISH, RABMSM-MODIFY*
Link_Configure_Thresholds	Command	*SMREG-PDP-MODIFY*

2.3.4.2.4 Protocol stack and primitives mapping of Satellite network

The support for satellite networks has not been provided in the IEEE 802.21 specifications. Hence, the mapping of satellite primitives with the *MIH_Link_SAP* primitives is proposed in this thesis. The INMRSAT BGAN system has been adopted for this mapping. The media dependent Link SAP for satellite (BGAN) is named as *MIH_SAT_Link_SAP*. The *MIH_SAT_Link_SAP* (media dependent) interfaces the MIHF with underlying BGAN protocol stack and MIH_SAP (media independent SAP) provides the interface between layer 3 and layer MIH function. Figure 2-18 shows the satellite network protocol stack with respect to the MIH reference model.
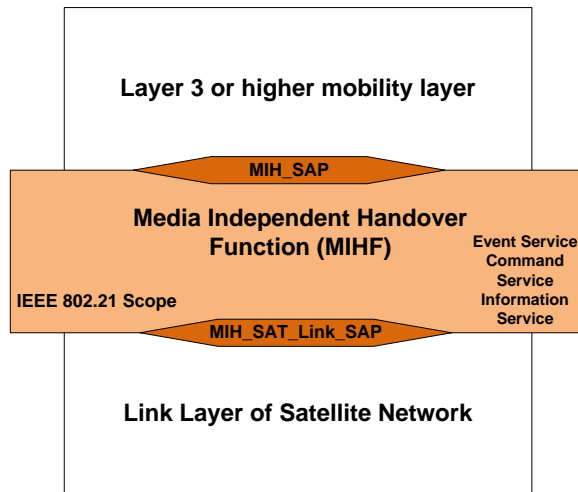
Figure 2-18: Satellite network protocol stack with MIH

Table 2-5 represents the *MIH_Link_SAP* primitives mapping for the satellite network (BGAN). The AT commands are used in BGAN for the control of communication device such as to initiate control operations like power on and power down etc. The AT and AT+ commands are generally used to collect the simple information in communication devices. It provides full control of information which is sent and received over the communication devices [31].

Table 2- 5: MIH primitives mapping for Satellite network (BGAN)

MIH_Link_SAP Primitives	Service category	Mapping BGAN AT commands
Link_Down	Event	+CGEREP: NW DEAC /NW DETACH /ME DEACT/ ME DETACH
Link_Event_Subscribe	Command	+CGEREP (Packet Domain event reporting)
		+CGREG (GPRS network registration status)
Link_Event_Unsubscribe	Command	+CGEREP (Packet Domain event reporting)
		+CGREG (GPRS network registration status)
Link_Get_Parameters	Command	+CLCC (List current calls)
		+CGDCONT?
		+CGDSCONT?
		+CGTFT?
		_ITFT?
		+CGEQREQ?
		+CGEQMIN?
		+CGEQNEG +CGEQNEG? UMTS Quality of Service Profile (Negotiated)
		+CGATT?
		+CGACT?
		+CGCMOD?
		+CGPADDR +CGPADDR=? (Show PDP address)
		+CGEREP?
		+CGREG?
Link_Action	Command	+CGEREP (Packet Domain event reporting)
		+CGDCONT (Define PDP Context)
		+CGDSCONT (Define Secondary PDP Context)
		+CGTFT (Traffic Flow Template)
		_ITFT
		+CGEQREQ UMTS Quality of Service Profile (Requested)
		+CGATT (PS attach or detach)
		+CGACT (PDP context activate or deactivate)
		+CGCMOD (PDP Context Modify)

2.4 Summary

This chapter provides an overview and brief description of the various wireless access technologies that are considered in this thesis. The various characteristics like coverage, throughput/bandwidth, standards, advantages and disadvantages of these different technologies were also presented. The different wireless access technologies inherently differ from each other in a number of characteristics like availability, throughput, QoS, latency, packet loss, jitters and delays. Therefore the QoS requirements and end-to-end interworking issues of wireless access technologies in heterogeneous access networks possess multiple challenges.

A discussion on the role of satellites in future networks was also presented. The main advantage of the satellite communications is the wide coverage area, however their high cost in deployment and usage and lower data-rates and higher delays stop them from being adopted on a large scale. Satellite can still retain exclusive status in some particular areas like the maritime and aeronautical markets due to its distinctive coverage feature. However the rapid growing demand for broadband and multimedia services anytime and anywhere requires the integration of satellite technologies with the terrestrial access technologies to achieve efficient service delivery and global coverage and for the exploitation of new services in densely populated big cities and rural areas with sparsely located population. In this research the IEEE 802.21 standardization is adapted for the interworking of heterogeneous wireless access networks by extending it to include the satellite networks. The IEEE 802.21 network architecture and the primitive mappings for the different wireless access technologies have been also presented in detail.

Chapter 3: LOAD BALANCING IN HETEROGENEOUS WIRELESS NETWORKS

3.1 Overview

The ever increasing user QoS demands and emergence of new user applications, make the job of network operators and manufacturers more challenging for efficiently optimisation and managing the available radio resources in pools of different wireless access technologies. Particularly in areas, where different wireless access networks are providing coverage simultaneously.

A group of strategies or mechanisms which are collectively responsible for the efficient utilisation of radio resources available within the RAT are termed as Radio Resource Management (RRM). RRM is composed of Handover Control (HC), Power Control (PC), Admission Control (AC), Packet Scheduling (PS) and Congestion Control (CC) operations. The traditional RRM strategies are implemented independently in each RAT, as each RRM strategy considers the attributes of a particular access technology. Therefore traditional RRM strategies are not suitable for heterogeneous wireless networks.
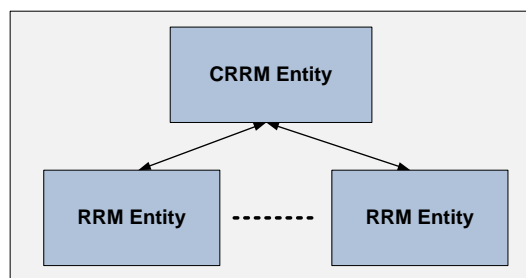
Figure 3-1: Two Tier RRM Model

Common Radio Resource Management (CRRM) [47, 48] or joint radio resource management strategies are proposed for coordinating the radio resource management between multiple RATs in an improved manner. The concept of CRRM is based on a two tier RRM model [49] as shown in Figure 3-1. The RRM manages the radio resource units within a single RAT and is located at the lower tier of the two tier model. The CRRM is located at the higher tier of the model and it controls different RRM entities and it can also communicate with other CRRM entities. The CRRM may make management decision based on the information collected from different RRMs representing different RATs.

There are two decision making methods in RRM such as RRM centred and CRRM centred. In RRM centred decision method the CRRM provides information to the individual RRM entities which then make the final decision. In the CRRM centred method, the CRRM entity makes the decision and informs the RRM entity to execute the decision for RAT selection. In this thesis the CRRM centred method is adapted for better scalability and other reasons which have been elaborated further in upcoming sections of this chapter.

From a network topology point of view, the CRRM functionality can be implemented in various different ways such as CRRM server approach [50, 51], integrated CRRM approach [52], hierarchical CRRM approach [53], CRRM functions in User Terminal (UT) approach [54] and a hybrid approach which can be combination of these approaches. While in the CRRM server approach, a separate CRRM server is added in the core network, in the integrated CRRM approach, the CRRM functionality is added within an

existing network entity like the base station (BS), the Radio Network Controller (RNC) or the Access Point (AP). The CRRM server is a centralised approach due to which it attains high scalability. The integrated CRRM requires minimum infrastructure changes and also reduced the communication delays between the local RRM and CRRM entities. However this approach is distributed and does not scale well due to the large number of connection between the various local RRM entities.

The hierarchical CRRM approach divides the problem into various layers and each layer is managed by a dedicated management entity. This approach adds further complexities due to a number of new entities additions in the architecture infrastructure. In the final approach, the CRRM functions are present in the end user terminal. This approach allows the mobile node to make decision for suitable RAT selection. In this case, the network needs to provide enough information to the mobile nodes, but this would require extra signalling.

In this thesis a hybrid of CRRM server and CRRM functions at the user terminal approach is applied to get advantages of both centralised and distributed approaches. Figure 3-2 represents the proposed CRRM approach for this research. Figure 3-2 is composed of three layers namely, the Core Network (CN), the access network entities and the User Terminal (UT). For the load balancing purpose each of these layer are equipped with IEEE media Independent handover (MIH) components i.e. CRRM server acts as Media Independent Information Server (MIIS) and similarly the CRRM entity in the mobile node or UT communicate with the RRM entities in the network side using IEEE 802.21 MIH reference model.
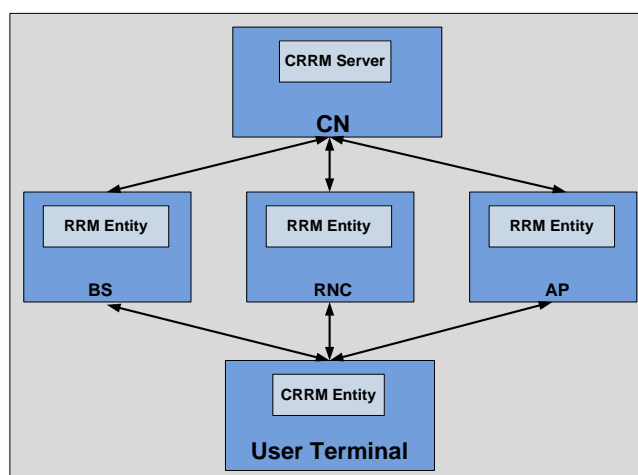
Figure 3-2: proposed CRRM approach

Before presenting our proposed load balancing framework for heterogeneous wireless access networks it is important to review existing related mechanisms. An efficient Radio Access Technology (RAT) selection mechanism provided by CRRM plays a key role in a load balancing framework as it enables a multi-mode terminal to select a suitable access technology from the different available technologies based on different criteria. This chapter covers a review of such RAT selection techniques in the heterogeneous wireless networks. This chapter also presents a review on some load balancing techniques in heterogeneous wireless networks. It also presents in detail the architectures and algorithms which have been previously presented for the load balancing in the future heterogeneous wireless networks. In the end the advantages and disadvantage for load balancing in heterogeneous wireless network are discussed.

3.2 RAT Selection in heterogeneous wireless networks

It is envisaged that future wireless access networks will comprise of co-existing multiple radio access technologies such as satellite networks and terrestrial networks like WiMax, UMTS and WLAN. To achieve seamless

interworking of these multiple RATs, a Common or Joint Radio Resource Management (CRRM/JRRM) is introduced in literature to provide efficient radio resource utilization [48]. The algorithms used for RAT selection in such integrated heterogeneous wireless networks form an important component for CRRM/JRRM. The role of RAT selection algorithms is to validate the suitability of available RATs in the heterogeneous wireless networks. The RAT selection algorithms must ensure the most efficient way for the utilization of available radio resources while provisioning the guaranteed required QoS for all active connections. The traditional RAT selection strategies do not provide solutions for heterogeneous wireless networks as they were designed for homogeneous radio access networks. For this purpose RAT selection strategies have been revised so that they can perform efficiently in heterogeneous wireless networks environment.

There have been several proposals for the RAT selection algorithms in heterogeneous wireless networks which have been classified into multiple categories. The authors of references [55] and [56] compare some of these RAT selection algorithm based on their advantages and limitations. In general, RAT selection algorithms in heterogeneous wireless networks can be categorised as follows:

- Random based RAT selection algorithm
- Fixed Policy based
    - Service-class based RAT selection algorithm
    - Service-cost based RAT selection algorithm
    - Signal path based RAT selection algorithm
    - Layer based RAT selection algorithm

- Dynamic decision based RAT selection

    o Utility Function based RAT selection

    o Artificial Intelligence (AI) techniques based RAT selection algorithm (fuzzy, neural-fuzzy, fuzzy with evolutionary optimisation etc.)

These various RAT selection algorithms are further explained with references in the following sub-sections.

3.2.1 Random based RAT selection algorithm

During the vertical handover (inter wireless technologies handover) procedure, this RAT selection algorithm will randomly select one from the available RATs. Generally the call will be dropped or blocked in situations where there are no available radio resources on the selected RAT. If the randomly selected RAT has enough resources, then the call shall be admitted. On the other hand if the selected RAT cannot serve the call due to lack of enough radio resources, the algorithm will randomly select another for this call, and the procedure repeats. In the worst case, if none of the RATs have enough resources for the call, then the call is simply blocked or dropped. This algorithm is usually used as a baseline algorithm for performance evaluation of other existing RAT selection algorithms. The advantages of using this algorithm is that it is simple and easy to implement however it has comparatively highest call dropping and blocking probabilities, less radio resource utilization [56].

3.2.2 Fixed Policy based

3.2.2.1 Load balancing based RAT selection algorithm

The main goal of the load based RAT selection algorithms is to distribute the networks/RATs traffic load uniformly between the heterogeneous wireless networks which have common or shared coverage area. The mechanism of balancing the load between such heterogeneous wireless networks provides the better radio resource utilization [48, 57]. The load balancing can be performed in different fashion in heterogeneous wireless networks such as: performing load balancing when the available resources have reached to a particular minimum threshold as in reference [58], or after certain time intervals like in [57], or performing load balancing upon certain events such as call arrival, departure or handover or by reaching a particular calculated decision value. The load balancing can also be carried out by forcefully handing over active call(s) from heavily loaded network to the least loaded network or it can be performed when new call or mobility based handover occurs. The former approach is called forced load balancing and the later approach for load balancing is known as unforced load balancing. The load balancing algorithm provides the high network stability advantage to the RAT selection procedure; however load balancing can sometimes lead towards the low user satisfaction and in case of forced load balancing the trade-off of handover signalling overhead goes high where number of forced handovers increases. The proposed algorithms in this research are also load based RAT selection algorithms. The RRM in the network assist the RAT selection algorithm by providing the required information, whereas the RAT selection

mechanism based on mobile node processes the information and make decision for load balancing base handovers.

3.2.2.2 Service-class based RAT selection algorithm

The service-class based RAT selection algorithm is based on concept that different access technologies are designed for provision of different classes of services. For example WLAN is designed for data services, GSM is designed for voice services and UMTS is designed for both data and voice services. The service class based RAT selection mechanism admits arriving calls to the corresponding RAT only, e.g.: streaming or data calls to UMTS or WLAN and voice calls to the GSM [59]. This RAT selection algorithm is connection-centric and achieves the high QoS provision to the users. There are high chances of admitting a large number of calls to a particular RAT when this algorithm is used and this may lead to high load variation among co-located heterogeneous wireless networks.

This mechanism can further be categorised into rigid and flexible service based RAT selection. In the first case, the algorithm will try to allocate the arriving call only to the corresponding RAT type and in case the RAT specified for this call type does not have enough radio resources, then the call will be dropped or blocked. On other hand, the flexible service class based RAT selection algorithm first attempts to admit the arriving call on the dedicated RAT for that call type and if there are no radio resources available in that type of RATs then the algorithm starts attempting to admit the call on another RAT type. As compared to the Rigid version of this mechanism, the

flexible version of this algorithm has low call blocking and dropping probability due to its flexible nature in admitting call to any type of RATs.

The authors of reference [59] evaluated the performance of the service based RAT selection mechanism for GSM and UMTS networks with overlapping coverage areas. Both RATs are assumed to have same service capacity in the simulation for three different classes of traffic such as streaming, voice and data. It was seen that the best performance was obtained in the case where streaming calls were allocated to UMTS with high priority, the voice calls are allocated to the GSM with priority and data users are allocated to the least loaded RAT [59].

The service class based RAT selection algorithm is also adopted in reference [60], where the UMTS and WLAN networks are considered with two traffic classes such as voice and data. The policy adopted in this paper is that for the overlapping coverage area, the UMTS network is given higher priority for voice calls and the data connections can only be admitted on WLAN. The strategy was to reduce the number of voice calls handovers between UMTS and WLAN and it was validated in results.

3.2.2.3 Service cost based RAT selection algorithm

The main objective of the service-cost based RAT selection mechanism is to allocate the arriving calls to the least expensive technology in order to reduce the costs for the user. The usage tariffs are usually different for different networks, e.g.: while the usage tariff for WLAN may be quite low, it would be very high for using a satellite network. This algorithm can result in highly unbalanced networks load situations as every user will generally prefer the

least expensive network, thereby overloaded it all the time. In reference [61] the service-cost based algorithm is evaluated and its benefits are demonstrated in the heterogeneous wireless network comprising of WLAN and UMTS. It showed that in the overlapping coverage area, by utilizing the WLAN network to its maximum capacity before admitting calls to UMTS can result in overall the least cost scenario for the end users. However the two networks would remain in highly un- balanced situation as WLAN would remain overloaded all the time and UMTS network would be lightly loaded. It also suggests that networks profits for the two RATs would be different as less people would use the UMTS network.

3.2.2.4 Path loss based RAT selection algorithm

The path loss based RAT selection strategy admits the arriving calls in the RATs with lowest path loss measurements. The path loss based RAT selection can introduce high number of handovers; however they have the benefit of high throughput and low bit error rate. In reference [62] the authors have computed the received power of the RATs such as UTRAN and GERAN in the mobile terminal during the establishment phase. The arriving call will be admitted to the GERAN if the path loss of UTRAN is higher than a particular threshold otherwise the call will be admitted to UTRAN. In reference [63] the authors presented a path loss based RAT selection algorithm for the heterogeneous wireless network considering CDMA/TDMA based network. In this algorithm the high path loss based mobile users are connected/allocated to the TDMA based network and the low path loss mobile users to CDMA network. A hysteresis margin is introduced in this proposed algorithm to avoid the Ping-Pong effects of undesired handover

between different RATs in heterogeneous wireless networks. The simulation results validated the reduction in call blocking and frequency of handovers when using the proposed algorithm.

3.2.2.5 Layer based RAT selection algorithm

In heterogeneous wireless networks where different wireless networks share common coverage area, the layer based RAT selection technique assigns arriving calls to a particular layer. If the layer could not service the arriving call due to low resources then the call is forwarded to the next level of layer. In this mechanism different access networks are considered as different layers in the RAT selection procedure. This technique is simple to implement but it might introduce highly unbalanced load situation in the heterogeneous wireless networks. In reference [64] a layer based predictive RAT selection algorithm is proposed for coexisting heterogeneous wireless networks. In this algorithm the arriving calls are admitted in layer k and if there are no resources in the layer k and arriving call is blocked then the call is forwarded towards the next layer by algorithm for seeking availability of resources. The algorithm keeps on checking for the available resource for the call until it searches all the available layers. The call blocking and dropping is minimized using this algorithm. The heterogeneous wireless network composed of three RATs is considered which are UMTS, GPRS and WLAN. Comparison of the results using proposed algorithm and independent admission control in different RATs showed improvements in call dropping however this algorithm leads to highly unbalanced load among three layers of considered radio access networks.

3.2.3 Dynamic decision based RAT selection

3.2.3.1 Utility function based RAT selection algorithm

The particular RAT is selected in this approach based on the certain utility or cost functions as proposed in reference [65] or a fittingness factor as presented in reference [66, 67]. In reference [68] a utility function based RAT selection approach for heterogeneous wireless networks is proposed. The concept of arbitration probability is utilized in this approach which indicates the data user degree of willingness to use a particular network's resources. The factors like user's satisfaction on QoS, link quality and monetary cost are considered while computing the arbitration probability value. Once the user has computed the arbitrary probability values of all the available networks, the network with best value is selected in RAT selection process. This RAT selection algorithm has high computational overhead and is complicated to implement however it has benefit of high efficiency.

3.2.3.2 (AI) techniques based RAT selection algorithm

Artificial intelligent techniques may also be used for developing efficient RAT selection algorithms that may consider various different policies and QoS parameters. Fuzzy logic is the most commonly used approach for developing intelligent RAT selection techniques. Such techniques while have high efficiency and improved users satisfaction, they are usually more difficult to implement as compared to standard algorithms. A fuzzy control system is composed of the fuzzifier, the fuzzy rule base, the fuzzy inference engine and the defuzzifier.

In reference [69] the authors proposed a fuzzy logic based RAT selection algorithm which considered different input metrics like available candidate networks, application QoS requirements, user defined criteria, etc. In reference [70], a fuzzy Multiple Attribute Decision Making (MADAM) based RAT selection algorithm is proposed which considers parameters like user's preferences, battery status, latency, reliability, cost, signal strength and bandwidth. The main aim of the algorithm was to select a suitable RAT for a particular service class, based on previously mentioned norms. In reference [71], a fuzzy Multiple Objective Decision Making (MODM) approach is proposed for selecting a suitable RAT for handover calls in heterogeneous wireless networks. The algorithm considers parameters like data rate, network type, call arrival rate, transmission delay and coverage of the network for evaluation for decision making.

While all these use a Fuzzy based RAT selection approaches that consider various parameters, none of them considers network load conditions. Hence all of these techniques result in high load variation across the different networks with some networks being more utilised than others. This also in the long run results in such networks becoming full, thereby resulting in call blocking and call dropping. Hence it is important that an efficient RAT selection algorithm also considers the network load conditions to be able to avoid such overloading situations.

3.3 Load balancing in heterogeneous wireless networks

3.3.1 Need for load balancing

The deployment of heterogeneous wireless networks is increasing in order to meet the ever rising demands of the users for anytime anywhere network service availability. It is envisaged that such networks of the future shall involve a collaboration of multiple radio access networks [72, 73, 74]. Satellite networks technologies and terrestrial wireless network technologies such as UMTS, WiMax and WLAN are used to provide network access for both voice and data services. Usually more than one wireless networks may provide coverage to any given location in an urban area. In densely populated areas like town centres, shopping centres and train stations of big cities, a large number of mobile users may be connected to the more common UMTS even though other access technologies may be available. This results in unbalanced loading across these wireless networks. It has been seen that these popular networks may get overloaded in some situation leading to poor service. In order to solve this problem, it is important to be able to use different available networks thereby distributing the load amongst them. Now days, new smart phones may allow using different networks for different services. Hence for example when working in an office building, the mobile device of a user may be in the coverage of a UMTS mobile network and a WLAN office network. In such a situation, users may manually configure their devices to use the UMTS network for voice services and the WLAN access for data services.

However, this requires the user's intervention every time and also does not provide seamless transition between networks when such users are on the move. Therefore to achieve a more seamless, automatic and efficient load distribution across networks, a load balancing algorithm is required. This high load variation can be balanced by moving mobile users from heavily loaded networks to least loaded networks which involves execution of vertical handovers. The considerable benefits of the load balancing mechanism are that it can provide better services for the users, enlarge the wireless network capacity and improve the radio resource utilization.

Seamless vertical handovers across different wireless networks may be achieved using the IEEE 802.21 Media Independent Handover (MIH) specifications. Modern mobile devices like cell phones, smart phones, Tablet PCs already support multiple wireless technologies like UMTS, WLAN and Bluetooth and in the very near future could also support satellite and WiMax with multiple interfaces provision. While most of these devices are able to scan the different available networks, the user would manually select which network he or she may want to use. It is envisaged that in the future generation heterogeneous wireless networks, these devices may be able to apply some complex RAT selection techniques to find the most suitable network. Such a RAT selection technique may need to consider various parameters like received signal strengths, errors rates, costs, user preferences, load, QoS requirements, etc. Such a RAT selection technique would not only play an important part when a user switches on his or her mobile device but also when the user moves around.

While most of the current day mobile networks already support seamless handovers, these are restricted to handovers within the same technology, i.e. horizontal handovers. It is envisaged that to efficiently use the network services the future mobile devices shall also support handovers across different radio access technologies. This process of switching mobile devices connectivity from one technology to another type of technology is called vertical handover.

The JRRM strategies for next generation heterogeneous wireless networks are envisioned as user-centric. User centricity implies that user's preferences are considered in decision making for RAT selection. However user-centric JRRM algorithms often lead to highly unbalanced networks load, which cause congestion on overloaded network and eventually increase the call blocking and call dropping probabilities. The unbalanced load situation in co-located networks also causes the poor radio resource utilization as some networks remain under loaded and some get overloaded. RAT selection techniques for JRRM aim to find the most suitable network that a mobile node should be connected to, for achieving seamless services and meeting the QoS requirements of the user. Traditional RAT selection algorithms are mainly based on the Always Best Connected (ABC) paradigm whereby the mobile nodes are always directed towards the available network which has the strongest and fastest link. This however could create a high variation among the load across the different co-located networks; which cause congestion on overloaded network and eventually increase the call blocking and call dropping probabilities. The unbalanced load situation in co-located networks also causes the poor radio resource utilization as some networks

remain under loaded and some become overloaded. There is a need for the load balancing strategies to efficiently utilize the available radio resources and avoid the unwanted congestion situations on overloaded wireless networks. The load balancing strategies are required to efficiently utilize the available radio resources and avoid the unwanted congestion situations due to overloaded wireless networks.

3.3.2 Load balancing strategies

The load balancing in heterogeneous wireless networks can be achieve in different ways such as using network controlled handovers or using network assisted mobile controlled handovers. The load balancing using network controlled handovers can either be one of the following:

- Periodic balancing
- Event driven load balancing

In periodic load balancing the load balancing algorithm performs the load balancing operation on network side periodically after a specific amount of time. The event driven load balancing is triggered every time the particular events occur for example call established, handover or call completed. Both event driven load balancing and periodic load balancing have their limitations. In case of periodic load balancing the no load balancing is performed until the particular time has be elapsed and load in collocated networks remain unbalanced in that time period. The event driven load balancing can generate signalling overhead as load balancing operation requires other networks information. A hybrid approach can be beneficial as it can combine the advantages of both periodic and event driven load

balancing and mitigates the drawbacks. In this thesis a hybrid approach is adapted, which combines both periodic and event driven approach. The load balancing using network assisted, mobile control handovers is usually event driven and is coupled with mobile nodes mobility. Different MIH events trigger, as the mobile node moves across the coverage areas of different wireless networks. These events can be utilised to trigger the load balancing in mobile node.

3.3.3 Load balancing mechanism

The load balancing mechanism can generally be divided into two main parts in heterogeneous wireless networks [75] the load balancing algorithm and the network architecture. The later part which is network architecture is the basics for the efficient load balancing and good network architecture can improve the efficiency of load balancing mechanism. From the control mode perspective the load balancing mechanism can be categorized as distributed, semi-centralized semi-distributed and centralized load balancing mechanism [76, 77]. Both pure centralized and distributed approaches have issues as distributed approach in the heterogeneous wireless network architecture will have a huge overhead and the centralized approach will have low reliability [78].

3.3.3.1 Load balancing architectures

The approach presented in reference [79] provides a mathematical framework which could be used to represent and analyse the heterogeneous wireless networks that converge for the sake of interoperability. The authors generalized the bonacich centrality equation which measures the connectivity

between nodes by the number of routes between them towards heterogeneous wireless networks and used it for the study of heterogeneous wireless network architecture. In reference [77] the authors proposed a Semi-Distributed and Semi-Centralized Architecture (SCSDA), which is used in such a way that BSs exchange load information with other neighbouring BSs. This architecture in theory is able to reduce the overhead of control signalling but it was not proved in the paper by the authors using simulation or analytical model. A hybrid wireless network architecture design presented in reference [80] and whereas network architecture based on multiple mobile routers to support seamless mobility across future heterogeneous wireless network is presented in reference [81]. Both reference [80] and reference [81] verified the reduced overhead by NS2 simulations however the model was not derived in these approaches. Theoretical route overhead is presented in reference [82] by counting the number of control messages generated in the network entities such as BS or AP for route maintenance. In reference [83] the communication overhead of the presented mechanism was calculated and to minimize this communication overhead an algorithm was presented. The presented algorithm was proved effective with the help of simulation results. The general heterogeneous wireless network was considered in reference [84], where two basic network entities the mobile node and the AP are considered. This approach formulated the overhead for the discovery of AP by dividing it into RREQ messages and HELLO messages and proved the effectiveness of their proposed method with help of simulation results. This approach is not very beneficial in case where multiple networks line satellite, UMTS, WiMax and WLAN are forming the heterogeneous wireless

network. The mechanism presented in reference [85] proposed a hierarchical distributed architecture with three levels of hierarchy in mobility management. The three levels of mobility are as follows: i) the end terminal changes its point of attachment but remains connected to the same radio access network. ii) The end terminal remains associated to the same operator but changes its radio access network. iii) The end terminal changes its operator network. In this paper the authors also estimated the signalling cost during the QoS negotiation for handover process. In reference [86] a hierarchical semi-centralized architecture is presented which considered heterogeneous wireless networks including WiMax, WLAN and UMTS to share network load. This approach introduced new entities like resource allocator, information servers and resource statistics and named them collectively as Resource Management Unit (RMU). The authors have also provided list of signalling's between the newly introduced entities in the network architecture. The disadvantage of this approach is that it is not standardized approach for integration of heterogeneous wireless networks and requires upgrades in wireless networks architecture as well as in protocol stacks. On the other hand the proposed approaches in references [72, 73, 111] have also adopted semi-centralized architecture but utilized the enhanced MIH [2] which is a standardized and more efficient. In reference [87] the authors adopted the loosely coupled architecture for the cellular/WLAN integration [88] for the load balancing purpose between heterogeneous wireless networks. The authors adopted the two phase control strategies in the load balancing policies. The dynamic vertical handover during the traffic serving phase is used to make the performance variance smooth and call admission is used to

provision static QoS guarantee during the admission phase. The effectiveness of proposed strategy in this approach was proved with the statistical comparison of results with other similarly presented references. In reference [89] the authors presented a dynamic load balancing architecture for the load balancing in heterogeneous wireless access networks which considers WiMax, 3GPP Long Term Evolution (LTE) and WLAN. This approach adopted the central architecture for load balancing and introduced new entities in the wireless networks architecture, like Community Resource Manager (CRM), Local Resource Manager (LRM), Community Access Point (CAP) and Spectrum Manager (SM). The approaches presented in references [48, 90, 91] discussed the inter RAT load balancing algorithms and [92] presented the usage of radio enabler in IEEE-P1900.4. The discovery of RATs is assisted from the information provided by the radio enabler. In reference [5] load balancing approach has been presented which targets the Proxy Mobile IPv6 (PMIPV6) domain using MIH for heterogeneous networks. A comparison has been made between the scenario performing load balancing in extended PMIPV6 for handover signalling and the scenario using MIH signalling for load balancing. It was shown in the results that use of load balancing improves the efficiency whereas, MIH based load balancing improves data rate as compared to extended MIPV6 based load balancing. This disadvantage in this approach is when considering load-aware RAT selection; it is specifically designed for a MIPV6 architecture using Local Mobility Agent (LMA) and a new entity called Mobile Access Gateway (MAG) in the network. In reference [93] the distributed architecture is acquired by the authors, which uses a user-centric

Joint call admission control strategy for the load balancing in heterogeneous wireless access networks. The proposed architecture in this thesis for load balancing is semi-centralized and semi-distributed as the load balancing is performed at both ends such as mobile node and the network entity like BS, AP, RAN and RNC.

3.3.3.2 Load balancing algorithms

The load balancing approaches presented in reference [94] and reference [95] have considered load balancing in homogenous network targeting WLAN. The approach in reference [94] considers the Received Signal Strength Indicator (RSSI) value to distribute the load between different Access Points (APs) which have overlapping coverage areas. This approach uses two values in balancing the load which are RSSI between Mobile Station (MS) and AP and the average RSSI value of all the MSs currently connected with AP. The method given in reference [95] considers both RSSI and the number of MS associated with AP which makes it much effective for load balancing. The technique used in reference [96] presented a solution for load balancing in homogeneous wireless networks, by utilizing genetic algorithm. As the genetic algorithm's convergence directly proportional to the size of population (mobile nodes and APs) therefore this approach is effective for WLAN networks and not for the heterogeneous wireless environment where population size is comparatively large due to large coverage areas. All approaches given in references [95, 96 and 94] were designed to enhance the performance for homogeneous network environment particularly for WLAN.

In reference [97] load balancing approach has been presented which targets the PMIPV6 domain using MIH for heterogeneous networks. A comparison has been made between the scenario performing load balancing in extended PMIPV6 for handover signalling and the scenario using MIH signalling for load balancing. It was shown in the results that use of load balancing improves the efficiency whereas, MIH based load balancing improves data rate as compared to extended MIPV6 based load balancing. This disadvantage in this approach is when considering load-aware RAT selection; it is specifically designed for a MIPV6 architecture using Local Mobility Agent (LMA) and a new entity called Mobile Access Gateway (MAG) in the network. In reference [98] a general set of algorithms have been proposed which considers battery power of mobile users, received signal strength and load on available points of attachments in handover process to balance the load in co-located networks overlapping their coverage areas. In this approach load balancing is done only at network side without any interaction with the mobile node. On the other hand our proposed approach considers both; mobile nodes and network entities such as AP, BS and satellite ground station for load balancing thereby resulting in more efficient load balancing across the neighbouring networks.

In reference [99] a detailed algorithm has been presented for network selection in heterogeneous wireless networks. The algorithm presented in reference [99] has been divided into two parts, one runs at mobile terminals and other part of algorithm runs at network entity such as base station (BS) or Access Point (AP). This approach considers received signal-strength, battery power, speed, and location of mobile user but does not considers

MIH which could have improved the handover process while moving the mobile nodes between different networks.

In reference [57] a Next Generation Networks (NGN) based approach has been presented in which hierarchical joint call admission control algorithm is extended to send newly added load reports from Hierarchical Call Admission Control (HCAC) entity to Vertical Call Admission Control entity (VCAC). The main goals of proposed approach in reference [57] are simplicity and scalability, however this approach performs balancing operation periodically and therefore may not performs very efficiently with abrupt load changes in different sub networks in the hierarchy. It also requires the implementation of HCAC and VCAC entities in the network. This approach performs load balancing only at network side which requires and does not consider the RAT selection at mobile node side as the proposed approach in this thesis is performing.

3.4 Advantages and disadvantages of load balancing

The process of load balancing in heterogeneous wireless network has its advantages and also some minor shortcomings. Some of the major advantages of load balancing in heterogeneous wireless networks are as follows:

- It reduces the call blocking and dropping probabilities.
- It reduces the congestion in the network by sharing the load between co-located wireless networks.
- It minimizes the number of handovers performed by the average mobile node.

- It reduces the total handover latencies observed by the average mobile nodes in the network.

- It offers an efficient way of utilizing the available radio resource.

- By avoiding congestion it offers high throughput and minimized drop ratio.

The load balancing process offers a number of advantages, but also has some drawbacks; some of them are given as follows:

- It does not always guarantee the best network such as the network with lowest latencies.

- Sometimes it may also allow higher end-to-end delays but acceptable to the application running on mobile nodes.

- Additional processing is required which needs to upgrade or integrate a module in the existing protocol stack.

- Additional signalling overhead is introduced while sharing the network information.

There may be some other shortcomings of using load balancing with a number of advantages. However this trade-off of acquiring major advantages with minor disadvantage is always adaptable.

3.5 Summary

In this chapter different aspects of load balancing in heterogeneous wireless networks are explored. Starting from the RRM and then importance of load balancing in RRM and different approaches for performing load balancing are also discussed with the help of previously proposed approaches for load

balancing in heterogeneous wireless networks. The shortcomings and benefits of different load balancing strategies are also explained such as periodic, event driven and hybrid load balancing strategies. It is concluded that important components of load balancing framework such as suitable architecture, improved strategy and efficient algorithm are some of the essential building blocks which need to be considered while designing a load balancing framework for heterogeneous wireless networks. The distributed load balancing framework using RAT selection on mobile nodes is more efficient as in this case the signalling overhead is reduced. The proposed load balancing architecture is semi-centralised and semi distributed. As the CRRM server and mobile node based RAT selection is used in the proposed load balancing architecture. Three different types of RAT selection algorithms such as baseline, fuzzy and fuzzy neural algorithms have been presented in this thesis targeting the load balancing during the RAT selection procedure.

CHAPTER 4: LOAD BALANCING FRAMEWORK IN HETEROGENEOUS WIRELESS NETWORKS

4.1 Overview

This chapter explains the load balancing framework and algorithms design. The chapter begins with a detailed explanation of the proposed load balancing framework which adopts the IEEE 802.21 Media Independent Handover (MIH) reference model for seamless handovers across different access technologies. The design of the load balancing mechanism is also briefly described elaborating the location of cognitive and non-cognitive load balancing algorithms in the protocol stack with respect to the MIH reference model. The mappings between the access technologies specific signalling and the MIH signalling and the handover procedures are explained in detail with the help of Message Sequence Charts (MSCs).

4.2 Architectural design for load balancing

More than one wireless networks may typically provide coverage to any given geographic area. For example, when in town centres or other public places like train stations, a mobile node may be in the coverage of a Universal Mobile Telecommunication System (UMTS) [11, 12] mobile network and a WLAN [8] network. A user may manually configure the mobile node to use the UMTS network for voice services and the WLAN access for data services. The area may also be in the coverage of other technologies like satellite and Worldwide Interoperability for Microwave Access (WiMax) [9]. In such overlapping areas, a Radio Access Technology (RAT) selection helps in finding the most suitable network based on received signal strengths, errors

rates, costs, user preferences, QoS requirements, etc. An efficient RAT selection process should aim to balance the load across the different networks in order to avoid over utilisation of a particular network while the others are underutilized. Load balancing techniques have been explored for more than two decades in the field of computing but it is still a relatively new area in wireless communication networks [100, 101, 102 and 103]. In computing, load balancing techniques are used extensively for balancing the load across different back-end servers. Whereas the need for the load balancing in the field of wireless communication is for efficiently utilizing all available radio access technologies and avoiding unwanted situation such as congestion, call blocking and call dropping which are caused by unbalanced utilization of radio access technologies. In this thesis, the load balancing framework considers both mobile nodes and as well as the network for load balancing. The algorithm running on mobile nodes make sure that mobile nodes select the least loaded network based on the considered parameters and the algorithm running on the network side keep on monitoring the network load and initiate the load balancing process upon unbalanced and overloading states.

Figure 4-1 presents the target network architecture considered in this research. It shows an MIH enabled multi-interface mobile node which can use any of the four available wireless access networks (satellite, WiMax, WLAN and UMTS) [73] supported by its interfaces. It is assumed that a single operator is controlling all the wireless networks hence all four wireless networks share a common core network. The core network is in turn connected to the Internet. The mobile node can communicate with a

correspondent node over the internet, using any available wireless network which it supports. On-going sessions would be handed over to another available network without losing any connectivity if the mobile node moves out of its current network coverage and enters into another network.
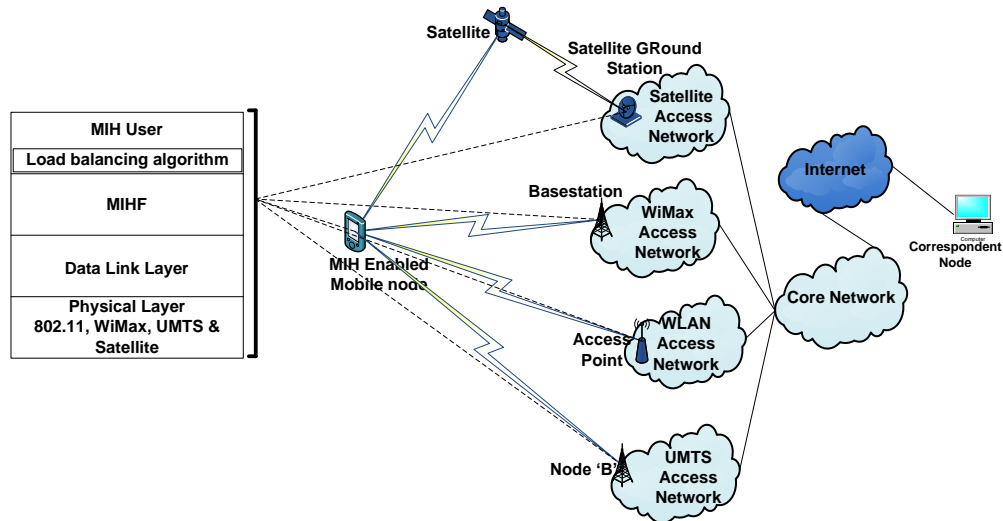
Figure 4-1: Load Balancing Architecture design

The key phenomenon in the MIH reference model is the introduction of Media Independent Handover Function (MIHF) between layer 2 and layer 3 of the OSI layer model. The MIHF receives and transmits the information about the network condition and configurations of the access networks around the mobile node, regardless of the MIHF location such as mobile node or network elements. The information handled by the MIHF originates at different layers of protocol stack in mobile node or in network elements. The MIHF is composed of a set of handover enabling functions which provide service continuity while a MN traverses between heterogeneous wireless access link layer technologies. In the MIH Reference model [2]. The MIH user makes use of the MIHF function to support seamless handovers. Hence as shown in Figure 4-1, the load balancing module acts as the MIH user. The

following sub-sections describe in general the proposed framework of the load balancing algorithms that are running at the mobile node and the network entities.

4.3 Load Balancing Framework

This thesis proposes a load balancing framework that is necessary to provide the efficient load management strategies across different heterogeneous wireless networks which share/overlap their coverage areas. A common example of such coverage areas overlapping of heterogeneous wireless networks can be observed at the urban areas, especially in busy town centres, train stations and market places.

The load balancing framework comprises 3 main components:

- Load-aware RAT selection algorithm on the mobile node.

- Network load balancing algorithm on the radio access network

- Extensions to the MIH framework to support load balancing

These three components are described in detail in the following sub-sections.

4.3.1 Load-aware RAT Selection Algorithm Design

The load-aware RAT selection algorithm considers the network type, signal strength, data rate, user preference, network cost and network load as primary decision parameters for selecting a suitable access network technology. It tends to uniformly distribute the load among available heterogeneous wireless networks in order to maintain the load equilibrium on all networks which have overlapped coverage areas. When a mobile node is moving out of its serving network coverage and entering on other network

coverage, this algorithm will be applied in order to handover on-going sessions to the best available radio access network.

Assuming that all considered networks and mobile nodes support the IEEE 802.21 MIH standard, the proposed approach has taken advantage of MIH Media Independent Information Service (MIIS) specifically for the exchange of network load information besides exchanging other network related information like link type, link data rate, link capability, offered security and QoS and cost [2].

The flow chart shown in Figure 4-2 summarises the load aware RAT selection algorithm which runs at the mobile node. The mobile node compares the load conditions of the new available networks and the one to which it is currently connected. A list of networks is generated for those networks which are visible to the mobile node such that the received signal strength from those networks is higher than the minimum threshold. In the next step load, cost, offered QoS and other network related information of each network in the list is obtained from MIIS. This information, together with the received signal strength of each individual network will be used to finalise the order of the network list. The top element in the list will be selected to be the target network for handover. Three different algorithms, namely, baseline (least loaded), Fuzzy and Fuzzy Neural Network, are applied to generate the ordered network list. In case of baseline or non-cognitive algorithm the most preferred network from the list is the one with lowest load and highest offered data rate, whereas for the cognitive algorithms all the parameters such as signal strength, load, offered data rate of network, cost of network, coverage area of the network, speed of mobile node, user preferred network and

required data rate of mobile node are considered. The terms "HO" and "Conn" in the following flow chart represent handover and connection.
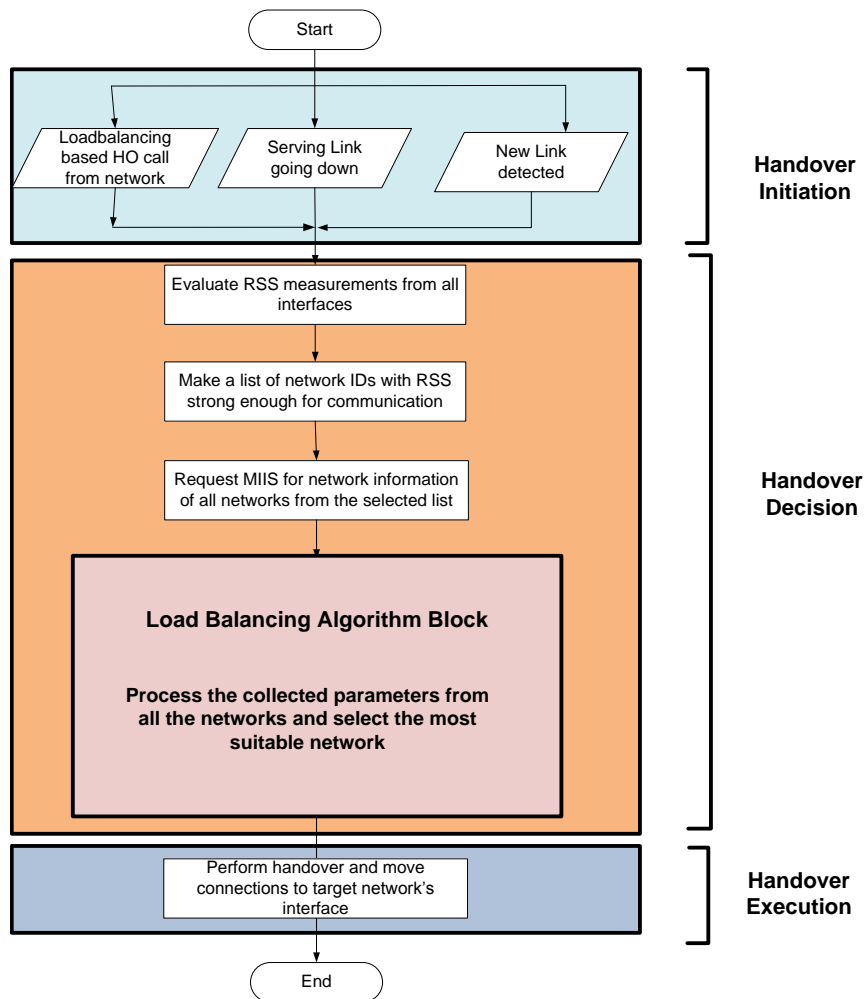
Figure 4-2: Load-aware RAT selection algorithm

The mobile node side algorithm can also be seen as different phases of a handover process: handover initiation, handover decision and handover execution. In the handover initiation phase, a mobile node detects new network or existing link getting weak. In this phase the process of load aware handover is initiated using MIH event signalling. The second phase is handover decision in which the mobile node compares all the considered parameters from available network and decides the target network for handover. The second phase also comprises of an important component

which is the load aware RAT selection algorithm. The last phase is the handover execution in which the mobile node performs the load aware handover and moves all the active connections to the target network.

4.3.2 Network Load Balancing Algorithm Design

While the RAT selection algorithm described in the previous sub section runs when a mobile node moves in or out of the coverage area of any access technology, it is important that the networks periodically monitor their own loads and the loads of the other networks in the common coverage area to make sure the loads can be uniformly balanced. In order to support this, a load balancing algorithm is proposed that runs on the access network entities such as RNC, BS, RAN or AP. The flowchart shown in Figure 4-3 represents the network side load balancing algorithm. In network entities the load balancing algorithm continuously keeps on updating the MIIS about its current load status and receives load information of its neighbouring networks. This process of updating the MIIS with network information by any particular network, have been discussed in references [104, 105, 106 and 107]. This updating process runs on every time when a new connection starts or ends in the network or periodically. The network entity also requests for the neighbouring networks load information from the MIIS when it sends out the local information. Upon receiving the neighbouring networks load information the network entity makes and filters the list to keep only those networks which are providing coverage to locally registered mobile nodes. Similar to Figure 4-2, the terms "HO" and "Conn" in the following flow chart represent handover and connection.
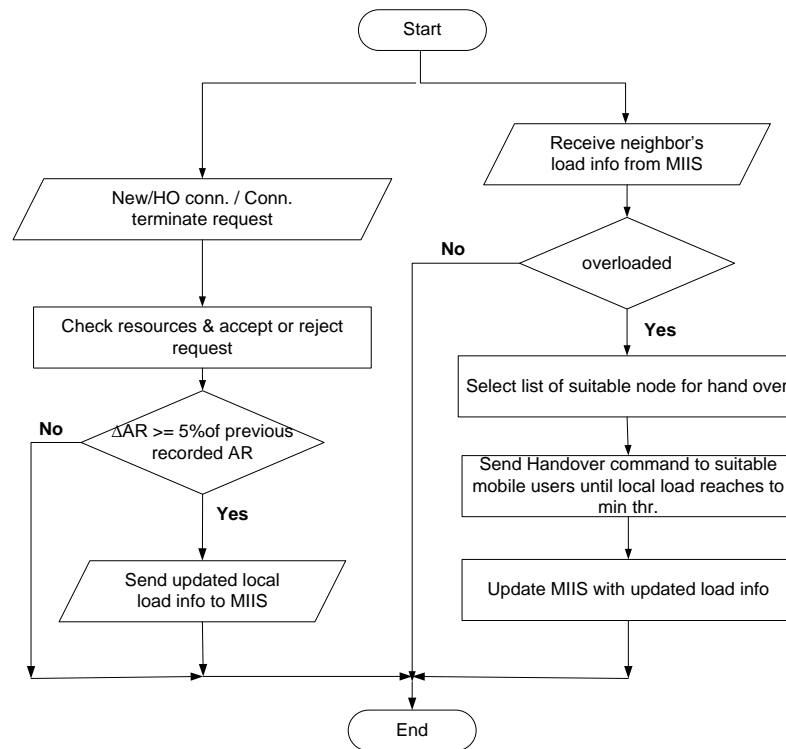
Figure 4-3: Network Load balancing algorithm

The most loaded network entity starts moving out the suitable mobile users to appropriate networks. Here the suitable nodes are those mobile nodes which can see the coverage area of other neighbouring networks apart from the serving network and the networks which have same of high load as compared to the serving network. If the load variation between the current network and neighbouring networks is higher than the threshold of 50% free resources margin. For example the percentage of free resources in remote (neighbour) network is greater than or equal to the double of available resources percentage at current (local) network. In other words, the overloaded network find itself as overloaded when any of its neighbouring networks have free resources percentage at least twice the local free resource percentage. The load balancing algorithm keeps on migrating out the suitable mobile nodes from overloaded network to the least loaded networks until the load in overloaded network becomes equal to or lesser

than the average load in all the neighbouring networks of overloaded network, or all the suitable mobile nodes have been moved out.

The selection of suitable nodes in the overloaded network can be easily done by keeping record for list of nodes in the network which can see other networks using the IEEE 802.21 MIH framework. In IEEE 802.21 MIH each mobile node sends list of available networks to the serving network upon detecting a new network or upon detecting the weakening of signal strength from serving network. The process of calculating the free resource at the network and process of suitable node selection on the network side is represented with the help of analytical equations in the following sections.

4.3.3 Extensions in IEEE 802.21 MIH Protocol

To support load-aware RAT selection and network load balancing across heterogeneous networks three major extensions have to be made to the standard IEEE 802.21 MIH specifications. The new additions are as follows:

i)   Extending the IEEE 802.21 to include the satellite networks. The IEEE 802.21 standard [2] considers UMTS, WiMax and WLAN but no satellite network is supported. The inclusion of satellite network requires the mapping of primitives between satellite SAPs at the MAC layer and the MIH link SAPs which is briefly described in the previous chapter.

ii) Introduction of a new primitive for sending load information from the network entity i.e.BS, AP, RAN or RNC to the common MIIS. The new primitive *MIH_Set_Information.indication* is introduced in the IEEE 802.21 for sending the information like load to the MIIS. The contents of the newly introduced primitive *MIH_Set_Information.indication* are shown in the Table 4- 1 given below.

Table 4- 1: MIH_Set_Information.indication parameters

Parameters	Description
DestinationIdentifier	Destination MIHF
InfoSetBinaryData	(Op.) TLV query.
InfoSetRDFData	(Op.) RDF query.
InfoSetRDFSchema	RDF Schema URL query (Required only when value in this field is "True")
InfoSetRDFSchemaURL	(Op.) RDF Schema query.
SetNetworkType	The type of network, who is updating its information in the MIIS

In the *MIH_Set_Information.indication* primitive only one parameter should be specified just like in case of *MIH_Get_Information*, where only one parameter is specified from a list of available primitives such as InfoSetBinaryDataList, InfoSetRDFDataList, InfoSetRDFSchemaList, InfoSetRDFSchemaURLList. The *MIH_Set_Information.indication* message is used by the MIHF located at the networks element to send their network information towards the MIIS so that their information in the MIIS database can be updated. This process of updating of the network information occurs every time when there is a 5% or greater change in the available resource as compared to the amount of resources available when the previous time this message was sent to the MIIS. All the networks keep track of the amount of available resource at time when the *MIH_Set_Update.indication* message was sent to the MIIS.

4.3.4 Analytical components for load balancing

The handover strategy used for the load balancing purpose is "network assisted mobile controlled" handover. Most of the information is collected at

network side and forwarded to the mobile node. Then the load aware RAT selection algorithms running on the mobile node utilize this information in decision making. Therefore most of the processing for information gathering is performed at network side. The analytical representations of some important procedures performed at network side are given below:

4.3.4.1 Resource Utilization

The Resource Utilization (RU) of a network is evaluated at the network side to calculate the available resources on that network. The element a_{xy} of an association matrix A is used to describe the association status between MN_y and BS_x. If the MN 'y' is associated with BS 'x' then a_{xy} = 1, else a_{xy} = 0. Suppose there are m mobile nodes in the system registered with the different base stations. A base station, x, is denoted by BS_x then the equation for the resource utilization in each base station or the network entity can be represented as RU which is shown in Eq. [1].

$$Axy = \begin{matrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1y} \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots & a_{2y} \\ a_{31} & a_{32} & a_{33} & a_{34} & \cdots & a_{3y} \\ a_{41} & a_{42} & a_{43} & a_{44} & \cdots & a_{4y} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{x1} & a_{x2} & 3 & a_{x4} & \cdots & a_{xy} \end{matrix}$$

$$RU_x = \frac{\sum_{y=1}^{m}(a_{xy} * Throughput_y)}{\text{Bandwidth}_x} \qquad \text{Eq. [1]}$$

In Eq. [1] the $Throughput_y$ denotes the throughput of MN 'y' and $Bandwidth_x$ the bandwidth of network 'x'. The percentage of available resources (RA_x) can be derived as:

$$RA_x = (1 - RU_x) * 100$$

Eq.[2]

4.3.4.2 Selection of neighbouring network with common coverage area

The network element such as BS or RNC will periodically request for the load status of its neighbouring networks from the MIIS. The process of neighbouring networks selection is carried out using with the help of a simple distance formula, as shown below in Eq. [3].

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad \text{Eq.[3]}$$

The local network sends out its location information in the *MIH_Get_Information.Request* message to the MIIS in order to get the load status of its neighbouring networks. For simplicity it is assumed that the information it sends out contains the location coordinates of the network element (such as BS) in the *'x-y'* plane, which is network element location and the range of network coverage which is shown as radius in Figure 4-4. The MIIS upon receiving the location information of the network will look for the networks having overlapping coverage area with the requesting network. For this the MIIS checks all networks in its database one by one, which fulfil the condition given in the following Eq. [4].

$$d{<}R \qquad \text{Eq. [4]}$$

(where, R = r1+r2)

In Eq. [4], the symbol *'d'* is the distance between two points P2 and P1 as shown in Figure 4-4. P1 and P2 are the centres of the coverage areas for requesting network and the neighbouring candidate network. *'R'* is the sum of the radius of the candidate and the requesting networks; it is the range of the network coverage. MIIS will reply to the requesting network with network information of all those neighbouring networks for which Eq. [4] is satisfied.

4.3.4.2.1 Neighbour discovery example

Figure 4-4 shows an example of three networks *N1, N2* and *N3* with different coverage and overlapping areas. The purpose of this diagram is to validate the Eq. [4] with the help of an example. It can be seen in Figure 4-4 that value of *d* is less than *R* for networks with overlapping or common coverage areas and for networks with no common coverage areas the value of d is always greater than or equal to *R*.
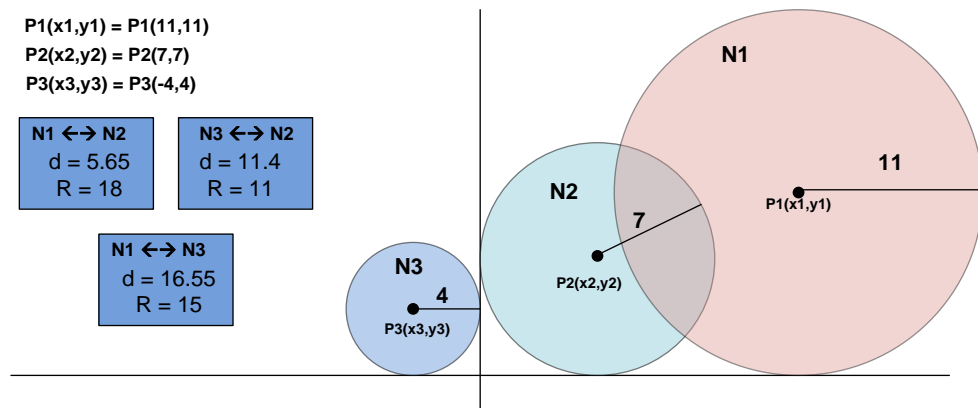

P1(x1,y1) = P1(11,11)
P2(x2,y2) = P2(7,7)
P3(x3,y3) = P3(-4,4)

N1 ←→ N2
d = 5.65
R = 18

N3 ←→ N2
d = 11.4
R = 11

N1 ←→ N3
d = 16.55
R = 15

Figure 4-4: Overlapping detection in network coverage areas

4.3.4.3 Suitable node selection

The overloaded network selects the least loaded network from its neighbouring networks and searches from its locally registered mobile nodes locations if there are any mobile nodes in common coverage areas. If there are some mobile nodes in the common coverage area, then the overloaded network will handover those mobile nodes to that least loaded network until the free resource in the overloaded network becomes less than or equal to the average free resource in all the networks. If no more mobile nodes left in current network which are located in the common coverage area of overloaded and least loaded network. When a network receives reply, from the MIIS, it compares the load information of its neighbouring networks with

its local load status. If the local load satisfies the condition to trigger load balancing, the network looks for suitable mobile nodes which can be moved to other neighbouring networks. The decision of suitable mobile nodes selection for load balancing based handover is achieved by distance equation shown in Eq. [3]. In this case the current position of the mobile node, the centre of the neighbouring network and the radius of the neighbouring network coverage area is used. The network will search for the mobile nodes registered locally which are in the common coverage areas of local network and the 2nd least loaded network from the list it got from MIIS and so on until the load in the current/local network reaches less than or equal to the average load in all neighbouring networks.

4.4 Detailed handover procedures

This section explains the overall handover procedures across the different radio access technologies to achieve efficient load balancing in detail. The handover procedure can be divided into three phases namely handover initiation, handover decision and handover execution. The first two phases (handover initiation and decision) can be categorised into Mobile Controlled Handover (MCHO), Network Controlled Handover (NCHO), Mobile Assisted Handover (MAHO) or Network Assisted Handover (NAHO). There can be other hybrid schemes evolved from these basic schemes such as mobile assisted network controlled and network assisted mobile controlled handovers. In the final handover execution phase, the connections of the mobile node are released from the serving network and seamlessly moved to the target network. The signalling exchange procedure between the mobile

node and network required for the handover execution phase can be of two types such as backward and forward. While the backward handover scheme utilizes serving network link for signalling exchange, the forward handover scheme establishes and uses new signalling link with target network [108,109].

Handover can also be classified into three categories namely; hard handover, soft handover and softer handover. In hard handover, a mobile node disconnects itself from the current serving network before connecting to the target network leading to the break-before-make handover scenario. In contrast, soft handover is a make-before-break handover where a mobile node connects to target network before disconnecting itself from current serving network. The backward handover which is initiated via the serving base station is the soft handover, whereas the forward handover which is initiated via the target base station is hard handover. In case of softer handover, the mobile node stays connected to the serving network but retunes its communication frequency or communication channel. In UMTS softer handover, the mobile node moves between different sectors of the same base station. Softer handover is also known as intra cell handover. Handover can also be categorized as horizontal and vertical handovers, in which horizontal handover represents the process of migrating mobile node from one network to another provided that both serving and target networks are of the same type, whereas in vertical handover the target and serving networks are of different types [110, 72].

The handover scenarios considered in this section cover the seamless vertical handover procedures between satellite (BGAN) network, UMTS,

WiMax and WLAN networks. It is assumed in the following handover procedures that the mobile node supports the multi-interfaces and can therefore utilize the networks such as satellite, UMTS, WiMax and WLAN upon detection/availability. It is also assumed that the mobile node establishes and maintains a connection with the remote source node which is located beyond the core network [72, 73,111].

4.4.1 General mobile initiated handover

Figure 4-5 represents the general handover scenario's message sequence chart in which a mobile node handovers from a serving network to a target network after establishing a connection with the target network.
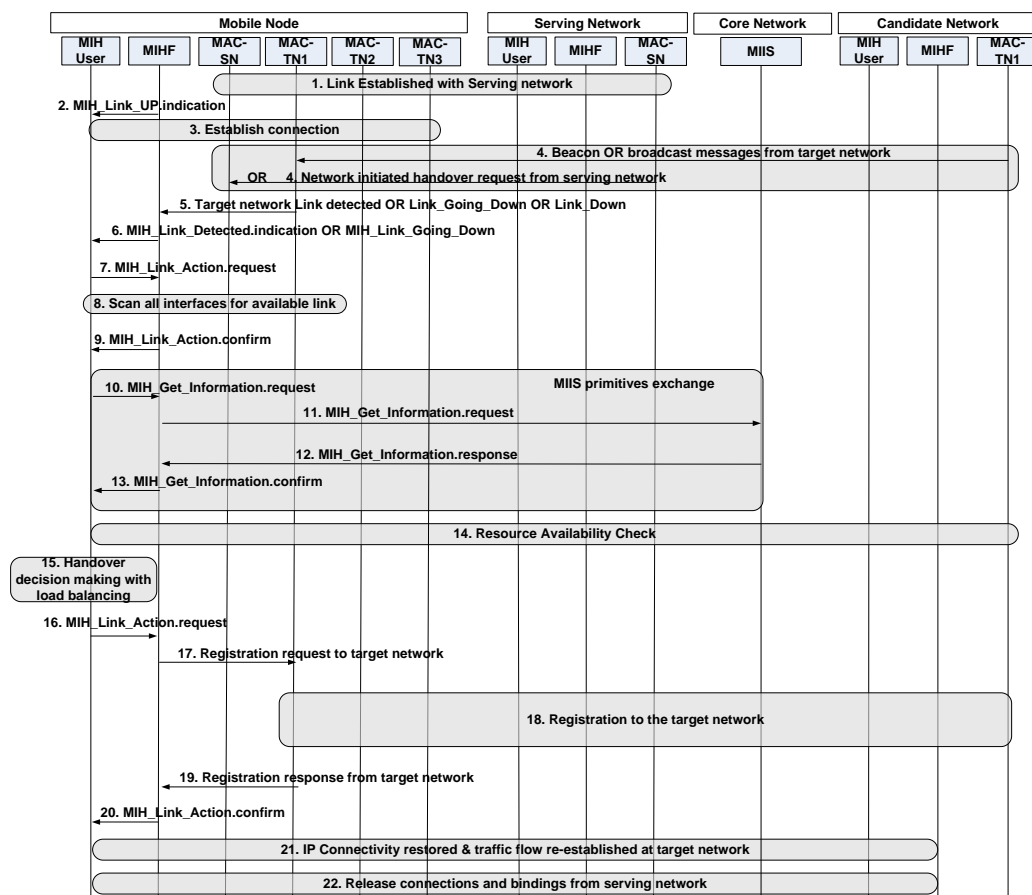

Figure 4-5: General Handover scenario (Mobile initiated)

In Figure 4-5 MAC-SN represents the MAC layer for serving networks, i.e. MAC-TN1, MAC-TN2 and MAC-TN3 stands for MAC layer of target network 1, target network 2 and target network 3 respectively. First the mobile node registers and establishes a connection with the serving network (shown in step 1,2 and 3 of Figure 4-5) a handover may be triggered if a new network is detected or it may initiated by the serving network in order to balance the loads.

In steps 5 and 6, the MIH user is informed about the handover trigger which is received in step 4. The sequence of messages from step 7 to step 9 represents the procedure of scanning all the interface of mobile node for the available networks. In the messages from step 10 to step 13 the MIH user at mobile node extracts the information about the detected networks which were detected in scanning process. Step 14 shows the resource availability process in the detected networks [2]. The target network selection decision is made in step 15 by utilizing the load aware handover algorithm. The mobile node registers itself with the selected network by using its corresponding interface in the sequence of steps from step 16 to step 20. Step 21 shows the connections handover from the serving network to the selected target network. All the bindings with the old serving network are released in step 22.

4.4.2 General network initiated handover

For load balancing from the network side, the handover procedure is triggered from the network side which is depicted in the message sequence chart shown in Figure 4-6. This message sequence chart is representing the general diagram for network initiated handover for load balancing purpose.
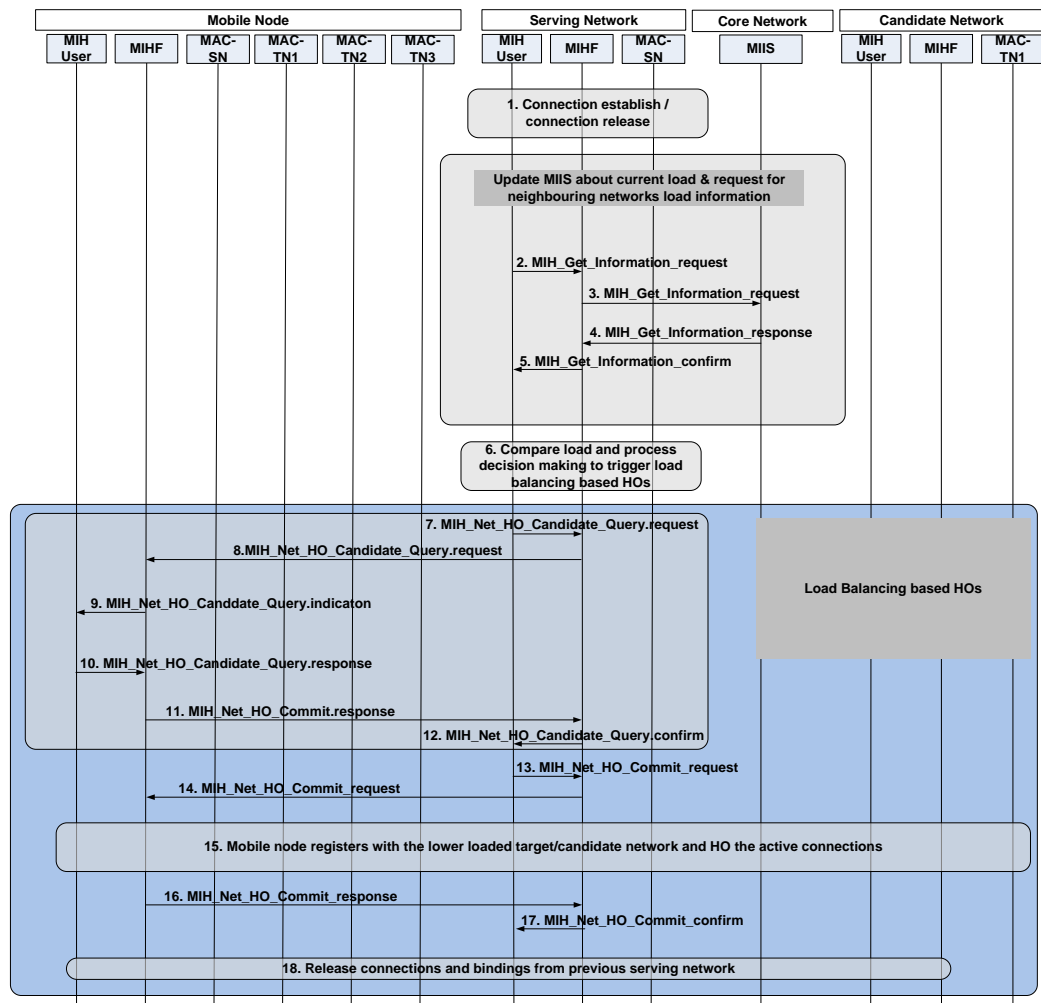
Figure 4-6: General scenario for Handover (Network initiated)

The step 1 in Figure 4-6 represents the situation where the network updates MIIS about its load status after change in available resource due to an active connection has been released upon completion or handover, or a new connection have been established. The network also requests for the load information of the neighbouring networks from MIIS at the same time, which is depicted in the sequence of messages until step 5. In step 6, the serving network analyses the load information which is received from the MIIS and decides whether load balancing HOs should be executed or not. Once the decision is made to trigger the load balancing based HOs from network side, the serving network collects the available networks from the suitable nodes

which are shown in sequence of steps from step 7 to step 12. The network then instructs the selected mobile nodes to initiate HO process to their candidate networks having lower load. The sequence of steps from step 13 to step 17 represents the network indicating mobile node to handover. Step 18 represents the final operation where mobile node releases bindings with serving network after registering with target candidate networks and establishing traffic flows.

4.4.3 Handovers between UMTS and WiMax

This section describes the steps involved in the handovers between UMTS and WiMax access technologies. The message sequence charts for the two scenarios such as handover from UMTS to WiMax and handover from WiMax to the UMTS are described in detail as follows:

4.4.4 Handover from UMTS to WiMax

Figure 4-7 shows the procedure and the SAP primitives involves in the handover from UMTS to WiMax. The step 1 in Figure 4-7 informs the MIHF that the mobile node is registered with the UMTS network. Step 2 informs the MIH User about the activation of link on mobile node's UMTS interface.

In step 3, a connection is established between the mobile node and the source using the UMTS network. Step 4 shows WiMax interface receives broadcast messages from WiMax BS. Step 5 signals the MIHF in the mobile node about WiMax network detection. Step 6 informs MIH User about WiMax link detection. In sequence of steps from 7 to 10 the mobile node acquire the neighbouring networks information from MIIS (which is located in the core network) using a set of MIIS primitives, and step 11 checks the availability of

required resources in WiMax network. Step 12 decides whether or not to perform handover to WiMax. In steps 13 to 22, the mobile node registers itself on WiMax. Step 23 is to establish a connection between mobile node and source using WiMax network. In step 24, mobile node releases its connections to the UMTS network.
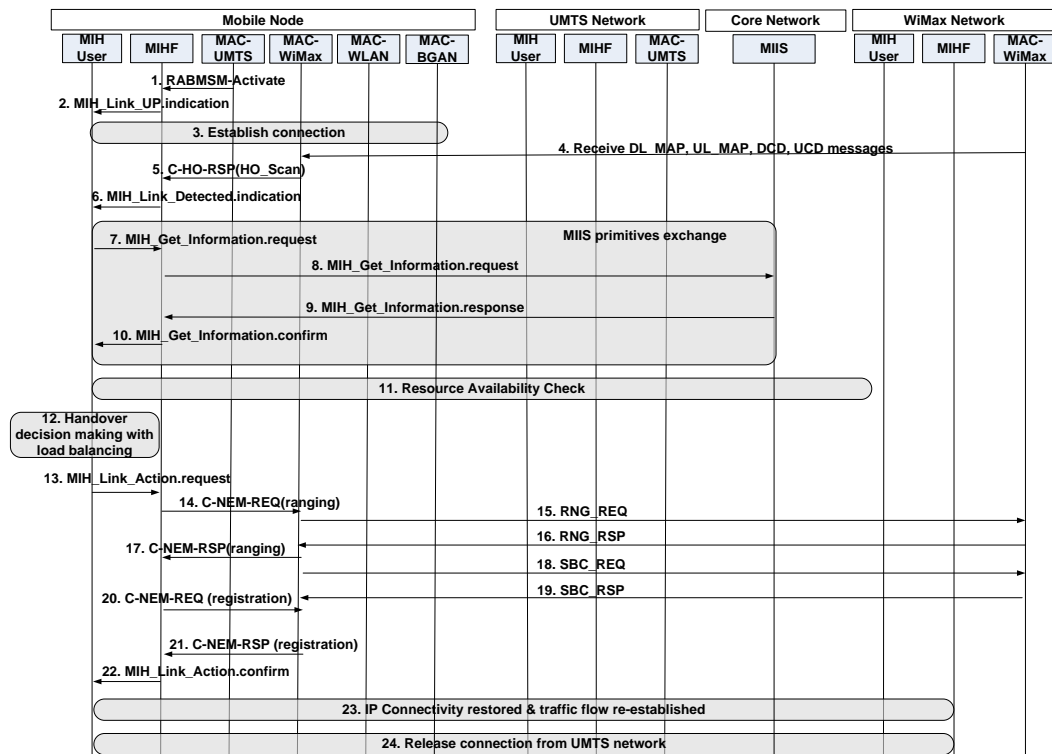

Figure 4-7: Handover from UMTS to WiMax

4.4.4.1 Handover from WiMax to UMTS

The message sequence chart shown in Figure 4-8 represents the handover procedure when mobile user moves from a WiMax-UMTS common coverage area to an area covered by UMTS only or when the network initiated handover is triggered for the load balancing purpose.
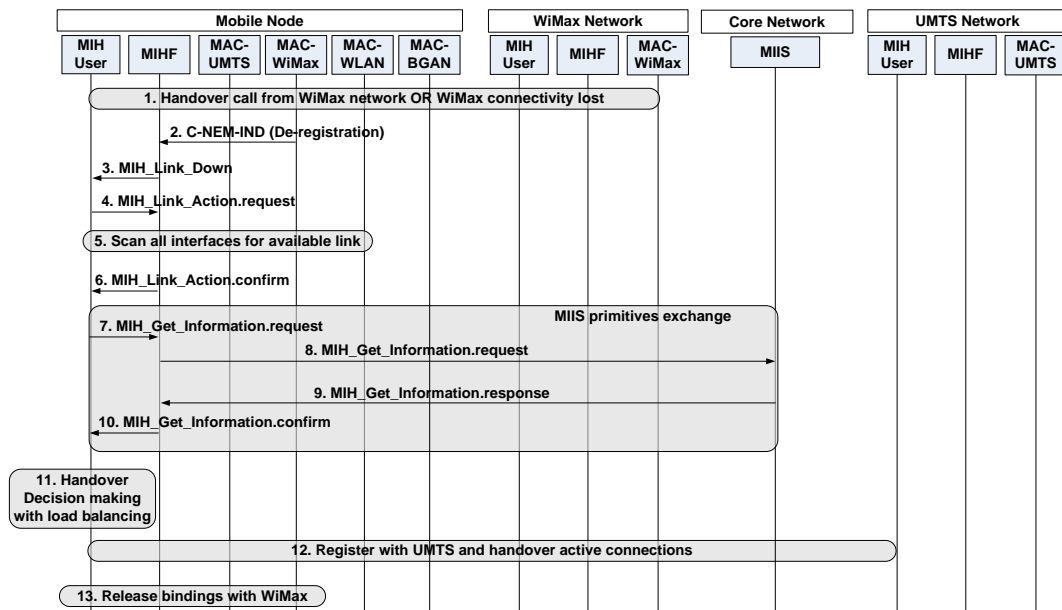
Figure 4-8: Handover from WiMax to UMTS

Here mobile node handovers to the UMTS from the WiMax network. Step 1 represents that the mobile node has lost the WiMax connectivity or it has received the handover call from the network for the load balancing purpose. In step 2 the WiMax MAC sends link down equivalent primitive to MIHF, this triggers the *MIH_Link_Down* primitive from the MIHF to MIH User in step 3. In messages from step 4 to step 6 scanning for the other available links performed. Sequence of steps from step 7 to step 10 shows that mobile node acquire neighbouring networks information from MIIS using MIIS primitives. Step 11 shows the decision making process for selecting the candidate network for handover, UMTS is selected as it is the only available network. Step 12 represents the mobile node's handover to UMTS. In step 13 all bindings with WiMax are released.

4.4.5 Handover between WiMax and WLAN

This section explains the steps involved in the message sequence charts for the handover between WiMax and the WLAN access technologies.

4.4.5.1 Handover from WiMax to WLAN

Figure 4-9 shows the SAP primitives used in the handover procedure from WiMax to WLAN. As represented by Figure 4-9, the 802.11 MAC layer in the mobile node, after detecting and registering with WLAN network, it sends *MLME-LinkUp.indication* message to the MIHF.
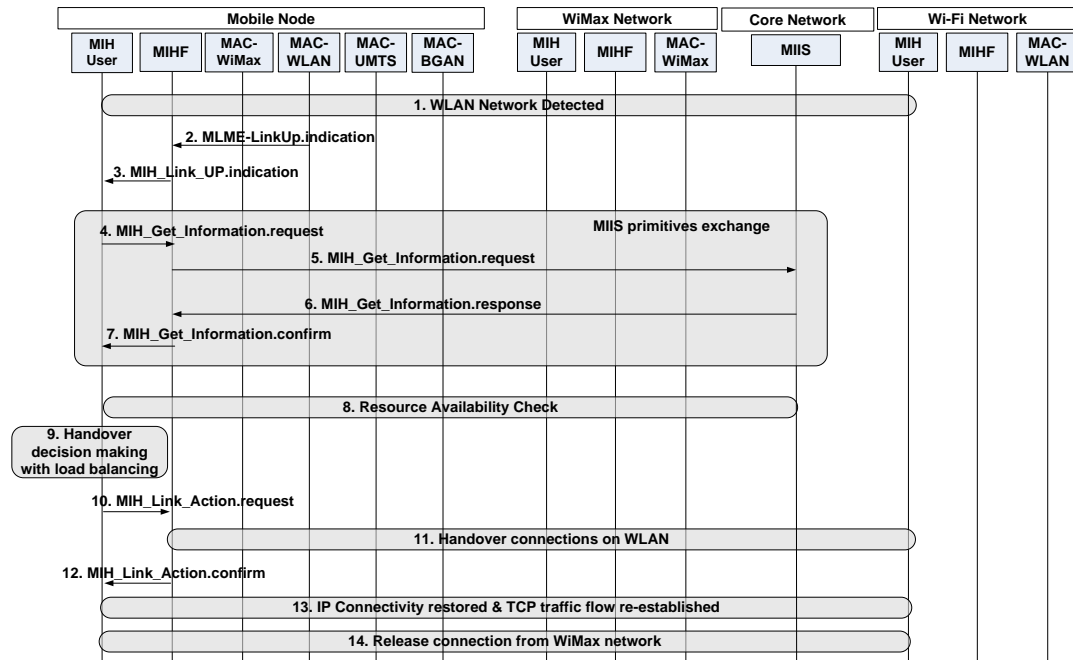

Figure 4-9: Handover from WiMax to WLAN

In step 3, MIHF sends *MIH_Link_UP.indication* to MIH User. A set of messages from step 4 to 7 acquire the neighbouring networks information. Step 8 checks for the required resources in WLAN for handover. The MIH User decides whether to perform handover or not in step 9. Steps 10 to 12 show the handing over of the connections to the WLAN network. Finally, step 13 and step 14 make sure that traffic flow has been re-established between the mobile node and source and then releases bindings with the WiMax network.

4.4.5.2 Handover from WLAN to WiMax

Figure 4-10 shows the handover procedure when mobile user moves away from the WLAN coverage area and enters a WiMax coverage area. The first step in Figure 4-10 is the message *MLME_MREPORT.indication* from MAC WLAN to MIHF. This is the periodic message which carries parameters of link. In step 2 the MIH User is being updated with link parameters report. Step 3 shows the message *link-Going_down* from WLAN MAC to MIHF, which represents that mobile node, is gradually losing the connectivity with WLAN.
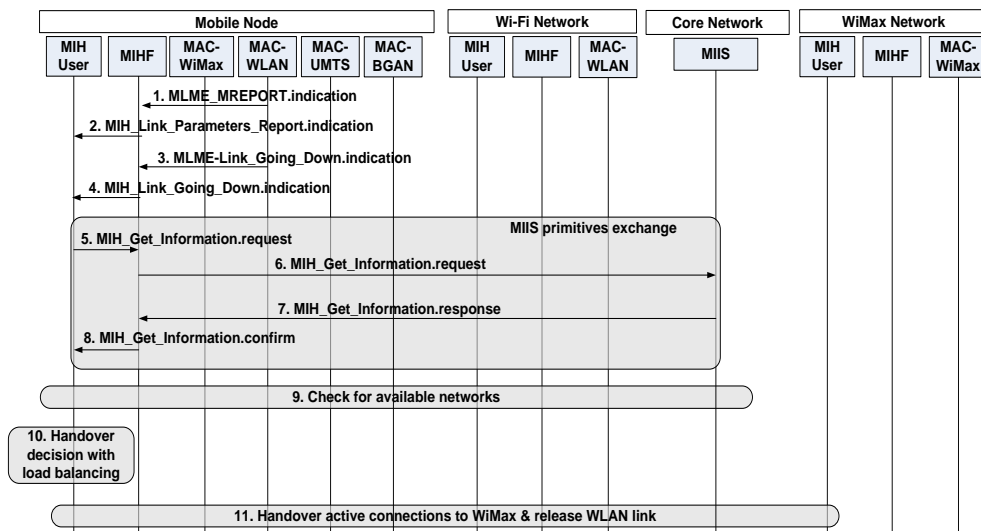

Figure 4-10: Handover from WLAN to WiMax

Step 4 informs the MIH User about link going down event. From step 5 to step 8 the messages are used to acquire neighbouring networks information from MIIS. Step 9 shown as bubble represents the process of scanning on all interfaces supported by mobile node. Step 10 and step 11 are for selecting the WiMax network and handover all active connections to WiMax.

4.4.6 Handover between UMTS and Satellite

This section explains the step involved in handover between UMTS and satellite with the help of message sequence charts.

4.4.6.1 Handover from UMTS to Satellite

Figure 4-11 represents the message sequence chart which shows the sequence of message for the handover procedure when the mobile node handovers from the UMTS network to the satellite (BGAN) network.
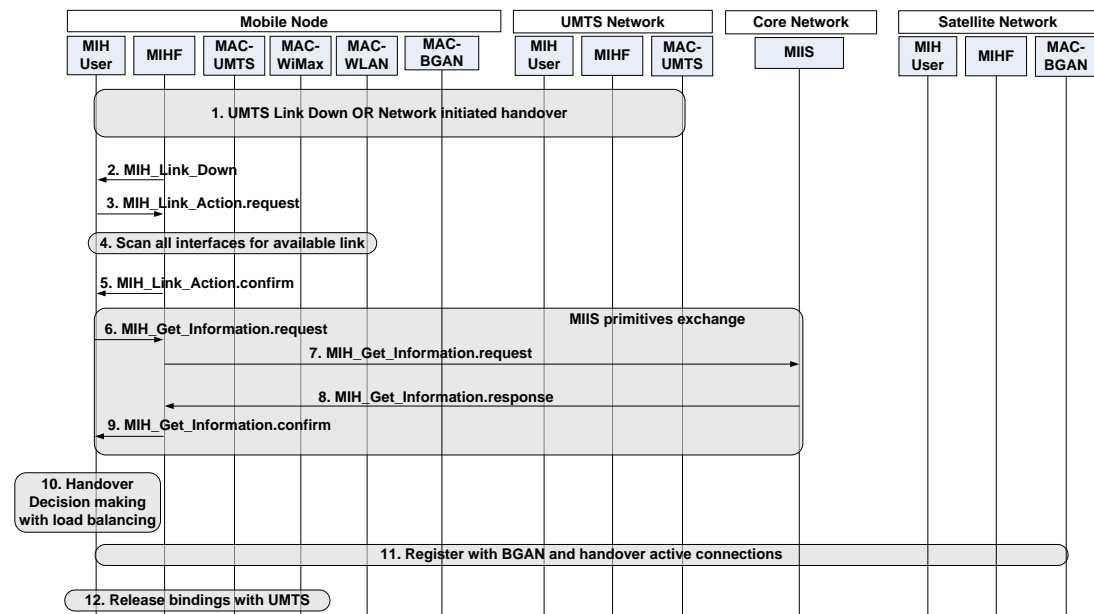

Figure 4-11: Handover from UMTS to Satellite (BGAN)

The step 1 shows that the mobile node has lost UMTS connectivity or the UMTS network initiated the mobile node handover process for the load balancing purpose, which triggers the *MIH_Link_Down* message from MIHF towards the MIH user in the step 2. The messages from step 3 to step 5 shows the phenomena where MIH user send the *MIH_Link_Action* command to all MAC layer interfaces of the mobile node via MIHF to perform scanning for the availability of the networks. Once the scanning is performed the MIH user gets the information about all the available networks in step 5. The

sequence of messages from step 6 to step 9 carries the operation for information exchange between MIH user and the MIIS which brings the information such as network load, cost and offered QoS of different networks. In the step 10 the handover decision is made using the load aware algorithms. Once the decision is made in step 10 to move the mobile node to the satellite network the mobile node registers itself to the BGAN network and move the active connections to the BGAN network as represented by the bubble in step 11. Step 12 shows that all the bindings with UMTS are released once the connections have been re-established over BGAN network.

4.4.6.2 Handover from Satellite to UMTS

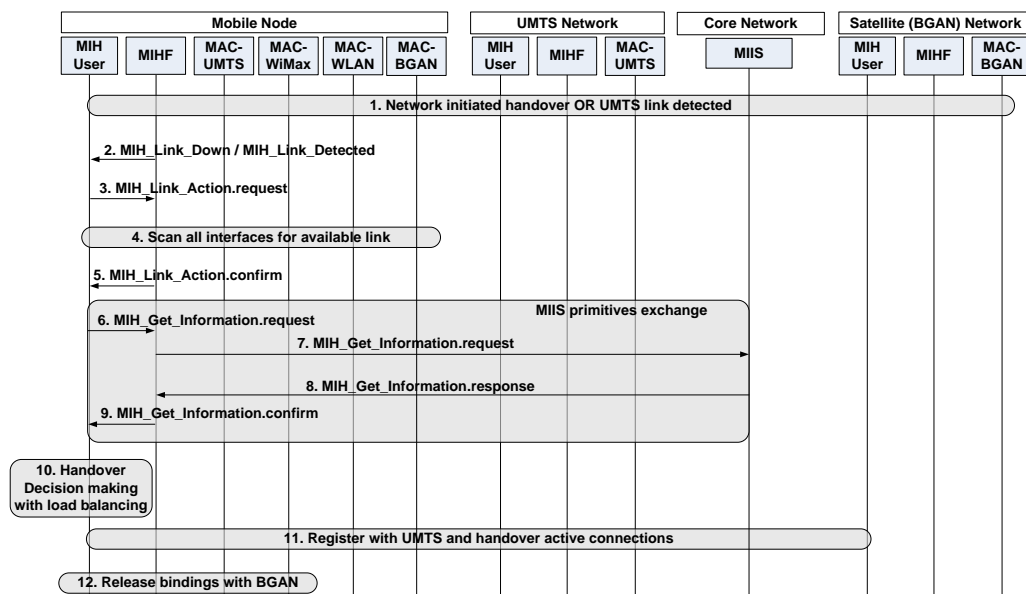Figure 4-12 shows the scenario where mobile node handovers from satellite network (BGAN) to UMTS.

Figure 4-12: Handover from Satellite (BGAN) to UMTS

It is assumed that the handover is triggered from the BGAN for load balancing purpose or the mobile node entered the UMTS coverage which is less loaded or the connectivity with BGAN is lost. The step 1 shows that the BGAN link gone down due to one to the above mentioned reasons. Step 2 in Figure 4-12 shows the message from MIHF to MIH user which informs MIH user about the BGAN link gone down. From step 3 to step 5 the MIH user in the mobile node preforms the scanning operation at all the interfaces. In step 6 to step 9 the MIH user in the mobile node extracts the information about the detected network from the MIIS. In step 10 the decision for the handover is made by utilizing the load aware RAT selection algorithm. In step 11 the mobile node register with UMTS network which is selected at step 10. Once the mobile node is registered with UMTS, it handovers the active connections to the UMTS from satellite network. In step 12 the bindings with the BGAN are released.

4.4.7 Handover between Satellite and WiMax

The details of the steps involved in the handover process between satellite and the WiMax are described in this section.

4.4.7.1 Handover from WiMax to Satellite

Figure 4-13 represents the scenario where mobile node performs the handover operation from WiMax to the satellite network (BGAN). In the step 1 of Figure 4-13 it is assumed that mobile node has lost the connectivity with WiMax network or the WiMax network has sent the De-Registeration message to the mobile node for the load balancing purpose. Step 2 shows

that MAC layer of WiMax at mobile node forwarded the De-Registration message to the MIHF which triggers the *MIH_Link_Down* in step 3 from MIHF to MIH user.
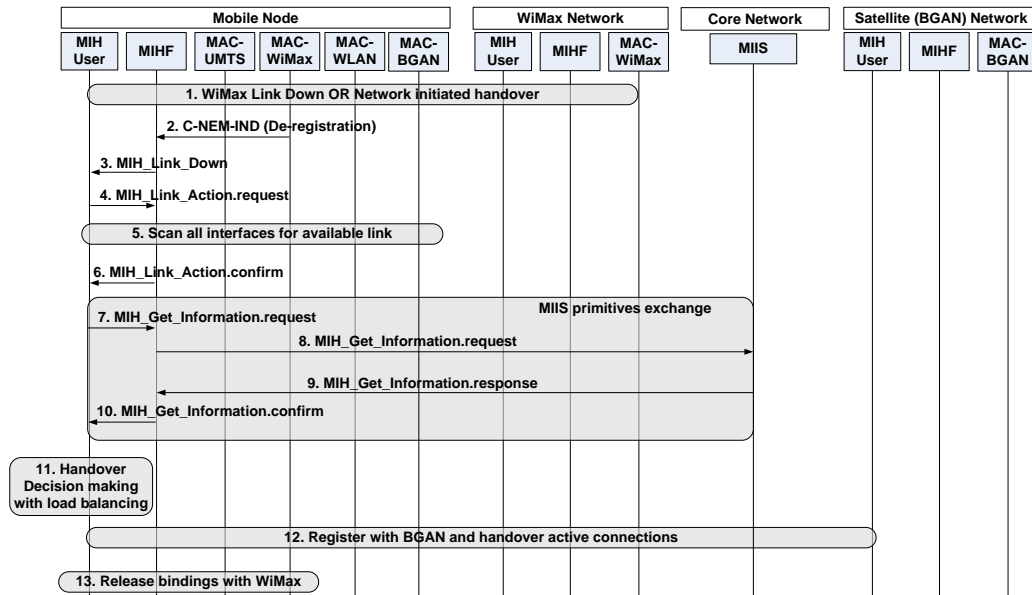

Figure 4-13: Handover from WiMax to Satellite (BGAN)

The sequence messages from step 4 to the step 6 in Figure 4-13 represent the process of scanning on all the interface of mobile node and sending scanning reports towards the MIH user. The steps from 7 to 10 are used to gather information from the MIIS about the networks which were detected during the scanning operation at mobile node. The handover decision is made in step 11 and the target network is selected for handover operation. In step 12 the mobile node register itself on BGAN network and handover the active connections to the BGAN. All the bindings with the WiMax network are released in step 13.

4.4.7.2 Handover from Satellite to WiMax

Figure 4-14 shows the message sequence chart for the handover process of a mobile node from the satellite network to the WiMax network. It is assumed

that the mobile node enters the coverage area of the WiMax network which is under loaded or the mobile node lost connectivity with the currently connected satellite network (BGAN) or the BGAN network initiated the mobile node handover for the load balancing purpose. The step 1 in Figure 4-14 shows that mobile node lost the connectivity with BGAN network.
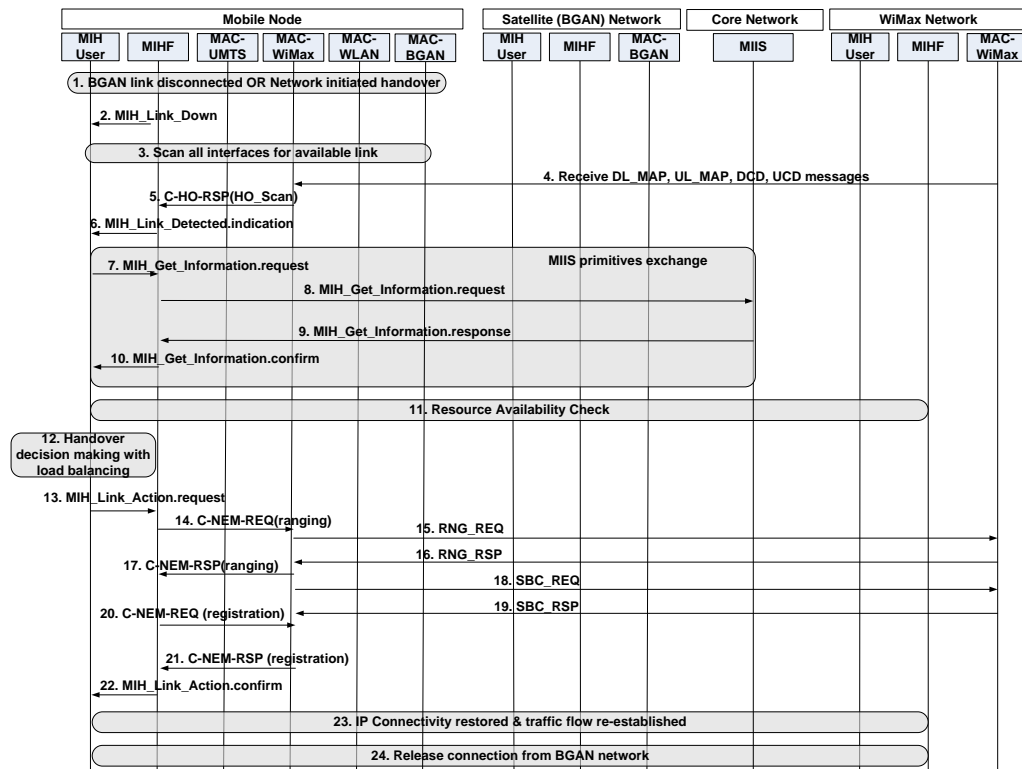

Figure 4-14: Handover from Satellite (BGAN) to WiMax

Step 2 represents the message generated from MIHF towards MIH user upon detecting the BGAN connectivity lost at MIHF. Step 3 shows the process where MIH user instructs all the interface of mobile node to perform scanning. In step 4 the WiMax interface receives the beacon messages from the WiMax network. Step 5 and 6 forward the received network information of the WiMax network to the MIH user from the MAC layer of WiMax in mobile node via MIHF. From step 7 to the step 10 the MIH user in the mobile node extracts the information of the detected networks including WiMax from the

MIIS. Once all the information is gathered at the MIH user the decision for the handover is made which is shown in the step 12. The sequence of steps from step 13 to step 22 represents the registration of mobile node at WiMax network. The bubble at step 23 shows the process of re-establishing the active connection over WiMax. The bindings with BGAN are released at step 24.

4.5 Proposed Load aware RAT selection Algorithms

To achieve efficient load balancing across different radio access networks, novel algorithms have been proposed as part of the proposed overall load balancing framework, The load balancing strategies proposed in this research are novel load aware RAT selection techniques which uniformly distribute the network load between co-located heterogeneous wireless networks. It utilizes parameters collected using MIH to seamlessly handover mobile users between heterogeneous wireless networks for load balancing purpose. The advantage of this proposed algorithms is that it minimizes the call blocking and dropping probabilities, number of packet drop/lost and delays during the handover process and enhances the network utilization by continuously balancing the load in co-located networks The following three algorithms are proposed in this thesis for performing load balancing in heterogeneous wireless networks during RAT selection:

- **Baseline algorithm:** is a dedicated load balancing algorithm with rather simple decision making rules instead of involving complex computational overhead.

- **Fuzzy algorithm**: Fuzzy logic based load balancing algorithm for the RAT selection in heterogeneous wireless networks. This algorithm utilizes the fuzzy logic controller to obtain the most suitable result by efficiently considering all the parameters.

- **Neural-fuzzy algorithm**: The benefits of fuzzy and neural network algorithms are combined together to increase the efficiency of load balancing in fuzzy neural based algorithm for heterogeneous wireless networks RAT selection.

All three load balancing algorithms are deployed on the MIH based network architecture for efficient load balancing in heterogeneous wireless networks. These proposed algorithms are described further in detail in the following sub sections.

4.5.1 Baseline Algorithm for Load Aware RAT selection

This is the simple load aware based RAT selection technique which does not involve overhead of complex decision making and heavy computation. Instead this approach follows a set of simple rules which are easy to implement and faster to execute as it does not involves the complex calculation like fuzzy logic technique. In this proposed baseline load balancing algorithm the RAT selection strategy for heterogeneous wireless networks takes the following input metrics: signal strength, data rate, network coverage, load of network, network cost, required QoS of arriving call, user preferred network and speed of mobile node. This algorithm addresses the load balancing problem in RAT selection for handover calls as well as for the new arriving calls. The baseline load balancing algorithm falls into the

category of "Load based RAT selection" in RAT selection algorithms categories which are described earlier in previous chapter.

The proposed baseline load balancing technique runs on mobile node in the MIHF user of the protocol stack [72, 73, 111]. The flow chart represented in Figure 4-15 represents the proposed algorithm which runs at mobile node. At the mobile node, the proposed technique first makes a list of available network IDs which are visible to mobile node such that received signal strength from those networks is higher than the minimum threshold required for communication. In the next step, the network coverage, offered data rate, cost and load values of each network in the list are obtained from MIIS using *MIH_Get_Information.request* primitive of IEEE 802.21 MIH. The mobile node speed, preferred network and required QoS are also considered for further processing. Then in the following step, it compares the data rate offered by each network in the list with the required data rate at the mobile node and list of networks is sorted as the most suitable on the top. The entries of networks which do not fulfil required QoS level are deleted from the list. The list of suitable networks is again checked for coverage and the speed of mobile node, and is sorted in such a way that most suitable network is on top of the list. Now the list is again sorted for user preference considering the network cost. This step would remove satellite network entries from the list if user does not want to pay for expensive satellite networks. In the last step of processing the parameters the list of network is sorted in such a way that least loaded network stays on top and most load network goes at the bottom in the list of suitable networks. The most preferred network from the list is the one with lowest load and highest offered

data rate and coverage and obviously the network cost also effects the decision. The following Figure 4-15 is representing the steps involved in the baseline load balancing algorithm for the RAT selection in heterogeneous wireless network.
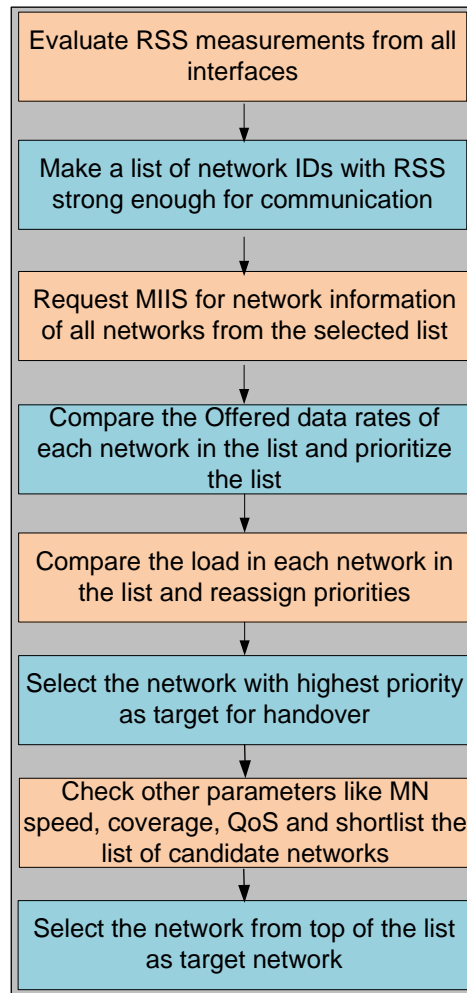

Figure 4-15: Baseline load balancing algorithm in mobile node

The performance of baseline algorithm has been analysed in reference [72, 73 and 111] which clearly show that the baseline load balancing algorithm improves the radio resource utilization and throughput considerably but results in a slightly higher overall end-to-end delays. This slight raise in the end-to-end delay is acceptable for most delay tolerant applications.

4.5.2 Fuzzy logic based load aware RAT selection algorithm

Fuzzy logic techniques provide an efficient way in decision making where multiple parameters play their role in obtaining the final decision. Fuzzy logic based algorithms have been used in various decision making systems like Call admission control, signal processing, data analysis, etc. The main advantage of fuzzy logic technique is that it can use the expert knowledge as fuzzy rules. This makes it easier to successfully automate the system for already known contexts. The fuzzy reasoning aims at modelling of reasoning schemes based on imprecise or uncertain, unlike the reasoning based on classical logic which requires exact information [112, 113].
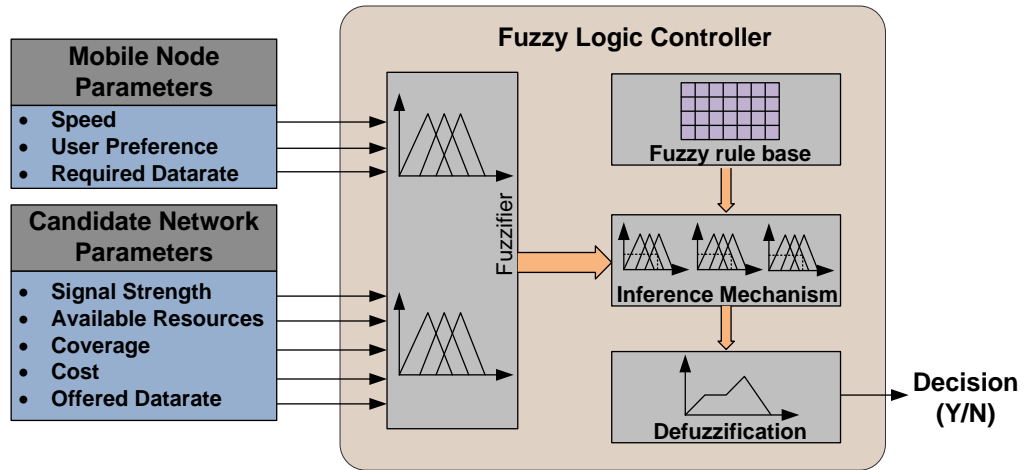

Figure 4-16: Fuzzy logic controller for load aware RAT selection

Figure 4-16, presents the proposed fuzzy logic controller for load aware RAT selection. It considers the following 8 input parameters:

- Signal strength of the available networks (SS),

- Available resources (AR),

- Network coverage area (CA),

- Cost of network (C),

- Offered data rate (ODR),

- User's network preferences (UNP),

- Required data rate bandwidth for user connections (RDR), and

- Speed of mobile node (S).

The fuzzy logic controller is used by each of the mobile node for load balancing based handover decision. The Fuzzy logic controller consists of the following four main components:

- The fuzzifier: The fuzzifier is a membership function, which can of any shape such as a curve or a line. It converts the input data of each input to the corresponding linguistic value for fuzzy set operations, by mapping the input values to fuzzy sets. One input value in a fuzzy set can have membership of more than one set. For example on input value for the network load can be mapped onto low, medium or high set. The fuzzy sets are used in the fuzzy rules.

- The inference engine: The inference engine or inference mechanism is a process which involves utilizing membership functions, applying logical Min and Max operations and applying If-Then rules. Two commonly used inference methods are Min inference and Product inference. In Product inference method the output variable is scaled by the rule premise's degree of truth. Whereas in case of Min product method, the output variable is assigned with the minimum value of the rule premise's degree of truth.

- The rule base or knowledge base: it gives the knowledge of the appropriate fuzzy operation on the fuzzy sets. The knowledge base is usually defined as a series of IF-Then rules format. For example the IF-Then rule in RAT selection decision which consists of eight input

criteria is: IF (Signal strength is high and Available resources are high and Mobile node speed is high and Network coverage is high and User acceptable network Cost is high and Network data rate is high and Network cost is low and User required data rate is low) Then decision for selection the network is Yes.

- The defuzzifier: The defuzzifier converts fuzzy decision set into a precise quantity which can be applied to a target system. The input for the defuzzifier is a fuzzy set and the output is a single number. There are multiple defuzzification methods such as centroid, bisector, average of maximum, largest of maximum and maximum method. Most commonly used methods are centroid and maximum methods. The final output value is computed in centroid method by calculating the centre of gravity of fuzzy output variable. In the maximum defuzzification method the final value is the maximum value in its fuzzy decision set.

The system diagram of the fuzzy logic controller used in this research for load aware RAT selection is shown in Figure 4-16. Two different types of membership functions such as trapezoidal and triangular membership functions have been used in fuzzification. These membership functions for each input are also shown in the following set of figures from Figure 4-17 to Figure 4-24. Three membership functions i.e. Low (L), Medium (M) and High (H) have been defined for each of the input parameters used as the fuzzy input variable in the fuzzy logic controller. The handover decision (HO) is the output linguistic parameter, which has three membership functions named as Yes (Y), Probably Yes (PY) and No (N) respectively.
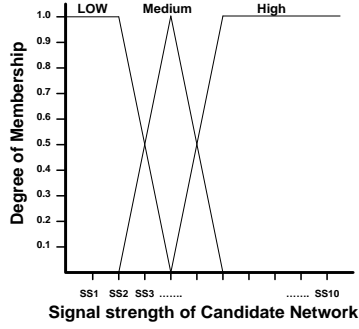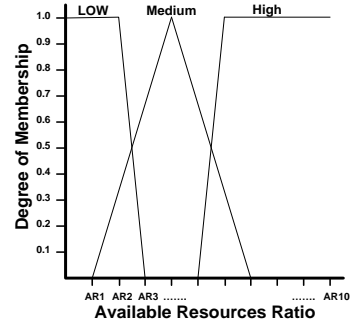
114

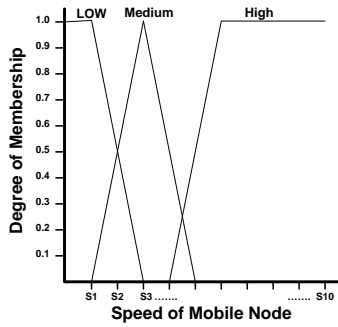Figure 4-17: Signal Strength (SS)

Figure 4-18: Available Resources Ratio
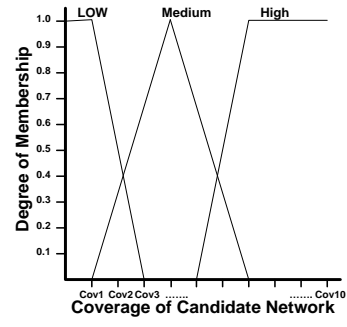

Figure 4-19: Speed (S)

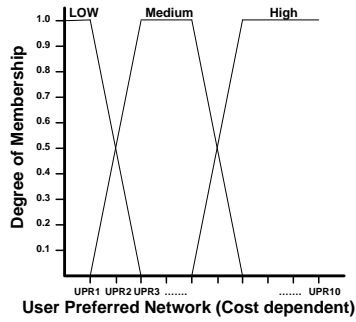Figure 4-20: Coverage (CA.)
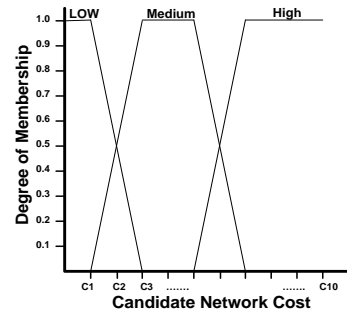

Figure 4-21: User Network preference (UNP)
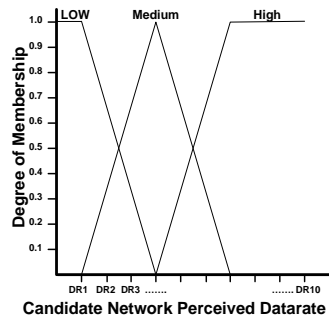

Figure 4-22: Remote Network Cost (C)

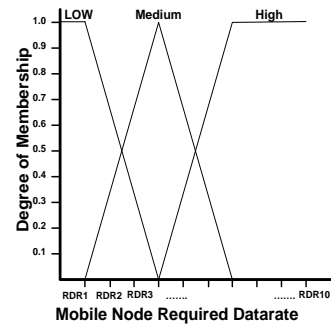Figure 4-23: Offered Network Datarate (ODR)

Figure 4-24: Required Datarate (RDR)

The "mamdani" and "sugeno-type" implication methods [114] are commonly considered in fuzzy logic controllers. While the "mamdani" implication method

is intuitive, well suited to human input and has widespread acceptance, the "sugeno-type" implication method is computationally efficient, shows best performance for linear and adaptive techniques and is well suited to the mathematical analysis. The implication method used in the proposed fuzzy logic controller for load aware RAT selection mechanism is the "mamdani" implication method.

The term sets of SS, AR, CA, C, ODR, UNP, RDR and S are defined as follows:

T(SS) = {Low, Medium, High} {L, M, H}

T(AR) = {Low, Medium, High} {L, M, H}

T(CA) = {Low, Medium, High} {L, M, H}

T(C) = {Low, Medium, High} {L, M, H}

T(PDR) = {Low, Medium, High} {L, M, H}

T(UNP) = {Low, Medium, High} {L, M, H}

T(RDR) = {Low, Medium, High} {L, M, H}

 T(S) = {Low, Medium, High} {L, M, H}

T(HO) = {Yes, Probably Yes, No } {Y, PY, N}

The fuzzy inference engine utilizes the predefined list of fuzzy rules from the rule base to admit the incoming call or the suitable selected active call to the suitable network. These predefined set of rules are the series of 'If-Then' statements of rules. As shown above there are eight different input parameters and each input parameter has three member functions such as low, high and medium. There are 3^8 rules in the rule base which are being utilized by the inference engine of the fuzzy controller. The term used for the set of rules in fuzzy logic is antecedent (input rules) and the term used for

output is consequent. Assuming that there are 'r' "If-Then" propositions (rules), and each fuzzy antecedent in set A_k consists of eight input elements $\{i_1, i_2, i_3, i_4, i_5, i_6, i_7, i_8\}$. A set of consequents O_k represents corresponding output for each proposition in A_k. As the logic operation AND has been used between antecedents for rules the equation for the consequents will be as shown in the following Eq. [5] and Eq. [6].

$$IF\ A_k = [I_{k1}, I_{k2}, I_{k3}, \ldots I_{k8}]$$ Eq. [5]

$$THEN\ O_k = min[I_{k1}, I_{k2}, I_{k3}, \ldots, I_{k8}]$$ Eq. [6]

The three "IF-Then" conditional statements which considers all eight input parameters take account of all $r = 3^8$ (6561) rules for the corresponding outputs. The process of defuzzification comprises of the operation which converts the fuzzy output HO into crisp output (HO*). A number of defuzzification mechanisms are available in the literature such as centroid method, bisector method, weighted average method and middle, smallest and largest of maximum methods also known as Middle of Maximum (MOM), Smallest of Maximum (SOM) and Largest of Maximum (LOM) respectively [115,116]. The method where a vertical line divides the resultant region into two equal regions using a bisector line is called bisector method. MOM, SOM and LOM methods key off the maximum value assumed by the aggregated membership function. The centroid method is the most commonly used defuzzification method as this method draws the conclusion using most of the information from monolithic membership function. This implies that most of the rules from the fuzzy rulebase have been taken into consideration while generating the final output. The centroid defuzzification method has one drawback which is computationally intensive. The weighted average method

is used in the proposed fuzzy based load aware RAT selection technique as it gives results very close to the centroid method and requires less calculations or computation resources as compare to the centroid method. Figure 4-25 to Figure 4-28 represent the fuzzy RAT selection or handover decision for *r* fuzzy rules in the fuzzy logic controller using different fuzzy input variables on the x and y axis of the 3-D graph. These figures show how the value of load-aware RAT selection factor varies with respect to the variations in the different input parameters.
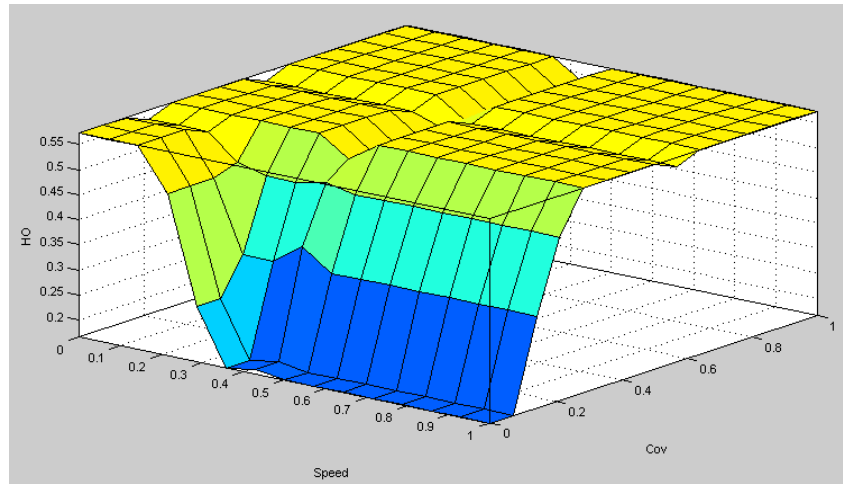

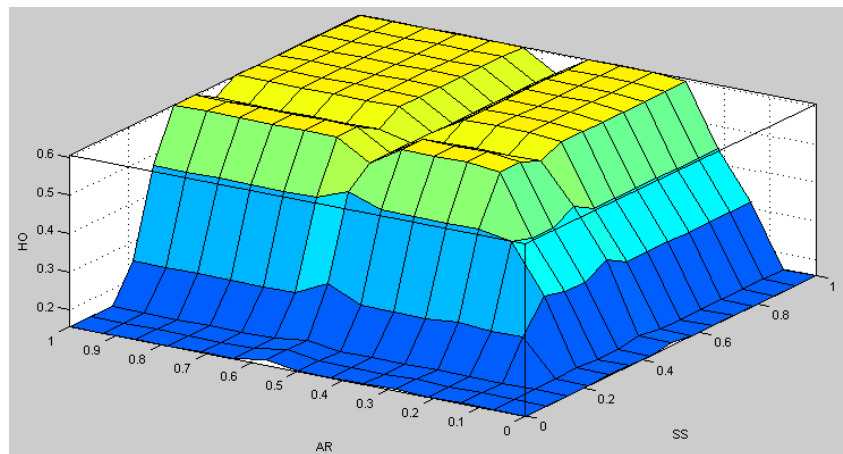Figure 4-25: Speed vs. Coverage area

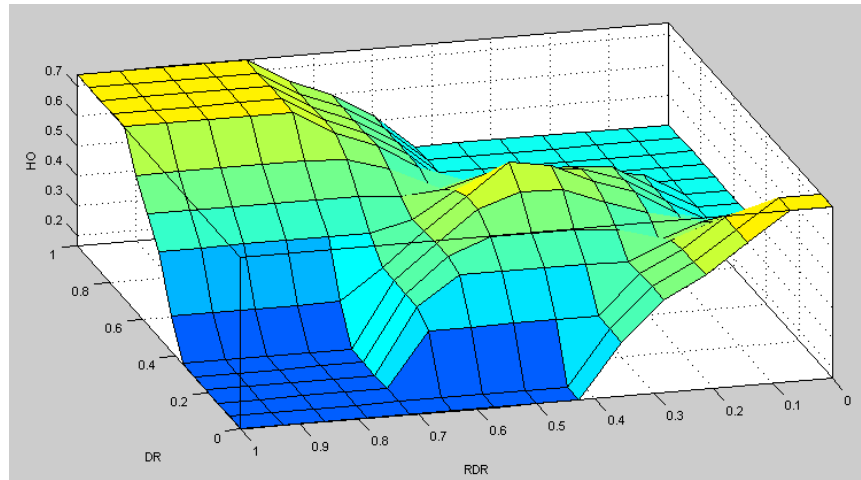Figure 4-26: Available resources ratio vs. signal strength

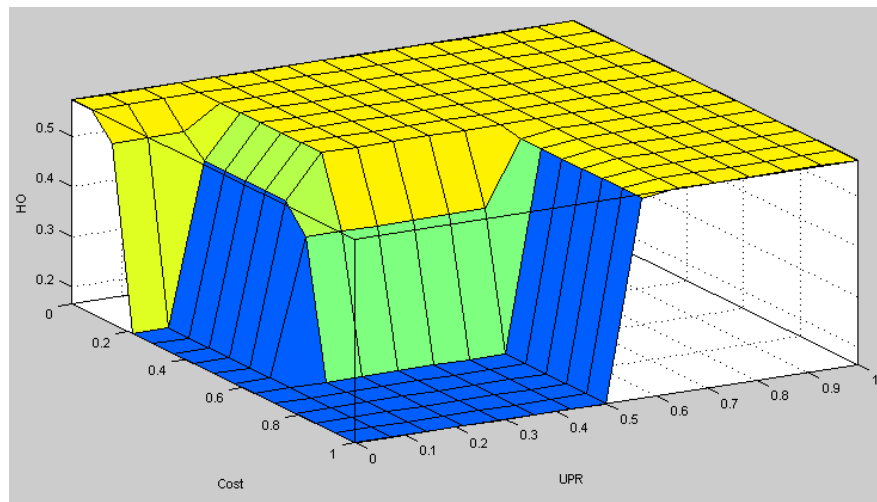Figure 4-27: Offered datarate vs. required datarate

Figure 4-28: Cost Vs. user preferred network cost

4.5.2.1 Load-Aware RAT Selection Example

In this section the process of fuzzy logic based load aware RAT selection is described with the help of an example. Assume that there are two new networks detected by the mobile node which are, network 1 (N1) and network 2 (N2).
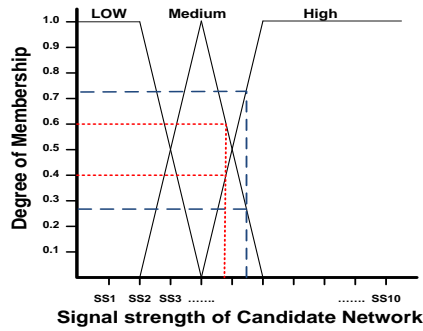
Figure 4-29: Signal Strength (SS)
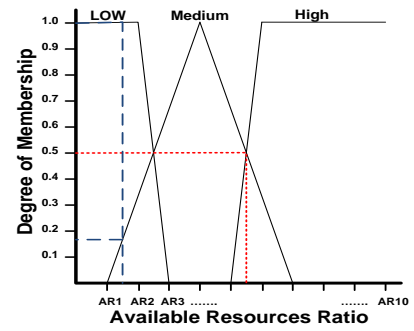
Figure 4-30: Available Resources Ratio (Local Resources/Remote Resources) in %

Figure 4-31: Speed (S)

Figure 4-32: Coverage (CA)

Figure 4-33: User preferred Network (UNP)

Figure 4-34: Remote Network Cost (C)

Figure 4-35: Offered Network Datarate (ODR)

Figure 4-36: Required Datarate (RDR)

The membership function values for each fuzzy input variable of these networks, shown from Figure 4-29 to Figure 4-36 are shown in the Table 4- 2

120

and Table 4-3. The membership values of N1 fuzzy input variables are shown in red lines with small dots and that of N2 are shown in blue lines with large dots. At the mobile node all the input parameters values are fuzzified and their d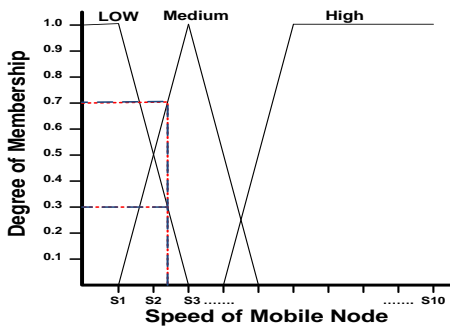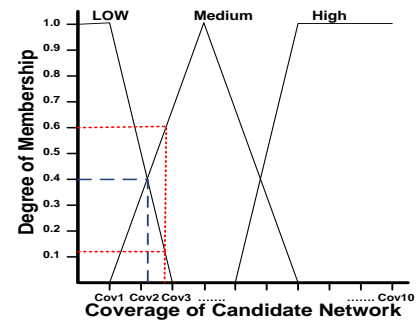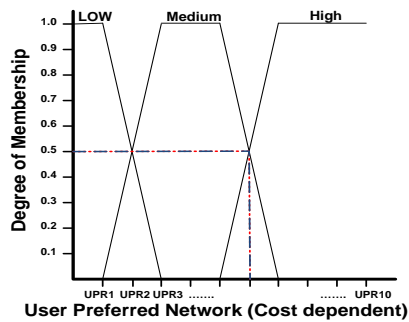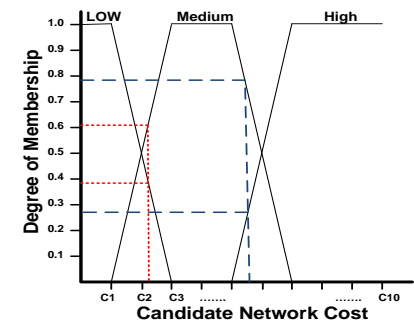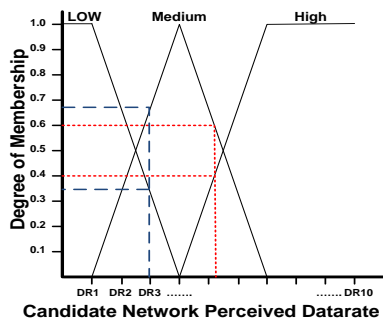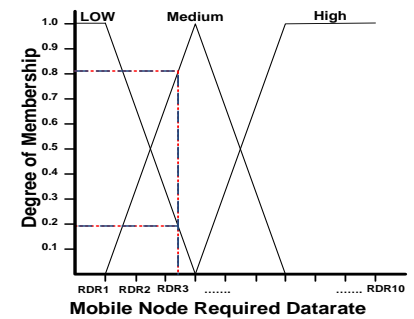egree of memberships have been measured for all three membership functions such as L, M and H. In this particular example N1 fuzzy input values are shown in Table 4- 2.

Table 4- 2: Membership obtained from N1 & mobile node

Criteria	Low (L)	Medium (M)	High (H)
Signal Strength (SS)	0	0.6	0.4
Available resources ratio (AR)	0	0.5	0.5
Speed (S)	0.3	0.7	0
Coverage (Cov.)	0.1	0.6	0
User preference (UPR)	0	0.5	0.5
Candidate network cost (C)	0.4	0.6	0
Candidate Network Datarate (DR)	0	0.4	0.6
User required datarate (RDR)	0.2	0.8	0

Similarly the fuzzy input values of same parameters from N2 are shown in the Table 4- 3 shown below:

Table 4- 3: Membership values obtained from N2 & mobile node

Criteria	Low (L)	Medium (M)	High (H)
Signal Strength (SS)	0	0.275	0.725
Available resources ratio (AR)	1	0.175	0
Speed (S)	0.3	0.7	0
Coverage (Cov.)	0.4	0.4	0
User preference (UPR)	0	0.5	0.5
Candidate network cost (C)	0	0.8	0.275
Candidate Network Datarate (DR)	0.35	0.68	0
User required datarate (RDR)	0.2	0.8	0

Once the membership values of all input variables have been assigned, then the set of these measured membership values are compared against the logical lookup table of *r* rules in the fuzzy rule base. As there are eight different input variables and for this example each having two different membership values, therefore results of If-Else rules gives us $2^8 = 256$ different combinations. In this scenario, the UNION operation of the fuzzy set will be used in determining the RAT selection factor. In this example the Average Weighted method has been utilized to defuzzify the obtained fuzzy decision values. Eventually to obtain the RAT selection factor, there is a need to construct another weighting matrix which defines the weighting of each decision element. If the weightings assigned to each possibilities such that Y, PY and N are 0.7, 0.4 and 0.1 respectively, then the RAT selection factor can be derived as follows:

$$\left(Z^* = \frac{\Sigma \mu_c\left(\overline{z}\right).\overline{z}}{\Sigma \mu_c \overline{z}} \right) \qquad \text{Eq. [7]}$$

In this example the Z* is the RAT selection factor, $\mu_{c\,(\overline{z})}$ is the membership value of each element in decision set and $\overline{z}$ is the weight assigned to each particular decision element such as Y, PY and N respectively. Substituting these values to the above equation gives us the following value as RAT selection factor.

$$RAT\ selection\ factor = \frac{(0.1 \times 0.1) + (0.4 \times 0.4) + (0.5 \times 0.7)}{(0.1 + 0.4 + 0.5)} = 0.52 \qquad \text{Eq. [8]}$$

In case of N2 different membership values for decision "N" are {0.175, 0.2, 0.275, 0.3 and 0.4} and the maximum value 0.4 is selected using UNION

operation. For "PY" and "Y" only one value 0.175 is available which is selected for both. Using these values the RAT selection factor for N2 is obtained in the following equation Eq. [9] using Eq. [7].

$$RAT\ selection\ factor\ = \frac{(0.4 \times 0.1) + (0.175 \times 0.4) + (0.175 \times 0.7)}{(0.4 + 0.175 + 0.175)} = 0.31$$

<div align="right">Eq. [9]</div>

In this example two networks N1 and N2 were considered, where RAT selection factor for N1 is 0.52 and that of N2 is 0.31, as shown in Eq. [8] and Eq. [9]. The RAT selection factor for N2 is "No" as the value is lesser than 0.4, and for N1 this value is 0.52 which suggests moving to this network if there is no other network available with higher RAT selection factor value.

4.5.3 Neural-fuzzy load balancing algorithm

The Artificial Neural Network (ANN) can be termed as a black box, which takes 'n' input values x_1, x_2, x_3, …, x_n and process the input vectors x_N to produce output vector z_N. The output vector 'z' represents the pattern or identification group. A trained artificial neural network represents the system that maps a set of input vectors x_N: N=1,2,3, …, n to a set of target output vectors z_N: N=1,2,3, …, n. This mapping enables the neural network to make interpolations and extrapolations to correspond any input x to the output z, which best matches the input pattern. After training the artificial neural network, it acts as a mathematical machine that implements the algorithm specified by the input/output nodes, nodes in the hidden layers, connecting lines, transforming nodes functions and the weight associated with the connecting lines of the artificial neural network. For a particular application an ANN must be trained to acquire the suitable weights on the connecting lines,

so that ANN can produce the close approximation of target result. The ANN is a better option as compared to the other artificial intelligence techniques such as Fuzzy logic (FL), neural network (NN) and genetic algorithm (GA) as it does not use pre-programmed knowledge base, have no restrictive assumptions, can handle noisy/imprecise data, robust and flexible. On other hand it has some drawbacks too, such as it requires high quality data for training/learning, variables must be very carefully selected a priori, risk of over fitting, requires a definition of architecture, long processing time for training and possibility for illegal network behaviour. The ability to embed the empirical data into the fuzzy control system can be achieved by utilizing training techniques of neural networks. This can greatly widen the application of fuzzy system as the ability to make use of both empirical and expert information. The dimensionality in the fuzzy systems is its limitation or in other words a severe drawback. The term "dimensionality curse" [117] is used for fuzzy system as for a fuzzy system the cost for implementation of rule base and deriving the output increases exponentially as the input space dimension increases. The expert information to model the input space relationship could be utilized to reduce the set of rules, as the expert knowledge may make the problem tractable. Relying only on expert knowledge is not enough to tune a fuzzy system for efficient and precise output. The use of training techniques based on error allows a fuzzy system to acquire the complexities hidden in the input data. A neural-fuzzy technique can be used for building a fuzzy system in multiple ways apart from training method. For example, it can be used for fuzzy membership function determination, in fuzzy rule selection and also in case of hybrid systems. A

124

hybrid system is which uses both fuzzy and neural network systems. There can be endless applications and combinations for the neural-fuzzy systems in hybrid way such as a neural network may intelligently associates the output of several fuzzy systems or a monitoring fuzzy system may choose the suitable output from multiple neural networks in a hybrid system [118, 113].
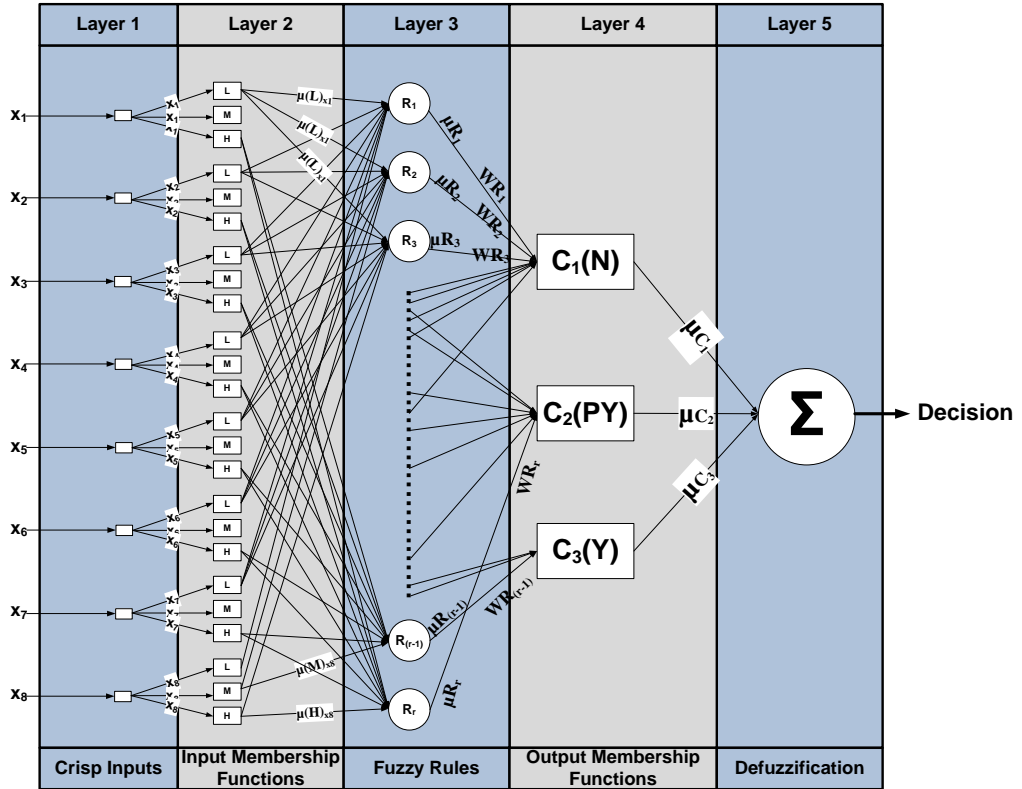
Figure 4-37: Neural-fuzzy system layers

Figure 4-37 describes the neural network representation of the fuzzy control system. The inputs x_1, x_2, x_3, x_4, x_5, x_6, x_7 and x_8 represent the input parameters of our neural-fuzzy control system such as: signal strength, available resources, speed of MN, coverage, user preference, cost, offered data rate and required data rate respectively. The purpose of utilizing the neural network here is to take advantage of the neural network intelligent techniques for deriving the reduced number of rules and get rid of the fuzzy

logic dimensionality curse which affects the efficiency of fuzzy algorithm due to the large number of input parameters in the target load balancing system.

Layer 1 in Figure 4-37 shows the eight inputs to the system. The membership of input parameters or fuzzification is performed at layer 2. The fuzzy sets which are used in the antecedents of fuzzy rules are represented by the neurons in layer 2. The fuzzification neuron after receiving the input, determines the grade to which the input belongs to neuron's fuzzy sets. Layer 3 represents the fuzzy rules or can also be called the fuzzy rules layer. Each fuzzy rule is represented by a unique neuron in this layer. The neuron (fuzzy rule neuron) in this layer receives input from fuzzification neurons from layer 2, which denotes fuzzy sets. The intersection operation can be implemented in neuro fuzzy systems, using product operator. Therefore the output of the ith neuron in the layer 3 can be determined as:

$$\mu R_i = x_{1i} \times x_{2i} \times x_{3i} \times x_{4i} \times x_{5i} \times x_{6i} \times x_{7i} \times x_{8i} \qquad \text{Eq. [10]}$$

Substituting the values in Eq. (10) for rule 1 we get the following:

$$\mu R_1 = \mu(L)x_1 \times \mu(L)x_2 \times \mu(L)x_3 \times \mu(L)x_4 \times \mu(L)x_5 \times \mu(L)x_6 \\ \times \mu(L)x_7 \times \mu(L)x_8$$

$$\text{Eq. [11]}$$

The firing strength of the neuron which denotes rule R1 can be represented by the μR$_1$. The layer 4 which can also be called the output membership layer denotes the fuzzy sets in the output of the fuzzy rules. The neurons in this layer receive inputs from the neurons in the fuzzy rules layer. Once the inputs are received from the layer three, the neurons in this layer combine all the received inputs using union fuzzy operation. The union fuzzy operation

can be implemented by exclusive OR operator which is also known as the probabilistic OR operator. The output of the neurons at this layer can be denoted by μC_i , where 'i' is the number of neurons in this layer. Eq. [12] represents the general format for firing strength of neurons in layer four.

$$\mu C_i = \mu R_1 \oplus \mu R_2 \oplus \mu R_2 \oplus \mu R_2 \ldots\ldots.\oplus \mu R_n \qquad \text{Eq. [12]}$$

Here 'n' represents the total number of rules which satisfy the output condition 'i'. The layer 5 which is also called the defuzzification layer is composed of a single neuron. The neuron in this layer represents the single output of the fuzzy neural system. The neuron in the layer 5 receives the input from the output of the neurons in the layer 4 and combines them into a single fuzzy set. A standard defuzzification method centroid, average weighted or any other can be applied in neural-fuzzy system. In this case average weighted. Equation shown as Eq. [13] represents the output decision of the fuzzy neural system using the weighted average method.

$$Decision = \frac{\mu C_1 \times aC_1 \times bC_1 + \mu C_2 \times aC_2 \times bC_2 + \mu C_3 \times aC_3 \times bC_3}{\mu C_1 + \mu C_2 + \mu C_3}$$

Eq. [13]

Here 'a' is centre and 'b' is the width of triangular activation/membership function. The neural network representation of the fuzzy control system shown in Figure 4-37 still requires a lot of computation power on the layer 2 where a large number of fuzzy rules are represented by the neurons. This huge computation requirement can be reduced to a large extent by converging fuzzy inputs into a set of general rules which are based on the fuzzy output of the rules set. The generalized rules representation operation reduced the total number of rules to 2000 from the large number of 6561.

The neural-fuzzy system with reduced number of rules is trained to perform efficiently with the set of rules data in hand. This training process of the neural-fuzzy system adjusts the weights to produce the required output for any particular input pattern. In fuzzy system the same process is carried out by tuning the membership functions, which was nearly impossible for the target load balancing system with such a large input parameters. Once the neural-fuzzy system is trained, a comparison is made between neural-fuzzy and the fuzzy inference system with the help of 500 input samples for each of the eight input parameters. Each of these inputs are randomly generated and provided to the neural-fuzzy and the fuzzy inference system. Finally the obtained decision factor from both neural-fuzzy and fuzzy inference system is plotted to compare the output of each system.
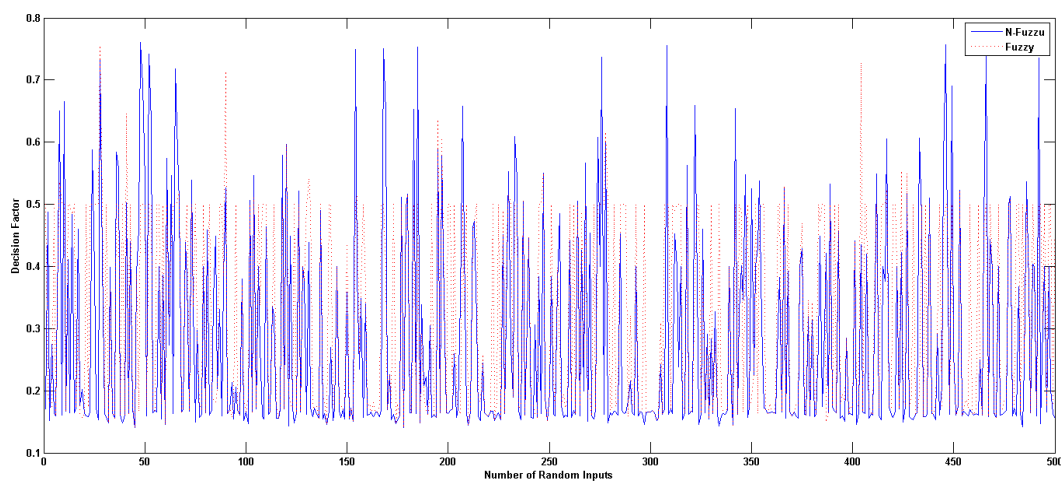

Figure 4-38: Comparison of Neural-fuzzy and Fuzzy Inference system using 500 random Input samples

Figure 4-38 shows the comparison of output decision factor for handover using both neural-fuzzy and fuzzy inference system. This output is obtained by applying 500 random input samples of the input parameters using MATLAB. The x-axis represents the total number of input samples and y-axis

represents the output decision factor which is used to make handover decision. The higher value of output on y-axis means more chances of handover and low value of output on y-axis reflects less chances of handover. The results show that in case of fuzzy inference system the number of handovers can be higher as the fuzzy system is not using properly tuned membership functions. The tuning process of membership functions for target fuzzy inference system is nearly impossible due the curse of dimensionality in fuzzy system as discussed in earlier section. On other hand the neural-fuzzy inference system is properly trained with the set of input output data of all possible rules therefore it give more precise results with controlled number of handovers as output value.

4.6 Summary

This chapter provides the detailed description of the proposed load balancing framework using the IEEE 802.21 reference model. The network architecture is explained in the beginning and major entities in the architecture have been described. The load balancing algorithm is implemented as an integral module and has been integrated into the MIH reference model so that the advantages of MIH reference model can be utilized for collecting the neighbouring networks information before registration or handover to that network. The proposed load balancing algorithm is composed of two independent part; one running in the mobile node and the other in the network entity such as AP, BS, RNC and RAN. Each part of the algorithm is explained with the help of flow chart in previous sections of this chapter. The MIIS entity located at the core network collects and distributes the network information such as load, cost, coverage area, latencies and QoS

parameters. As the handover process is the vital part of load balancing in the heterogeneous wireless networks therefore the combination of different possible handover scenarios have been explained with examples by showing MIH messages exchange. At the end, the proposed Baseline Fuzzy based and Neural-fuzzy based RAT selection algorithms have been explained in detail. All three algorithms take various different parameters as input and provide the output for decision of selecting the most suitable radio access technology from the available networks to the mobile node.

Chapter 5: Simulation framework

5.1 Overview

This chapter describes the simulation framework developed using Network Simulator 2 (NS2) to study the performance of the load balancing framework. Different load balancing algorithms have been implemented externally in Ansi C or Matlab and then integrated into the NS2 model. This chapter explains the details of the various tools used and also presents different simulation scenarios and parameters.

5.2 Tools

There are various software tools, programming languages and scripting languages that have been utilised in this research work. A brief explanation mentioning where and how these tools are utilised is provided in the following sub-sections:

5.2.1 Network Simulator

NS2 is an open source freely available network simulation tool which is being use by the vast community of researchers all over the world as it has a huge list of features and almost all the wired and wireless communication protocols implemented [119,120]. NS2 is an object oriented discrete event simulator which is mainly composed of two different languages which are TCL/OTCL and C/C++. While C/C++ is the back-end language used for implementing the various protocols and methods, TCL/OTCL is the front-end language which is used for defining the simulation topology and scenarios.

The end user in NS2 writes the code for the simulation scenario in TCL script which uses the C/C++ implementation of different protocols to simulate the

scenario and generates the trace file and the NAM (Network Animator) file. NAM is animation software which shows the graphical display for the scenarios simulated in NS2. The trace file is a text file where each row represents an event and each column represents the different attribute for that particular event, such as "event type", "time", "packet id", "source", "destination" and "packet size" etc.

5.2.2 C/C++ and MATLAB

While the baseline algorithm has been implemented purely in "Ansi C" [121], the fuzzy and the neural-fuzzy algorithms have been implemented in MATLAB [122] and "Ansi C". The fuzzy and neural-fuzzy algorithms have been integrated in NS2 using a new custom-made "Ansi C" library developed primarily for load balancing purposes in NS2. MATLAB was also used to plot 2-D.

5.2.3 Scripting Languages

While TCL scripting language [123] is used the simulation scenario definition, AWK and bash scripting have been used for tracefile analysis. As the results trace files were very big in size, custom scripts were required for analysing the results.

5.2.4 Graphs Plotting Utilities

After processing the trace file using scripting languages the required results is plotted using TraceGraph, GNUPlot and the MATLAB.

5.3 Implementation of Load Balancing Framework

The proposed load balancing framework is implemented using C/C++ within NS2. The NIST [98, 124 and 125] mobility model is used to take advantage of IEEE 802.21 Media Independent Handover (MIH) implementation in NS2. Figure 5-1 shows the overall load balancing framework developed in NS2. The load balancing framework is developed and integrated with this extended MIH implementation.
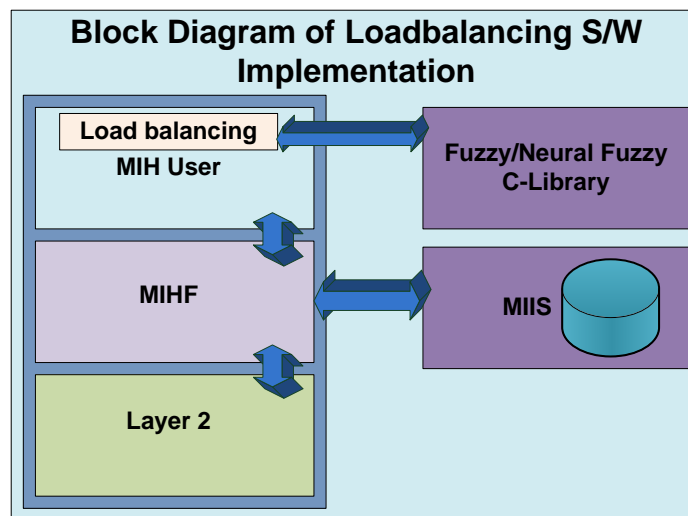

Figure 5-1: Integrated software components of MIH in NS2

The following have been implemented in this NS2 load balancing framework:

- A new load balancing module as a part of the existing MIH User

- Load balancing Algorithms library

    o The baseline least-loaded algorithm

    o The Fuzzy based load balancing algorithm

    o The Neural-fuzzy based load balancing algorithm

- Support for multiple connections over multiple interfaces in multi-interface MN

- Extensions to the MIH model

    o Addition of new primitives for MIHF-MIIS communication

- o Addition of satellite network support in MIH

- o MIIS implementation

The load balancing module utilises the custom-created C library for the fuzzy and neural-fuzzy systems for load balancing decision making purposes. The communication between MIHF and MIIS is achieved by the *MIH_Get_Information* primitive for obtaining network information from the MIIS and the *MIH_Set_Information* primitive used for updating the MIIS with local network information such as current load status.

The blue arrows in Figure 5-1 represent the flow of information in the implemented load balancing framework. For example for a mobile triggered load balancing scenario, layer 2 sends up the link updates to MIHF which then forwards it to the MIH User. At the MIH User, the load balancing module receives the link updates and queries the MIIS via the local MIHF for information on the available networks. Once the MIH User gets the response of this query, it then processes the information using any of the three load balancing algorithms depending on the simulation scenario.

5.4 Simulation topology and scenarios

The simulation topology implemented in NS2 to analyse the performance of the load balancing framework and algorithms is presented in Figure 5-4. This topology aims to simulate a real-life situation where a mobile may move across the overlapping coverage areas of different networks i.e. satellite networks, WiMax, UMTS and WLAN networks. Different networks contain a group of mobile users in their coverage areas at different time. A group of mobile users have been assumed to travel across different networks. The number of users in this moving group and their speed of movement can be

134

configured in the simulation. There are different set of test scenarios targeted on the topology shown in Figure 5-2.
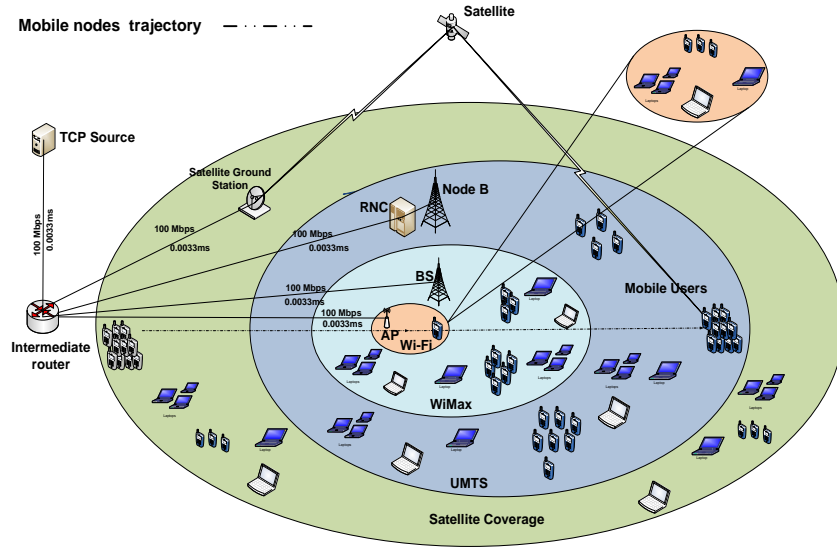

Figure 5-2: Satellite-terrestrial hybrid scenario

Table 5-1 represent the simulation parameters for the target simulation scenarios. In all scenarios the mobile nodes start from satellite only coverage area and pass through common coverage areas of all four wireless networks such as WLAN, WiMax, UMTS and satellite, as shown by the dotted trajectory path in Figure 5-2.

Table 5-1: Simulation parameters

Simulation parameters	Values
Satellite coverage radius	4000 meters
UMTS coverage radius	1000 meters
WiMax coverage radius	500 meters
WLAN radius	100 meters
Satellite data rate (per user)	492 kbps
UMTS data rate (per user)	384 kbps
WiMax data rate	45 Mbps
WLAN data rate	11 Mbps
Wired links capacity	100 Mbps
Propagation delays wired links	0.0033 ms
Propagation delay satellite	250ms
Application type	TCP - CBR
Application data rate	2 kB/s
Number of mobile nodes	50, 100
Speed of mobile nodes	2m/s, 25m/s

Different scenarios are simulated based on varying the following:

- Speed of the moving users i.e. 25m/s (high) or 2m/s (low)

- Number of moving users (50 or 100)

- Load balancing support (No load balancing, baseline load balancing, fuzzy load balancing and fuzzy neural load balancing algorithms.)

- Support for cost of service (No cost preference, with cost preference)

Table 5- 2: Simulation Scenarios

Scenario No.	Sub-Scenario	Number of nodes	Speed	Load balancing	Cost
1	a)	50	Slow	Baseline	Same
	b)	50	High	Baseline	Same
	c)	50	Slow	Baseline	Different
	d)	50	High	Baseline	Different
2	a)	100	Slow	Baseline	Same
	b)	100	High	Baseline	Same
	c)	100	Slow	Baseline	Different
	d)	100	High	Baseline	Different
3	a)	50	Slow	Fuzzy	Same
	b)	50	High	Fuzzy	Same
	c)	50	Slow	Fuzzy	Different
	d)	50	High	Fuzzy	Different
4	a)	100	Slow	Fuzzy	Same
	b)	100	High	Fuzzy	Same
	c)	100	Slow	Fuzzy	Different
	d)	100	High	Fuzzy	Different
5	a)	50	Slow	Neural-fuzzy	Same
	b)	50	High	Neural-fuzzy	Same
	c)	50	Slow	Neural-fuzzy	Different
	d)	50	High	Neural-fuzzy	Different
6	a)	100	Slow	Neural-fuzzy	Same
	b)	100	High	Neural-fuzzy	Same
	c)	100	Slow	Neural-fuzzy	Different
	d)	100	High	Neural-fuzzy	Different

Table 5-1 shows all the scenarios considered for results and analysis purpose. All the scenarios and their comparisons are briefly explained with the obtained results in the following section.

5.5 Evaluation Methodology

Different scenarios as described in the above section are simulation using the NS2 simulation framework in order to evaluate the performance of the proposed load balancing framework and algorithms. For each of the simulation scenario the following set of parameters are observed and analysed:

- Average throughput of mobile nodes

- Packets drop rate

- Throughput at different networks

- Average handover latencies

- Load at each network

The obtained results are individually analysed but then also cross compared to study the effect of the various algorithms across different scenarios. While some of the above parameters like packets drop rate, Average throughput at each mobile node, traffic on each network, total handover latencies observed by each mobile node and total number of handovers performed by each mobile node for the different scenarios can be easily compared, direct comparisons of parameters is not possible in some cases. The following subsection briefly introduces these issues and the methodology adopted for analysing the results

5.5.1 Comparison of network load

For the comparison of network load it is necessary to choose some time points in the simulation where the load on each network changes. Figure 5-3 shows the selected points in the simulation topology where the load of each

network is plotted for comparison between different simulation scenarios. The seven points shown in Figure 5-3 are selected for monitoring the load as the load in different networks changes on these points. The reason for change of networks load on these points is that these points are located at the boundary of different networks coverage areas. When mobile nodes enter or leave the coverage area of any network the load aware RAT selection is triggered and it handovers mobile nodes to appropriate available networks.
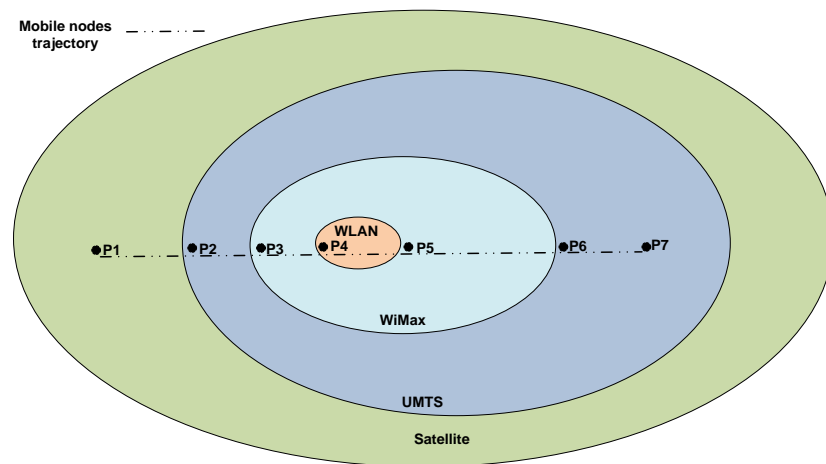
Figure 5-3: Methodology for comparison of network load

Position P1 in Figure 5-3 is the place where mobile nodes start their movement from the satellite only coverage area. Position P2 is the location where satellite coverage area overlaps the UMTS coverage area. Similarly position P3 represents the place where WiMax coverage area is begins to overlap satellite and UMTS common coverage areas. The points P4 and P5 represent the start and end of WLAN coverage area. P6 represents the place where WiMax coverage finishes on the mobile nodes trajectory. The point P7 is the final destination of all the mobile nodes and this area is the common coverage area of UMTS and satellite networks. Network load at these various

138

positions will be compared to analyse the performance of the proposed algorithms.

5.5.2 Study the Effect of User speed

Two different user speeds are considered in the simulation scenarios; low speed of 2m/s to simulated walking users and high speed of 25m/s to simulate users in a vehicle. Depending on the speed of the moving users, they would reach the above described positions (P1 to P7) at different times. The total time taken to travel the trajectory is also different for different speed. Hence direct comparison using simulation time is not feasible when comparing similar scenarios with also speed changes. For these comparisons also the above mentioned positions are used. Table 5-2 shows the time in simulation; when all the mobile nodes reach these different positions for both the high speed and low speed scenarios.

Table 5- 3: Time for each point with respect to mobile node velocity

Points	Times for 25m/s scenarios	Times for 2m/s scenarios
P1	8 seconds	8 seconds
P2	30 seconds	250 seconds
P3	51 seconds	511 seconds
P4	66 seconds	710 seconds
P5	74seconds	810 seconds
P6	90 seconds	1010 seconds
P7	250 seconds	1500 seconds

5.5.3 Effect of Cost preference

The performance of algorithms has also been compared when the cost of service is a parameter of concern for the user. Networks like the satellite networks may have a higher tariff for usage as compared to the terrestrial networks like UMTS. To observe the effects of cost on load balancing, it is

assumed in all scenarios that 70% of the mobile nodes in the scenarios do not want to pay for the satellite networks when other terrestrial networks are available to them. Hence as soon as these users come into the coverage of a cheaper terrestrial network they prefer to move away from the satellite networks. It is assumed that the remaining 30% mobile nodes are however willing to pay for these satellite networks. These users may be premium users who do not want to handover frequently or they could be on a pricing model that is uniform across networks.

5.6 Summary

In the beginning, this chapter first presents the implementation tools and software implementation of the proposed load balancing framework. It then presents the simulation topology and finally the target scenarios for simulation. At the end in "Evaluation Methodology" section this chapter explained the methodology of how the load at different networks is compared in different scenarios. The influence on load balancing mechanism by user speed and the cost of services offered by different wireless network access technologies are also explained on the load balancing mechanism.

Chapter 6: RESULTS AND ANALYSIS

6.1 Overview

In this chapter the results of the target scenarios are presented with the help of graphs and tables. The results from different scenarios are compared using different algorithms i.e. baseline load balancing algorithm, fuzzy load balancing algorithm and neural-fuzzy load balancing algorithm. Different evaluation parameters such as network load, packet drop rate, average bandwidth observed at mobile node side, average bandwidth utilized on each network and handover latencies are considered to performance evaluation with and without load balancing in different scenarios.

6.2 Results

This section presents the results obtained for the various scenarios simulated to evaluate the performance of each of the proposed algorithm. The results are presented in the following order:

- Baseline Least-Loaded load balancing algorithm (50 users, 100 users)
- Fuzzy based load balancing algorithm (50 users, 100 users)
- Neural-Fuzzy based load balancing algorithm (50 users, 100 users)

Finally the results are cross-compared in order to study the performance of each algorithm with respect to each other.

6.2.1 Scenario 1 – Baseline Least-loaded algorithm with 50 users

The results for the 50 mobile nodes scenarios using baseline load balancing algorithm are presented in this section.

a) Network load

The graphs shown in Figure 6-2 to 6-7 represent the load in each network such as satellite, UMTS, WiMax and WLAN at the different position of travel trajectory for the 50 mobile nodes scenarios with the mobile nodes moving at 25m/s and 2m/s.

These graphs show the load for the simulation scenarios when no load balancing is applied, baseline load balancing is applied and when baseline load balancing algorithm is applied with cost preferences. The x-axis of the graphs represents the selected time points where the load in each simulation scenario is monitored. The y-axis of the each load graph represents the total load in terms of number of users in that particular network.

It can be seen from these obtained load results that without load balancing most of the mobile nodes handover to the best available network in terms of cost and network latencies. For example in position 4 in Figure 6-1 and 6-2, we can see that all the users connected to WLAN. This however leaves the other networks under-loaded or underutilised. In the scenario where baseline load balancing is applied all the networks share the load where possible such as in the overlapped coverage areas. For example in the same position 4 in Figure 6-3 and 6-4, we can see that the users are distributed across the different networks.

Figure 6-5 and Figure 6-6 represent the load in different networks when baseline load balancing is applied with cost preferences. It can be seen that the network cost affects the load balancing in these scenarios. Looking at the points P2, P3 and P4 in Figure 6-3 and Figure 6-4 shows that load at these points is equally distributed among the available networks, however in Figure

6-5 and Figure 6-6 the same points show less load in satellite network and higher loads in other networks. This is due to the higher cost of satellite networks as compared to the other networks.
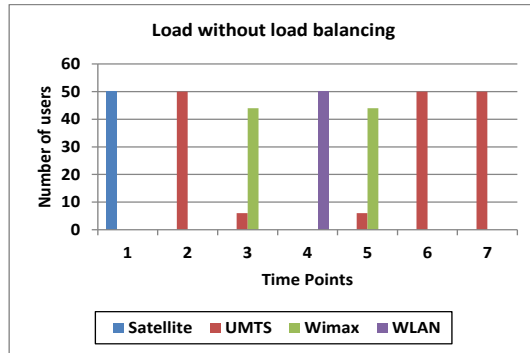
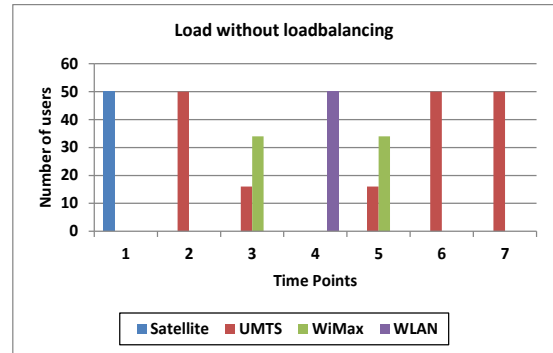Figure 6-1:Load distribution without load balancing (speed=25m/s)

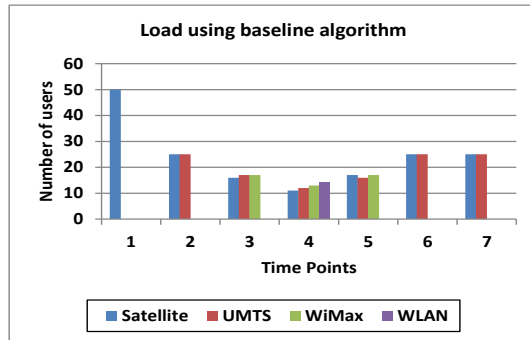Figure 6-2: Load distribution without load balancing (speed=2m/s)

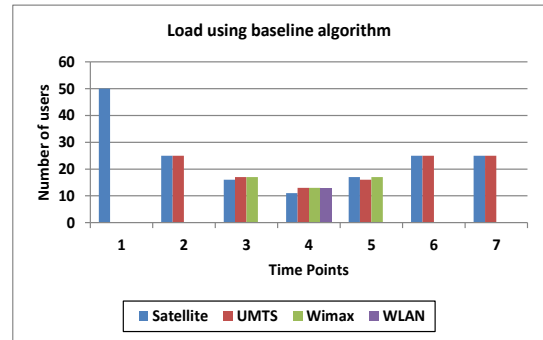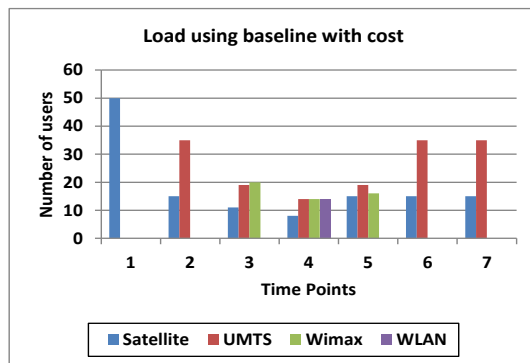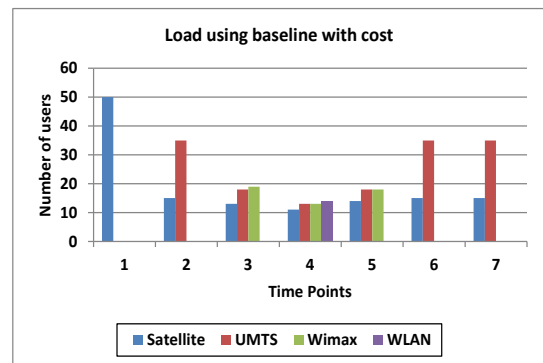Figure 6-3: Load distribution with baseline load balancing (speed=25m/s)

Figure 6-4: Load distribution with base line load balancing (speed=2m/s)

Figure 6-5: Load distribution with baseline with cost (speed=25m/s)

Figure 6-6: Load distribution with baseline with cost (speed=2m/s)

b) Packet drops

Figure 6-7 and Figure 6-8 shows the packets drop rate for this scenario. The packet drop rate represents the ratio of the number of packets dropped per

total packets transmitted. It can be seen from these figures that the packets drop rate when no load balancing is applied is the highest (0.2%). This is expected as most of the mobile nodes handover to the best available networks, which can cause congestion on that network thereby resulting in the large number of packet drops.

The packets drop rate is lower (0.15%) when the baseline load balancing algorithm is applied. This is because the networks are not overloading in this case. When cost preferences are also considered, the packet drops is (0.165%) which is slightly higher than the baseline case as in this case the load in networks are not perfectly balanced due the varying network service cost and mobile node preferences. It can also be seen from the comparison of graphs shown in Figure 6-7 and Figure 6-8 that as the speed is increased, the packet drop ratio also increases. The abbreviations used in the following graphs are as follows: NLB stands for "No Load Balancing" LL stands for "Least Loaded" and LLC stands for "Least Loaded with Cost".
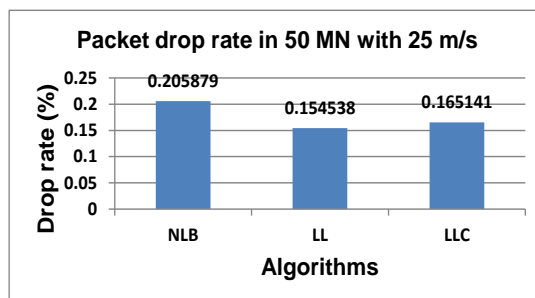
Figure 6-7: Packets drop rate (speed=25m/s)
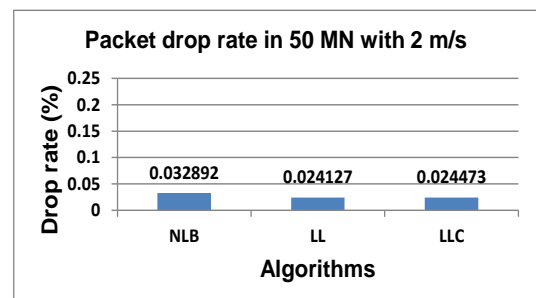
Figure 6-8: Total packet drops (speed=2m/s)

c) Handover Latency

Figure 6-9 shows the average handover latencies for this scenario. The x-axis in this graph represents the individual mobile nodes and y-axis represents the total handover latency observed by mobile nodes in seconds

for their complete journey. This latency is the sum of all the delays for the different handovers any given user would be subjected to during its movement across the travel path.
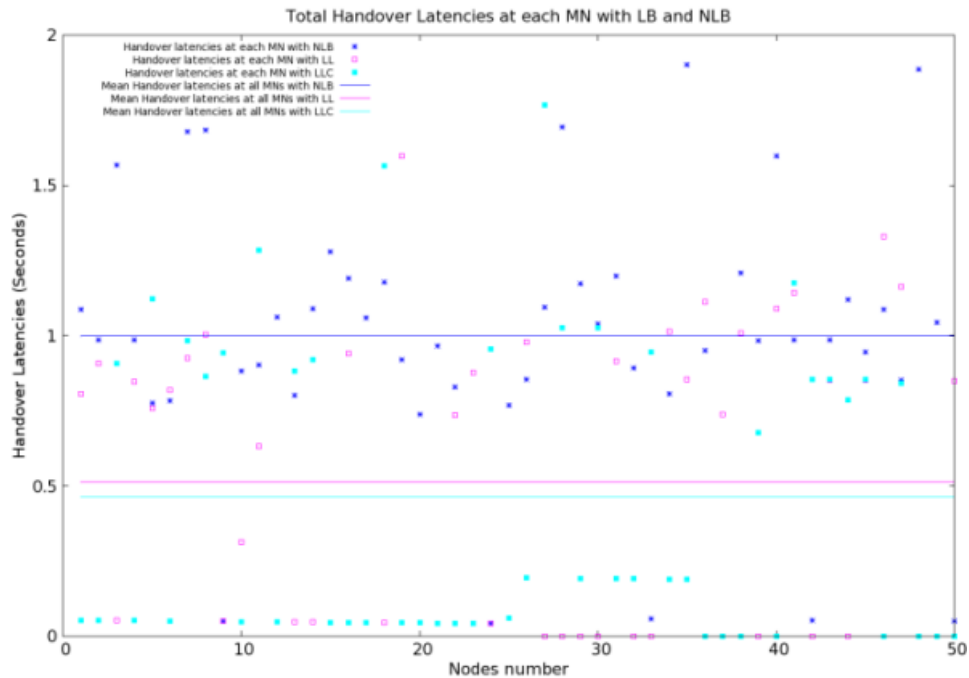

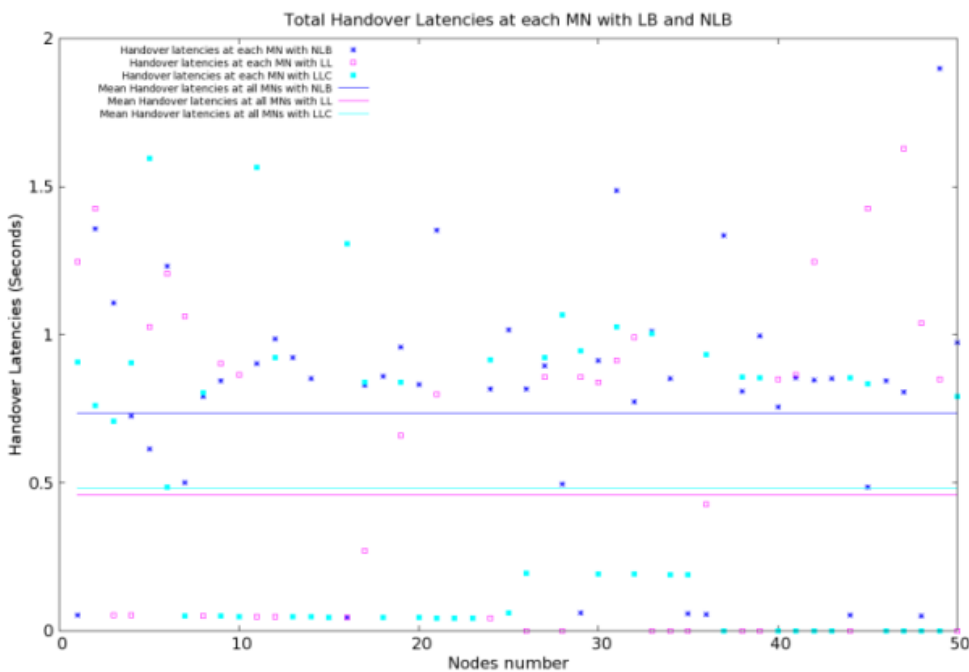Figure 6-9: Handover Latency (speed = 25m/s)

Figure 6-10: Handover Latency (speed = 2m/s)

The blue dots in Figure 6-9 represent the total handover latencies observed by each mobile node without load balancing. The pink dots represent the

total handover latency observed by each mobile node with baseline least-loaded load balancing algorithm and the cyan dots represents the handover latencies observed by each mobile node with the baseline least-loaded with cost algorithm. The blue, pink and cyan horizontal lines shown in this graph represent the mean of total handover latencies observed by all mobile nodes for the three cases.

It can be seen that the average handover latency for no load balancing is highest at around 1 sec, as in this case most of the mobile nodes handover to the best available network upon entering the common coverage areas. Hence a large number of handovers take place thereby resulting in this large overall delay. The mean values for the baseline load balancing with and without cost preferences are much lower than this at around 0.5 sec. This is due to the fewer handovers that take place in these cases. Similarly Figure 6-10 shows the handover latencies when the mobile nodes are travelling at 2m/s. It can be seen from this graph that without load balancing the handover latencies are highest and the handover latencies for baseline and baseline with cost are lesser as compared to the no load balancing scenario.

On comparing the graphs in Figure 6-9 and 6-10, we can also see that for the lower speed scenario the handover delays are lower. This is because as explained in the previous sub-section the packet drops are higher in the high speed scenario which also affects the handover procedure thereby requiring retransmissions of lost control messages during the handover process.

d) Average throughput at mobile node

Figures 6-11 and 6-12 represent the average throughput observed by each mobile node in this scenario where the speed of mobile node is 25 m/s and 2 m/s respectively.
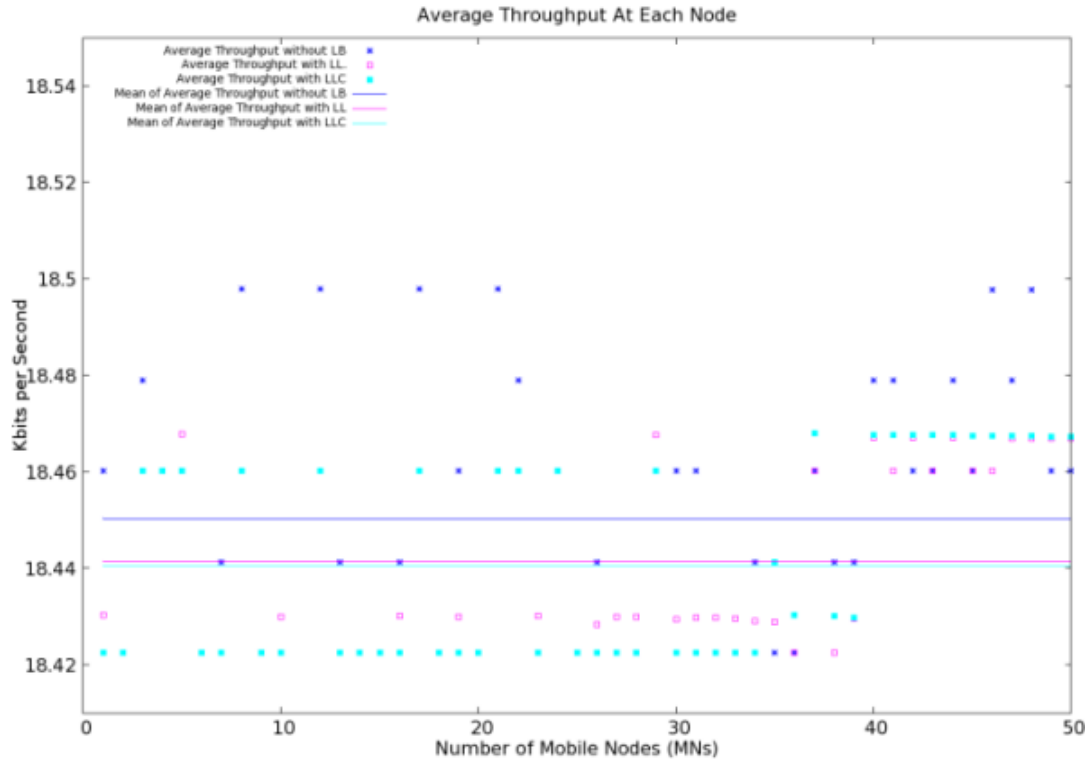

Figure 6-11: Average throughput (speed = 25 m/s)

The x-axis of the graphs in Figure 6-11 and Figure 6-12 represents the mobile nodes number and the y-axis represents the throughput in kilobits per second (kbps). The blue dots represents the values without load balancing, the pink dots represents the values with baseline least- loaded load balancing algorithm and the cyan dots represents the baseline with cost case. The average throughput for all the mobile nodes is shown as horizontal lines in respective colours. It can be seen from Figure 6-11 that the average throughput is higher in case of no load balancing at 18.45 kbps as the mobile nodes select the best available network. On other hand the average throughput for all the mobile nodes in case of baseline and baseline with cost

are almost similar at around 18.44 kbps and only slightly lower than that of no load balancing scenario. The reason for this is that load balancing tries to maintain the load equilibrium between the networks and this practice may result selection of network for some mobile node with high network latencies and lower data rate but only after making sure that the network can fulfil the required QoS of the mobile user.
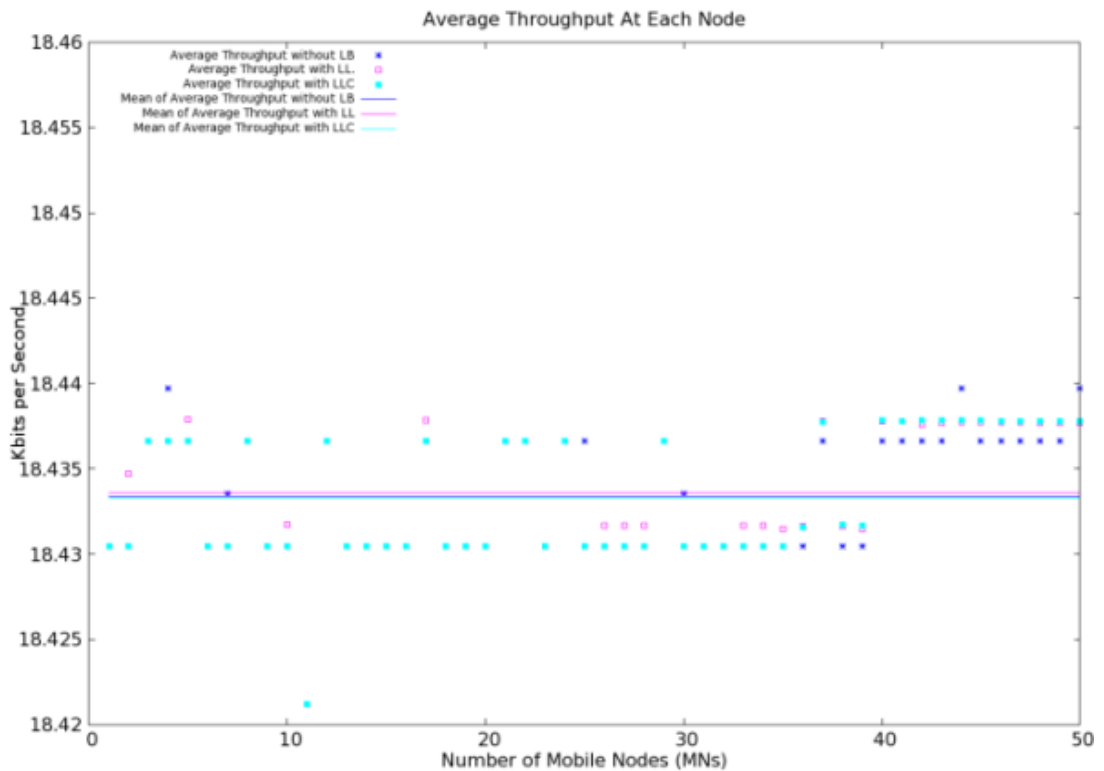

Figure 6-12: Average throughput (speed = 2 m/s)

Figure 6-12 shows that the average throughput is same in all the cases when moving at 2m/s. This shows that the use of load balancing does not really affect the throughput of the users. The main reason for closer average values is this scenario is the high total time of simulation due to the slow moving users. The scenario with low speed takes longer to travel the trajectory and therefore generates a large amount of traffic which causes congestion on the networks. This results in a slightly reduced mean values of average throughput of all the mobile nodes for different algorithms.

148

It can also be seen on comparing the two graphs that the average throughput is around 18.4335 kbps and 18.448 kbps when the users are moving at 2m/s and 25m/s, respectively. This shows that the use of load balancing does not really affect the throughput of the users who can still access their services properly while at the same time the loads across the networks are more uniformly balanced.

e) Network Throughput

Figure 6-13 to Figure 6-18 show the throughput at each network for this scenario. These graphs show how the networks are being utilised at different times. Figure 6-13, 6-14 and 6-15 show the throughput of all the networks when the nodes are moving with the speed of 25m/s without load balancing, with baseline or least loaded load balancing and with baseline algorithm using cost preferences from users respectively.

The graph in Figure 6-13 shows that all the nodes handover from satellite to the UMTS at time approximately 30 seconds during simulation. At approximately 51 seconds the mobile nodes enter the WiMax coverage area and leave the WiMax coverage area at time 90 seconds. The traffic in the WLAN network starts at approximately 66 seconds and ends at approximately 74 seconds. As without load balancing the mobile nodes handover to the best available network therefore the average throughput on each network shifts to the newly available better network, whenever the mobile nodes enter the network with low network latencies and high data rates. The graphs represent the number of TCP packets sent/received per unit time across the simulation.
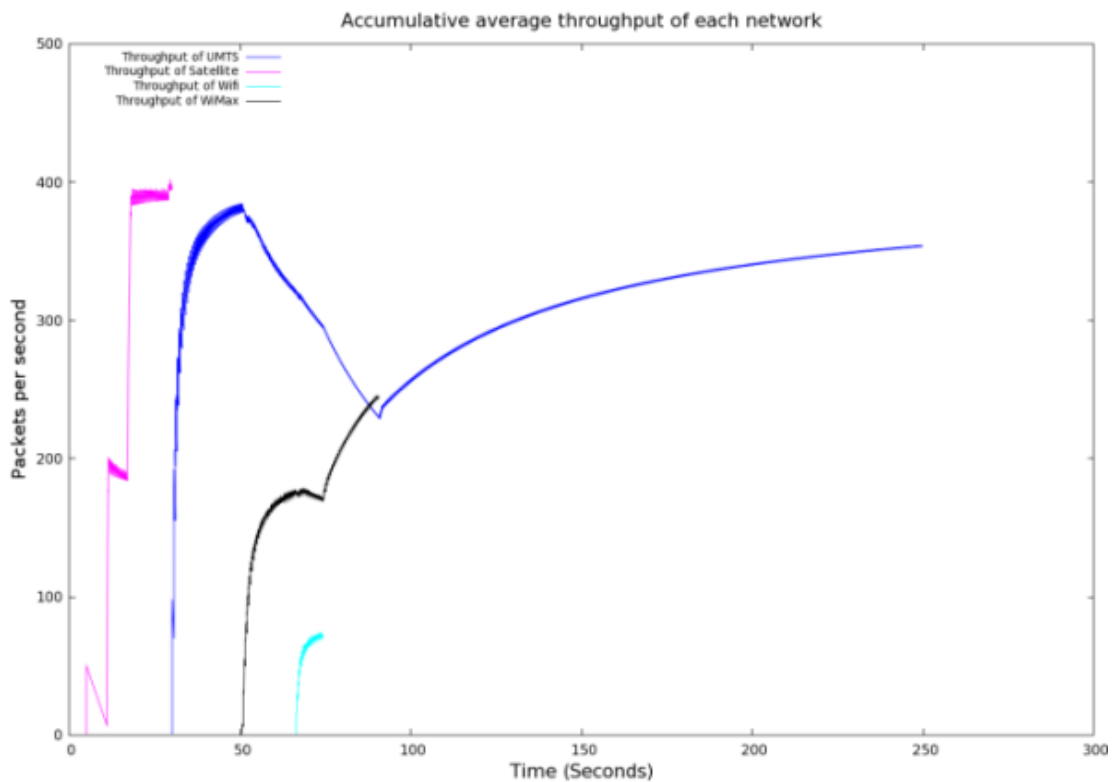
Figure 6-13: Network throughput with no load balancing (speed = 25m/s)

Figure 6-13 clearly shows that when no load balancing is used, in the beginning the throughput of the satellite network is 400 packets/second (1.84 Mbps) showing that the satellite link is heavily utilised as all users are on this link. Similarly at 30 seconds when mobile nodes enter the common coverage area of satellite and UMTS networks, the traffic shifts to UMTS and the UMTS throughput is approximately: 380 packets/second or (1.75Mbps).

When the users enter WiMax, we can see the network throughput for WiMax increases to around 180 packets/ (0.83Mbps) while the network throughput of UMTS network decreases. This is because there are still some users in the UMTS network. When the users enter WLAN coverage area, we can see from Figure 6-1 and 6-2 that all the user's move into WLAN. However the network throughput is only at around 80 packets/second (0.37Mbps); due to the high congestion in the network and resulting packet drops.
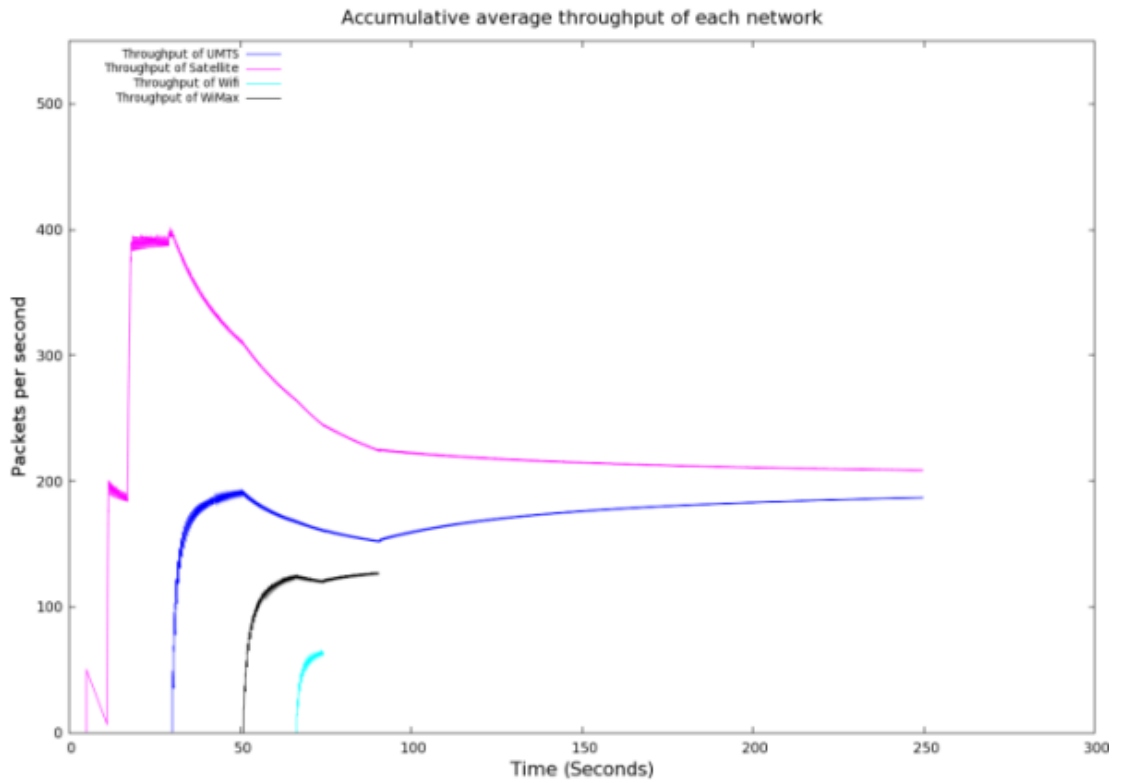
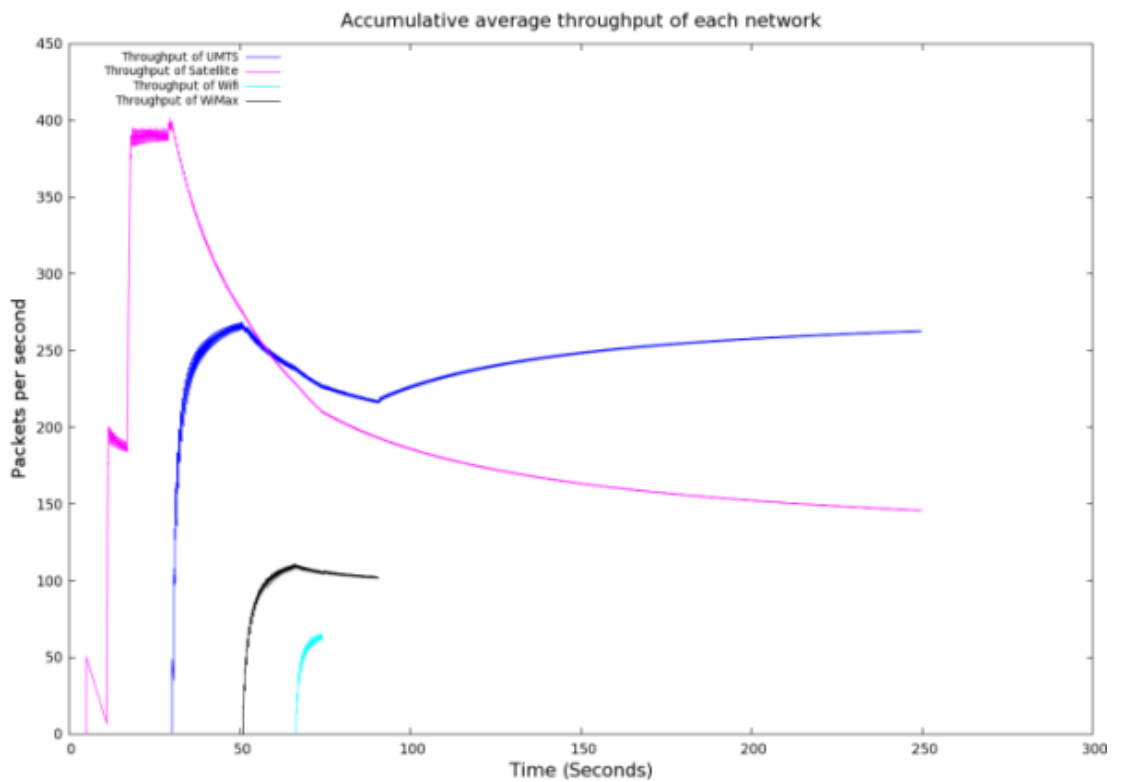Figure 6-14: Network throughput with baseline load balancing (speed = 25m/s)

Figure 6-15: Network throughput with baseline load balancing with cost (speed = 25m/s)

On the other hand Figure 6-14 shows the traffic on all networks when the

baseline least load balancing algorithm is applied. In this case until 30

seconds when all the mobile users are in satellite only coverage area, the satellite average throughput is approximately 400 packets/second. At time 30 seconds when the mobile nodes enter the common coverage area of satellite and UMTS networks the traffic in satellite network decreases and the traffic in UMTS network increases as the load is now shared between satellite and UMTS networks.

Comparison of Figure 6-13 and Figure 6-14 shows that when load balancing is applied the traffic in other networks is less (e.g.: UMTS is around 200 packets/second which earlier was around 380 packets/sec) as the satellite network shares the load with terrestrial networks throughout the simulation time. This shows the benefit of load balancing algorithm for sharing the load between networks to avoid the congestion situation.

Figure 6-15 represents the traffic in each network when load balancing with cost is applied. Comparing Figure 6-14 with Figure 6-15 shows the decreases in satellite traffic and increase in the other terrestrial networks traffic (e.g.: UMTS is around 250 packets/second which with no load balancing was around 380 packets/second but with baseline was around 200 packets/sec) due to the fact that most mobile nodes do not want to pay for satellite when they use terrestrial networks. This shows that the cost can degrade the efficiency of load balancing algorithms only slightly, but it is still far better than no load balancing.
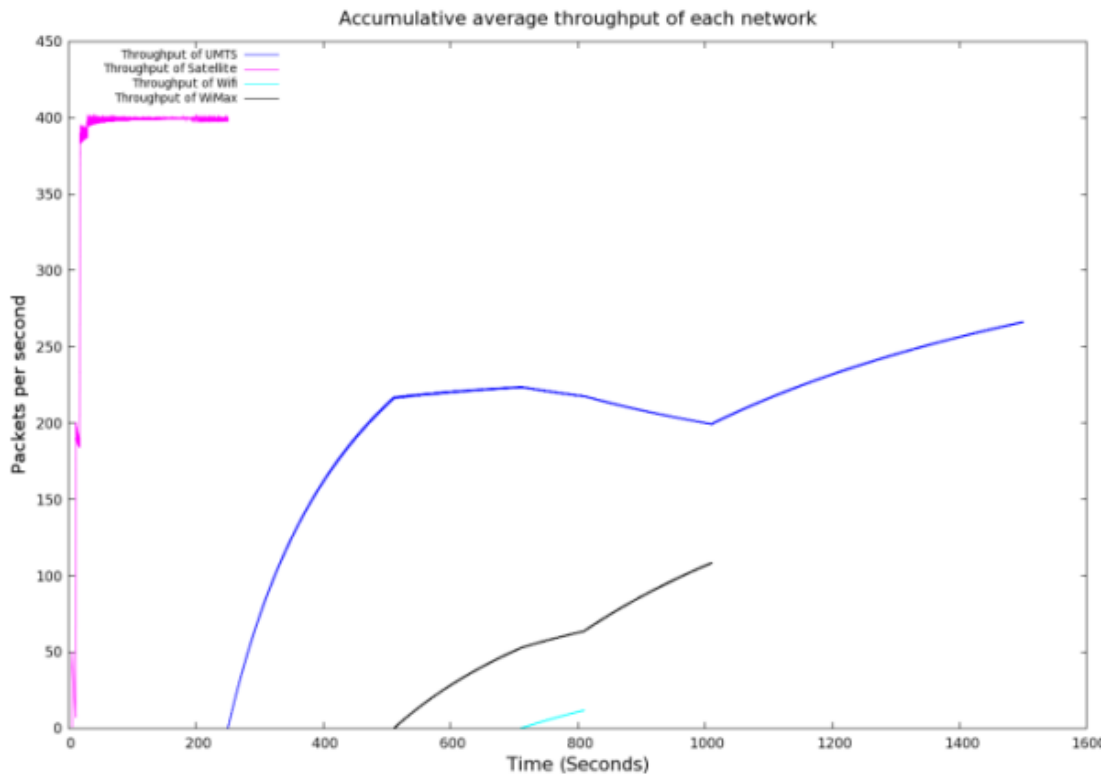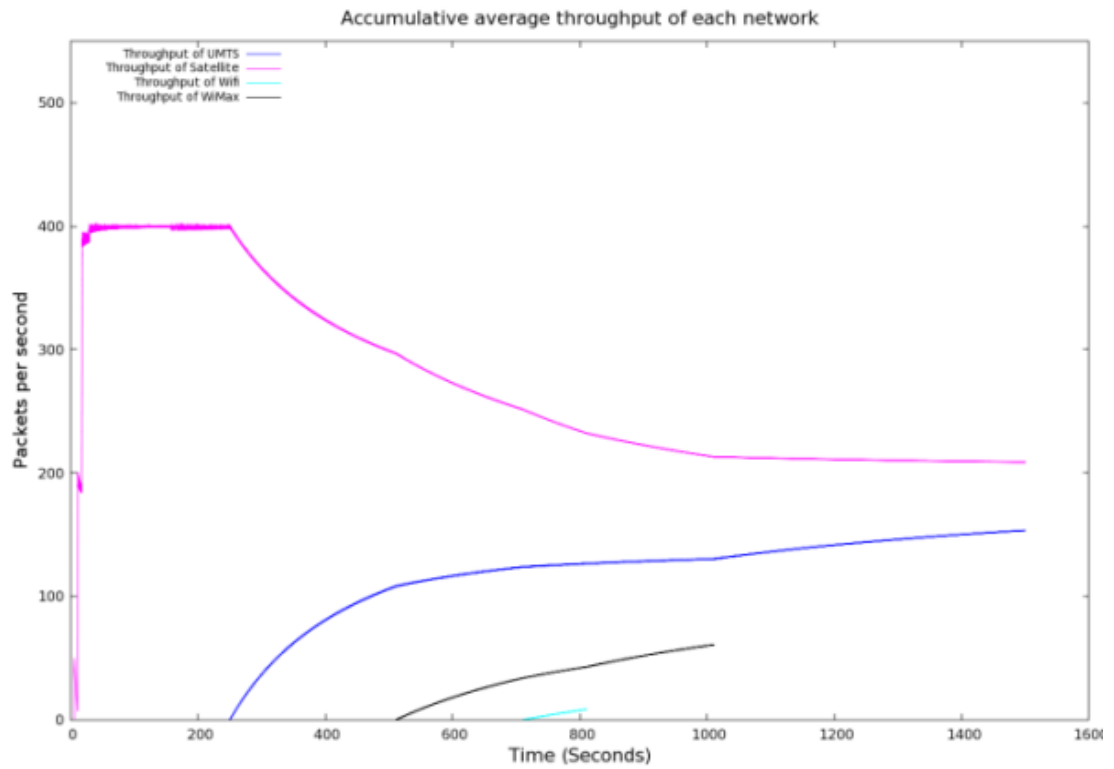
Figure 6-16: Network throughput with no load balancing (speed = 2m/s)

Similarly Figure 6-16, 6-17 and 6-18 represent the throughput in each network for the 2m/s scenario. Without load balancing the satellite network throughput is approximately 400 packets/second until 250 seconds. At 250 seconds the mobile nodes enter the common coverage area of satellite and UMTS and all the mobile nodes handover to the UMTS network making the throughput in UMTS to approximately 240 packets per second. At 511 seconds the mobile nodes enter the WiMax coverage area, therefore the traffic shifts from UMTS to WiMax making the WiMax throughput to approximately 70 packets/second and increasing. At 710 seconds the WLAN network appears in the trajectory of mobile nodes again the traffic shifts from WiMax to WLAN making the throughput at WLAN approximately 20 packets/second.

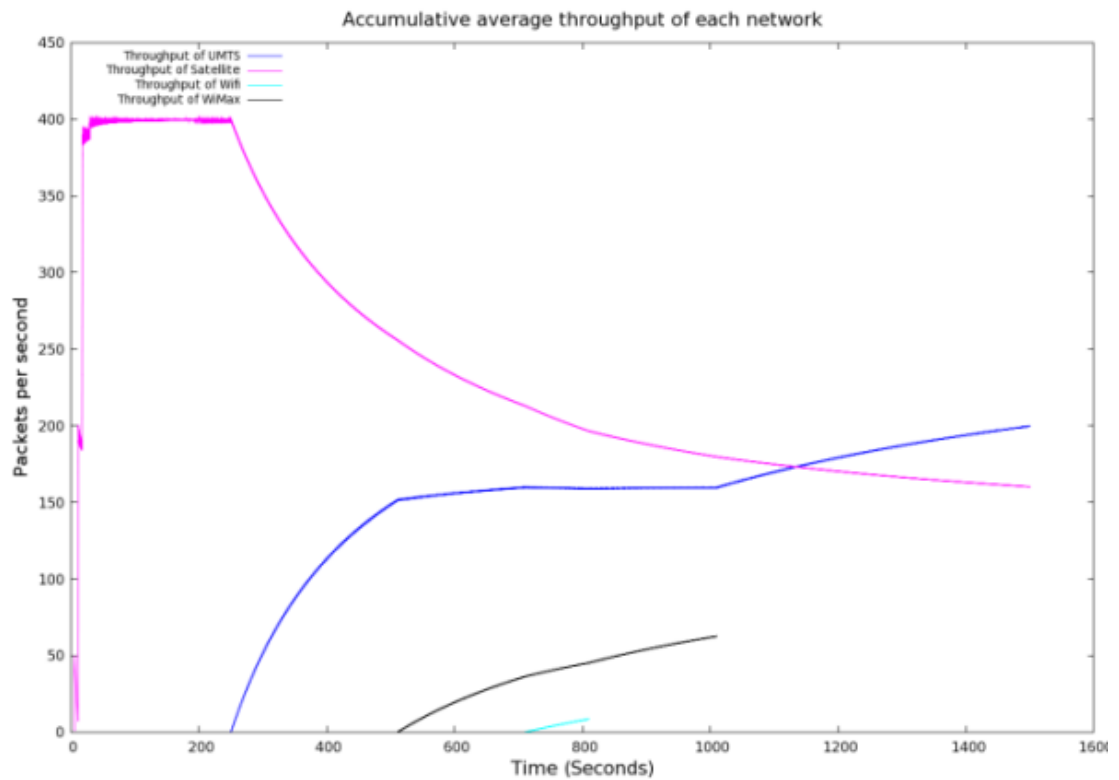Figure 6- 17: Network throughput with baseline load balancing (speed = 2m/s)

Figure 6-18: Network throughput with baseline load balancing with cost (speed = 2m/s)

At 810 seconds the mobile nodes leave WLAN coverage area shifting the traffic back to WiMax make the throughput on WiMax approximately 110 packets/second. At 1010 seconds mobile nodes leave the WiMax coverage area, shifting all the traffic to UMTS network making throughput at UMTS nearly 270 packets/second. In case of load balancing the traffic is shared between co-located networks as shown in Figure 6-17. It shows that when load is shared between satellite and terrestrial networks the average throughput at satellite, UMTS and WiMax network reduces to 200 packets/second, 180 packets/second and 70 packets/second. This reduction of traffic in each network by sharing the load between different available networks minimizes the chances of congestion and hence improves the performance. Figure 6-18 represents the average throughput in each network when load balancing is applied with cost. It shows the effects of cost on balancing the load in co-located wireless networks as the average throughput in satellite reduced to 150 packets/second and in UMTS and WiMax increased 220 packets and 80 packets approximately. This concludes that high network cost and users preferences towards using the inexpensive available networks may degrade the load balancing but the overall results with load balancing with cost are still improved as compared to the no load balancing. As the load balancing still share the traffic and avoids or minimizes the chances of congestion on the co-located networks using all available options.

6.2.2 Scenario 2 – Baseline Least-loaded algorithm with 100 users

Similarly for 100 mobile nodes scenarios different parameters such as network load, average throughput at each mobile node, total packet drops, throughput at each network and handover latencies have been monitored. This section briefly explains the results obtained from the 100 mobile nodes scenarios with mobile nodes velocity of 25m/s and 2m/s using no load balancing, least loaded load or baseline balancing and least loaded with cost algorithm. In case of cost preferences 30 mobile nodes are willing to stay and pay on satellite when they can see other network along with satellite but 70 mobile nodes do not want to pay for satellite when they can connect to the other terrestrial networks.

a) Network load

In 100 mobile nodes scenarios first set is simulated using high speed of 25m/s and the second set is simulated using low speed of 2m/s. The graphs shown in Figure 6-19 to 6-24 represent the load in each network such as satellite, UMTS, WiMax and WLAN at the different position of travel trajectory for the 100 mobile nodes scenarios with the mobile nodes moving at 25m/s and 2m/s.

These graphs show the load for the simulation scenarios when no load balancing is applied, baseline load balancing is applied and when baseline load balancing algorithm is applied with cost preferences. The x-axis of the graphs represents the selected time points where the load in each simulation

scenario is monitored. The y-axis of the each load graph represents the total load in terms of number of users in that particular network.
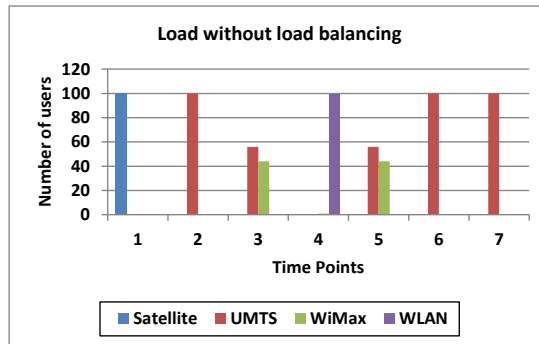

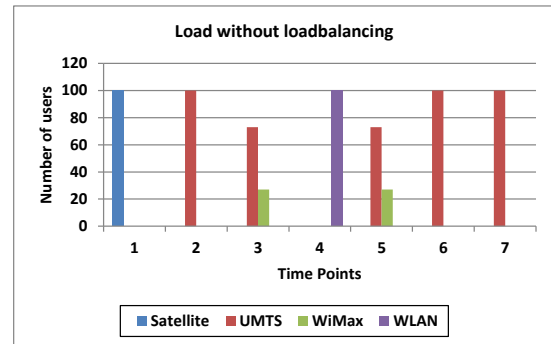Figure 6-19: Load distribution without load balancing (speed – 25 m/s)

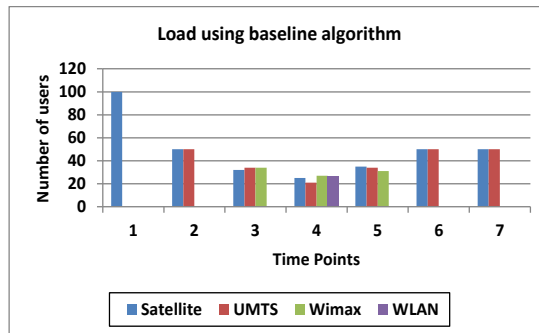Figure 6-20: Load distribution without load balancing (speed – 2 m/s)

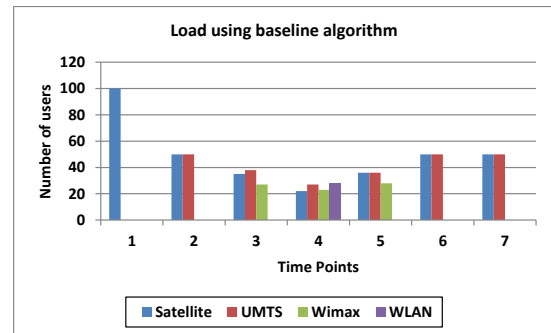Figure 6-21: Load distribution with baseline load balancing (speed–25 m/s)

Figure 6-22: Load distribution with baseline load balancing (speed–2 m/s)
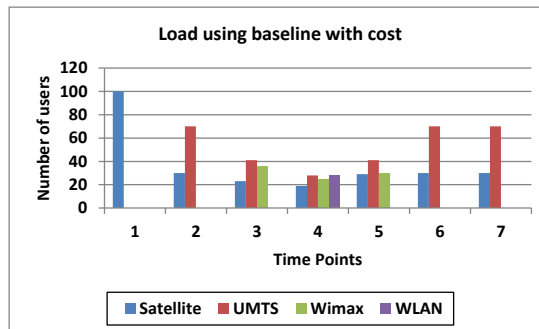

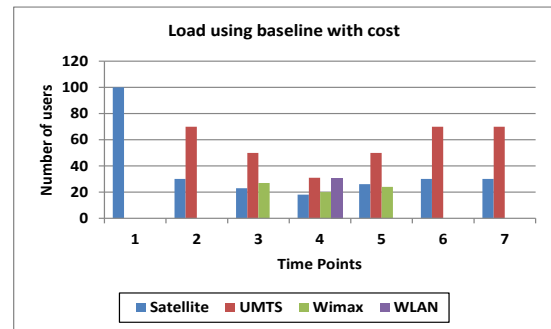Figure 6-23: Load distribution with baseline with cost (speed – 25 m/s)

Figure 6-24: Load distribution with baseline with cost (speed – 2 m/s)

It can be seen from these obtained load results that without load balancing most of the mobile nodes handover to the best available network in terms of cost and network latencies. For example in position 4 in Figure 6-19 and 6-20, we can see that all the users connected to WLAN. This however leaves the other networks under-loaded or underutilised. In the scenario where

baseline load balancing is applied all the networks share the load where possible such as in the overlapped coverage areas. For example in the same position 4 in Figure 6-21 and 6-22, we can see that the users are distributed across the different networks.

Figure 6-23 and Figure 6-24 represent the load in different networks when baseline load balancing is applied with cost preferences. It can be seen that the network cost affects the load balancing in these scenarios. Looking at the points P2, P3 and P4 in Figure 6-21 and Figure 6-22 shows that load at these points is equally distributed among the available networks, however in Figure 6-23 and Figure 6-24 the same points show less load in satellite network and higher loads in other networks. This is due to the higher cost of satellite networks as compared to the other networks.

b) Packet drops

The total packet drops in 100 mobile nodes scenarios are shown in Figure 6-25 and Figure 6-26 given as follows. These Figures represent that packet drops are higher in case where no load balancing is applied in the scenarios. The scenarios with baseline or least loaded load balancing suffer least packet drops and the scenarios using baseline with cost possess packet drops higher than baseline and lesser than no load balancing scenarios. The x-axis on graphs given in Figure 6-25 and Figure 6-26 represents the algorithms such as without load balancing, least loaded or baseline load balancing and baseline with cost. The abbreviations used in the following graphs are as follows: NLB stands for "No Load Balancing" LL stands for "Least Loaded" and LLC stands for "Least Loaded with Cost".
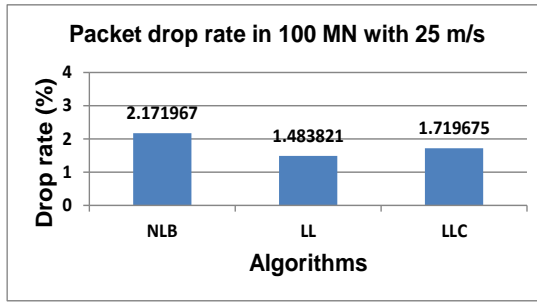
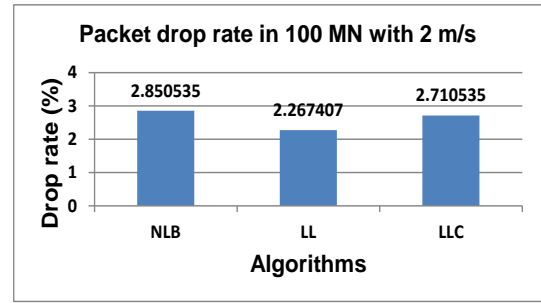Figure 6-25: Packets drop rate
(speed=25m/s)

Figure 6-26: Packets drop rate
(speed=2m/s)

The packet drop rate represents the ratio of the number of packets dropped per total packets transmitted. It can be seen from Figure 6-25 that the packets drop rate when no load balancing is applied is the highest (2.17%). This is expected as most of the mobile nodes handover to the best available networks, which can cause congestion on that network thereby resulting in the large number of packet drops.

The packets drop rate is lower (1.48%) when the baseline load balancing algorithm is applied. This is because the networks are not overloading in this case. When cost preferences are also considered, the packet drops is (1.71%) which is slightly higher than the baseline case as in this case the load in networks are not perfectly balanced due the varying network service cost and mobile node preferences. It can also be seen from the comparison of graphs shown in Figure 6-25 and Figure 6-26 that with 2mps in 100 mobile nodes scenario the packet drops are increased as the nodes stay longer in the networks and generate large amount of data causing congestion and ultimately resulting into higher drop rate.

c) Handover Latency

The total handover latencies observed by each mobile node in 100 mobile nodes scenarios with mobile nodes velocity of 25m/s and 2m/s are shown in

the graphs given in Figure 6-27 and Figure 6-28. The blue dot shown in Figure 6-27 and Figure 6-28 represent the total handover latencies observed by each node in 100 mobile nodes scenario with using no load balancing algorithm. The pink dot represents the total handover latencies by each mobile node using least loaded or baseline load balancing algorithm and the cyan dot represents total handover latencies using baseline load balancing with cost.
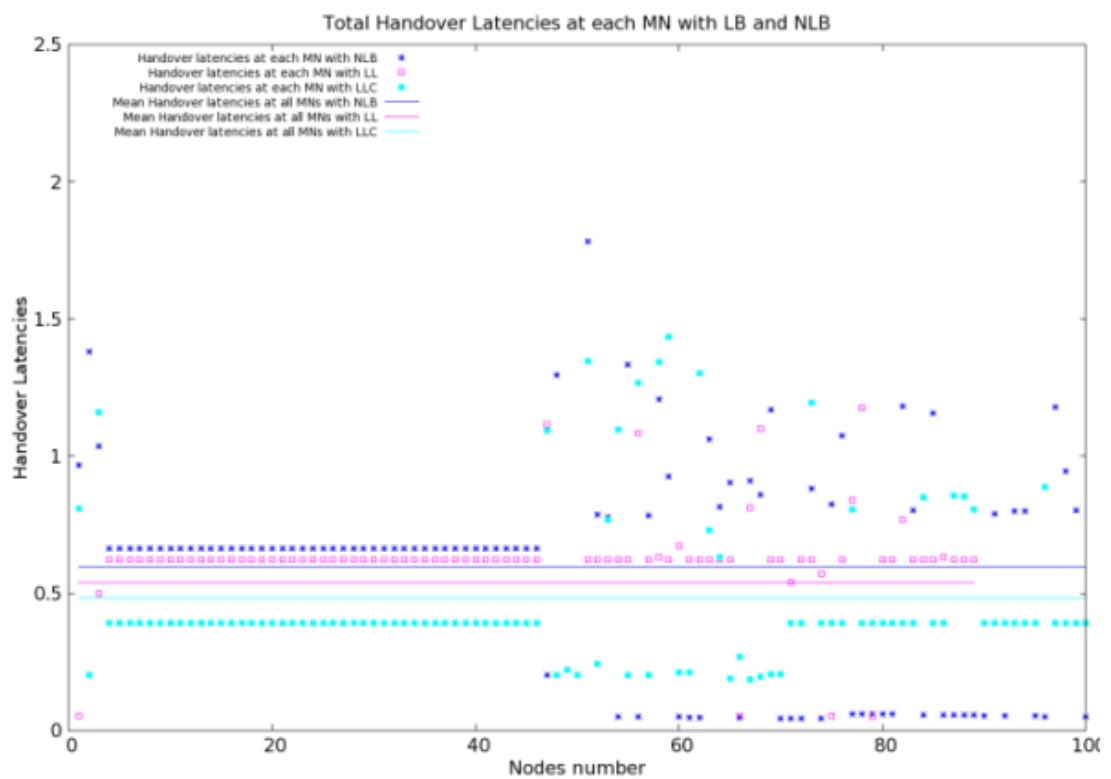

Figure 6-27: Handover Latency (speed - 25m/s)

The horizontal lines in Figure 6-27 and 6-28 represent the mean of the total handover latencies observed by each mobile node. Blue line shows the mean value for no load balancing, pink line shows the mean value for least loaded or baseline load balancing and cyan line shows the mean value for baseline load balancing with cost for the total handover latencies observed by each mobile node.
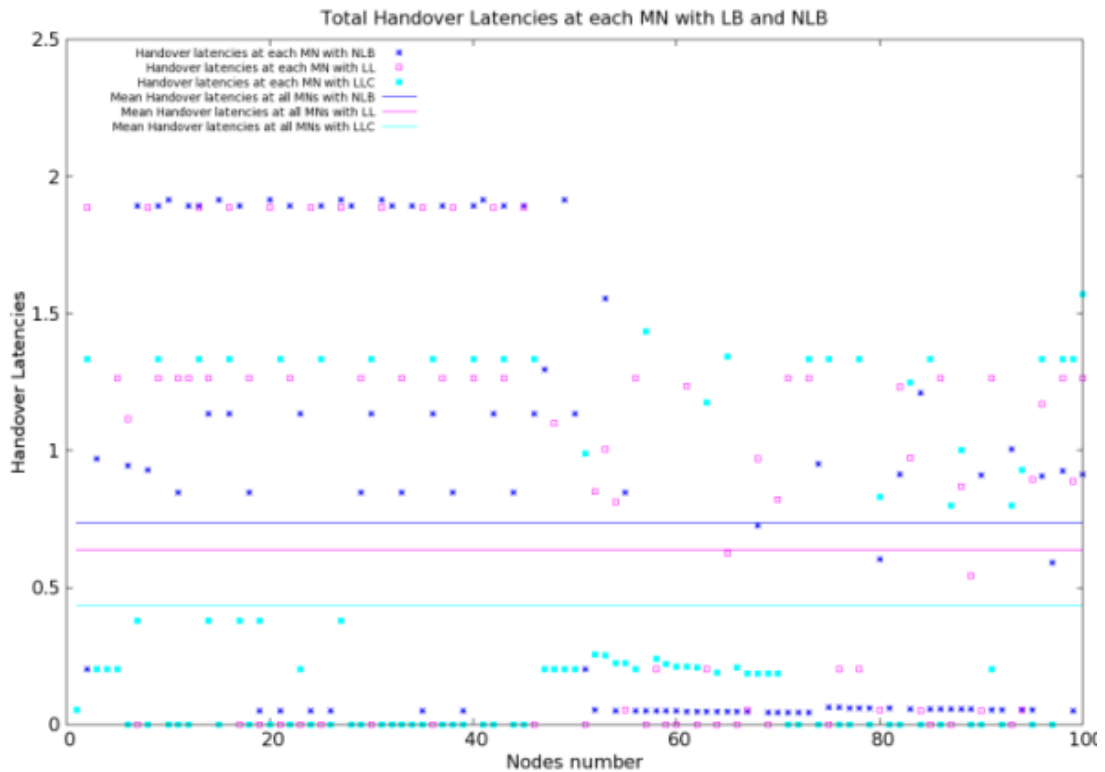
Figure 6-28: Handover Latency (speed - 2m/s)

The x-axis in this graph represents the individual mobile nodes and y-axis represents the total handover latency observed by mobile nodes in seconds for their complete journey. This latency is the sum of all the delays for the different handovers any given user would be subjected to during its movement across the travel path.

It can be seen in Figure 6-27 that the average handover latency for no load balancing is highest at around 0.6 second, as in this case most of the mobile nodes handover to the best available network upon entering the common coverage areas. Hence a large number of handovers take place thereby resulting in this large overall delay. The mean values for the baseline load balancing with and without cost preferences are much lower than this at around 0.55 second and 0.496 second. This is due to the fewer handovers that take place in these cases. Similarly Figure 6-28 shows the handover

latencies when the mobile nodes are travelling at 2m/s. It can be seen from this graph that without load balancing the handover latencies are highest and the handover latencies for baseline and baseline with cost are lesser as compared to the no load balancing scenario.

On comparing the graphs in Figure 6-27 and 6-28, we can also see that for the lower speed scenario the handover delays are lower. This is because as explained in the previous sub-section the packet drops are higher in the high speed scenario which also affects the handover procedure thereby requiring retransmissions of lost control messages during the handover process. The baseline with cost has least mean value for the total handover latencies observed by each mobile node. The reason for this is that in case of no load balancing all the mobile nodes perform handover whenever they detect better network or when they leave the coverage are of that network. However in case of baseline load balancing only a selected set of mobile users perform the handover in order to maintain the load equilibrium between the networks having common coverage areas or overlapping coverage areas. In case of baseline with cost the number of handovers are further reduced as most of the mobile nodes do not want to go back to satellite network when the terrestrial network is available.

d) Average throughput at mobile node

The average throughput of each mobile node and their mean values for the 100 mobile nodes scenarios with mobile nodes velocity of 25m/s and 2m/s are shown in Figure 6-29 and Figure 6-30. The blue colour dot represents the average throughput of each mobile node using no load balancing. The pink colour dot represents the average throughput of each mobile node using

baseline load balancing algorithm and the cyan dot represents the average throughput of each mobile node using baseline load balancing with cost. The blue, pink and cyan colour horizontal lines represent the mean values of the average throughput observed by each mobile node using no load balancing, baseline load balancing and baseline load balancing with cost. It can be seen from Figure 6-29 that the average throughput is higher in case of no load balancing at 18.45 kbps as the mobile nodes select the best available network.
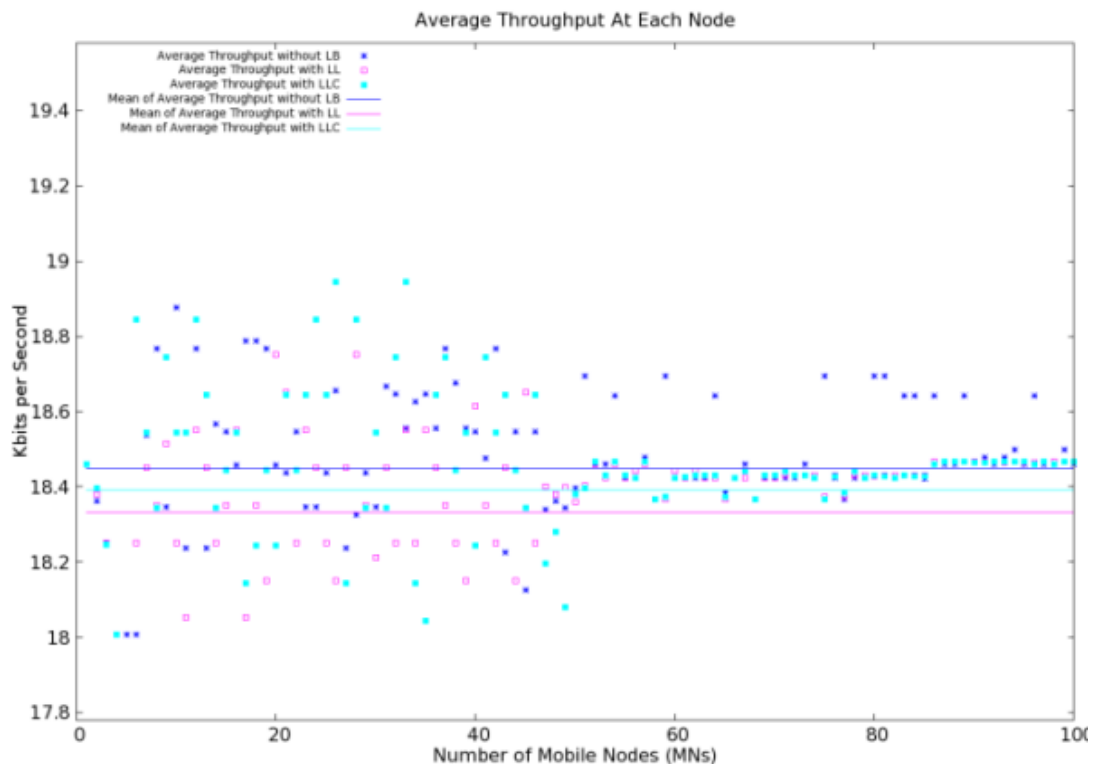

Figure 6-29: Average throughput (speed – 25 m/s)

On other hand the average throughput for all the mobile nodes in case of baseline and baseline with cost are almost similar at around 18.44 kbps and only slightly lower than that of no load balancing scenario. The reason for this is that load balancing tries to maintain the load equilibrium between the networks and this practice may result selection of network for some mobile

node with high network latencies and lower data rate but only after making sure that the network can fulfil the required QoS of the mobile user.
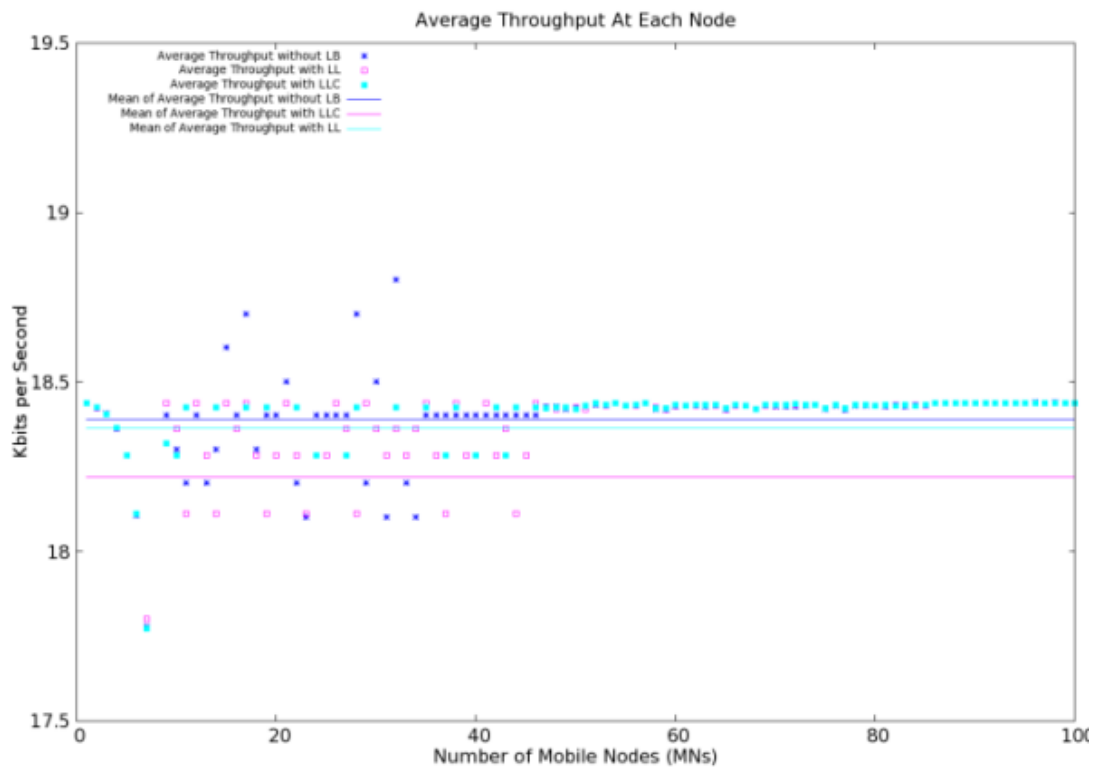

Figure 6-30: Average throughput (speed – 2 m/s)

Figure 6-30 shows that the average throughput is similar at (18.38 & 18.36 Kbps) in cases of no load balancing and load balancing with cost and with base line load balancing at (18.23 Kbps) when moving at 2m/s. This shows that the use of load balancing does not really affect the throughput of the users. The main reason for closer average values is this scenario is the high total time of simulation due to the slow moving users. The scenario with low speed takes longer to travel the trajectory and therefore generates a large amount of traffic which causes congestion on the networks. This results in a slightly reduced mean values of average throughput of all the mobile nodes for different algorithms. It can also be seen on comparing the two graphs that the average throughput is around 18.3 kbps and 18.4 kbps when the users

are moving at 2m/s and 25m/s, respectively. This shows that the use of load balancing does not really affect the throughput of the users who can still access their services properly while at the same time the loads across the networks are more uniformly balanced.

e) Network Throughput

The throughput of all the networks such as satellite, UMTS, WiMax and WLAN is shown in Figure 6-31, Figure 6-32, and Figure 6-33 for the 100 mobile nodes scenarios with mobile modes velocity of 25m/s. The x-axis on these figures represents the simulation time and the y-axis represents the throughput such as packets per second. Pink line represents the average throughput for the satellite network, blue line represents the average throughput for UMTS network, black line represents the average throughput for WiMax network and cyan line represents the average throughput for WLAN.
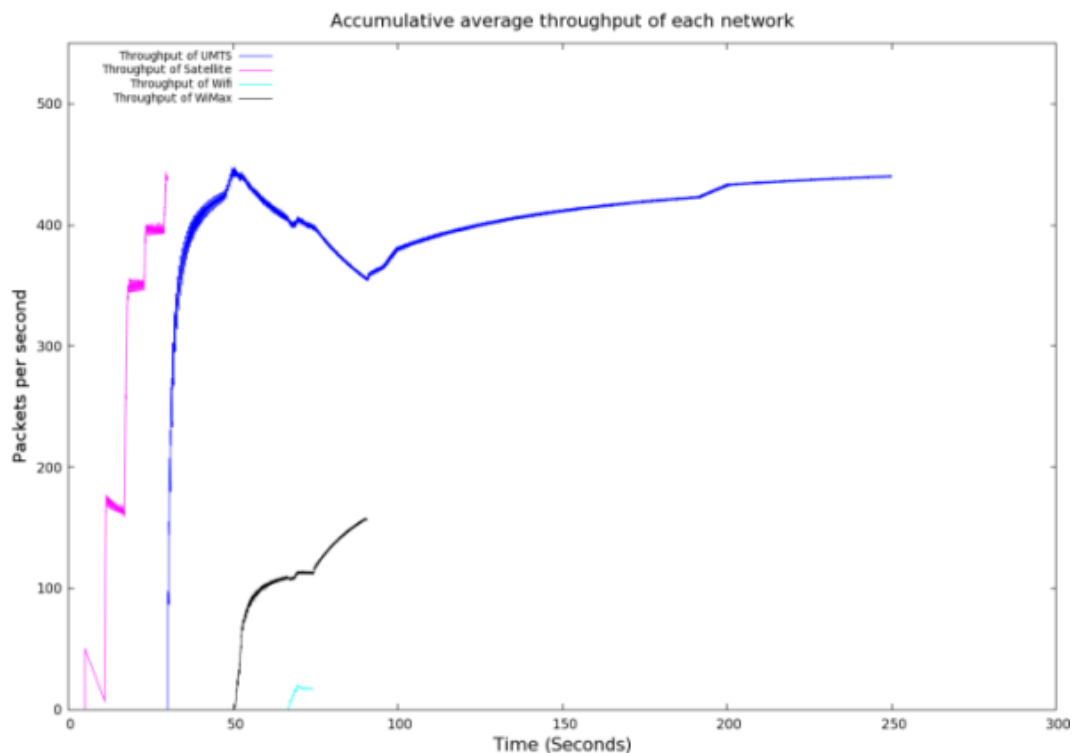

Figure 6-31: Network throughput with no load balancing (speed = 25m/s)
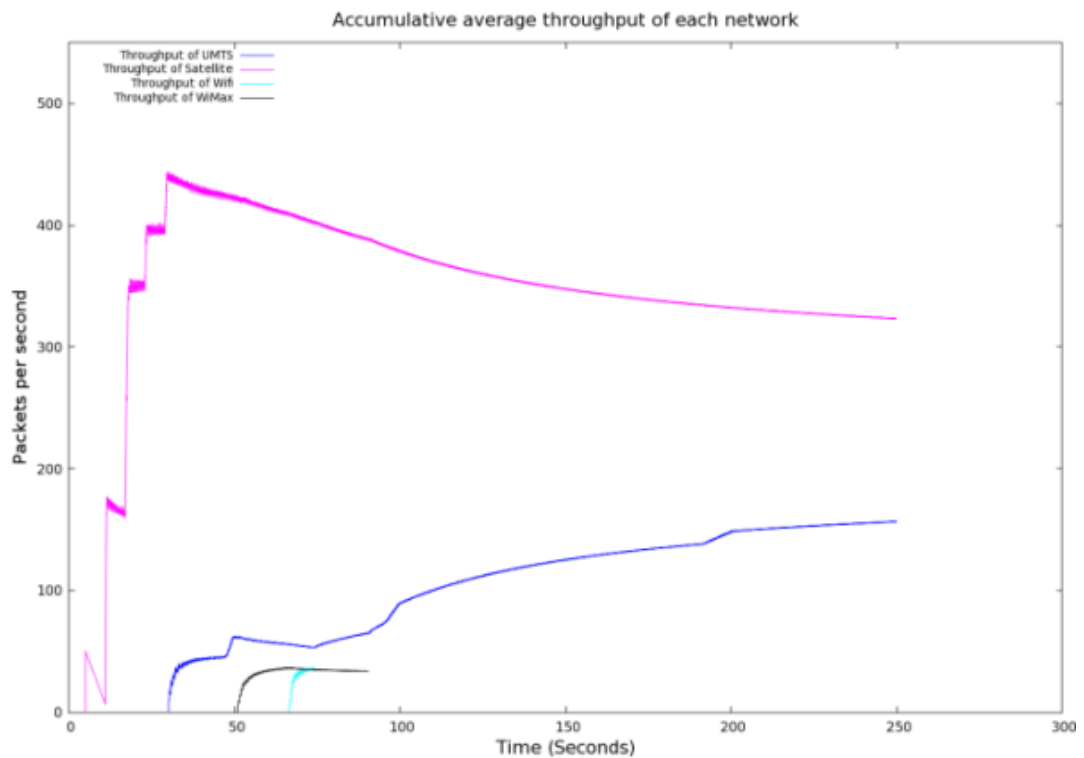
Figure 6-32: Network throughput with baseline load balancing (speed = 25m/s)

Figure 6-31 shows the average throughput of different networks for 100 mobile nodes scenarios using no load balancing with mobile nodes velocity of 25m/s. It shows that with no load balancing traffic is shifted to the best available networks when the mobile nodes enter or leave coverage areas or detect a new network. Figure 6-32 represents the average throughput of different networks when baseline load balancing is applied. In this case the traffic is partially shifted to the newly detected networks in order to maintain the load equilibrium between different networks having common coverage area or overlapped coverage area. In this way all the networks are being utilized on availability.

Figure 6-33 shows the average throughput of different networks when baseline load balancing with cost is applied. In this case the satellite network serves less number of mobile nodes when the mobile nodes are in the common coverage area of satellite and other terrestrial networks due to the

cost. Therefore the load in UMTS network goes comparatively higher when mobile nodes are in common coverage area of UMTS and satellite networks. This shows a considerable growth in average throughput of UMTS and for satellite it shows lower average throughput as compare to the baseline load balancing scenario shown in Figure 6-32.
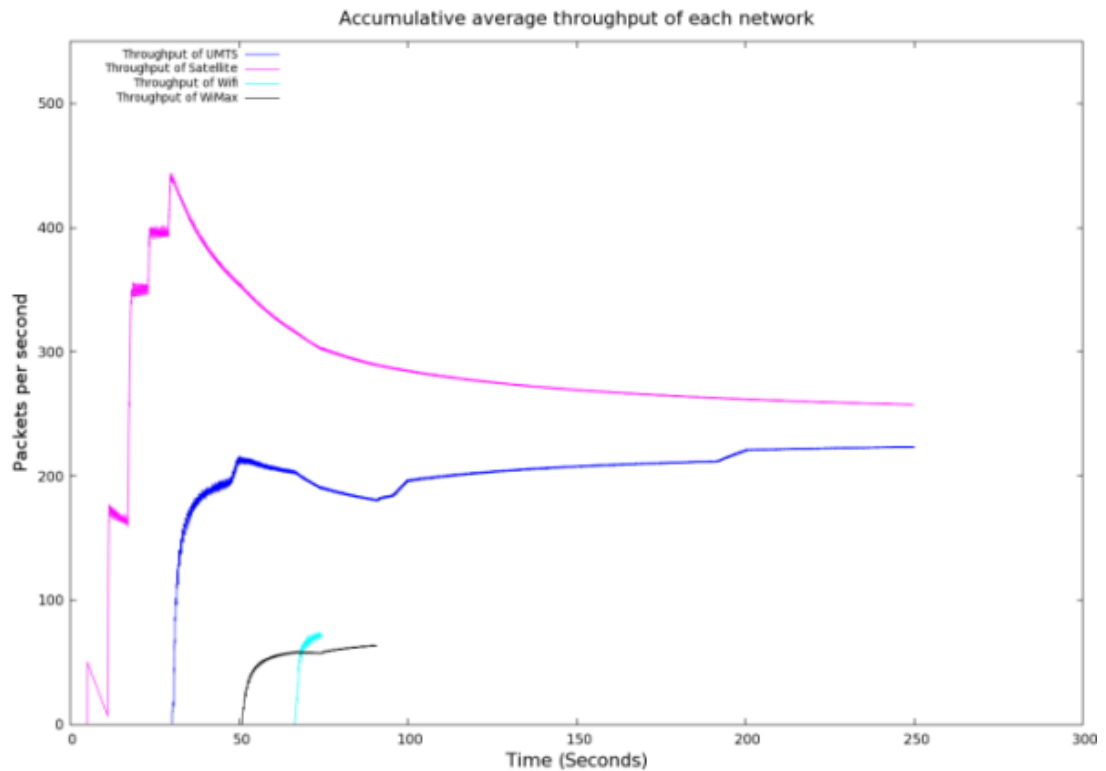

Figure 6-33: Network throughput with baseline load balancing with cost (speed = 25m/s)

For 100 mobile nodes scenarios with mobile nodes velocity of 2m/s, the following Figures from Figure 6-34 to Figure 6-36 are showing the average throughput at all the networks.
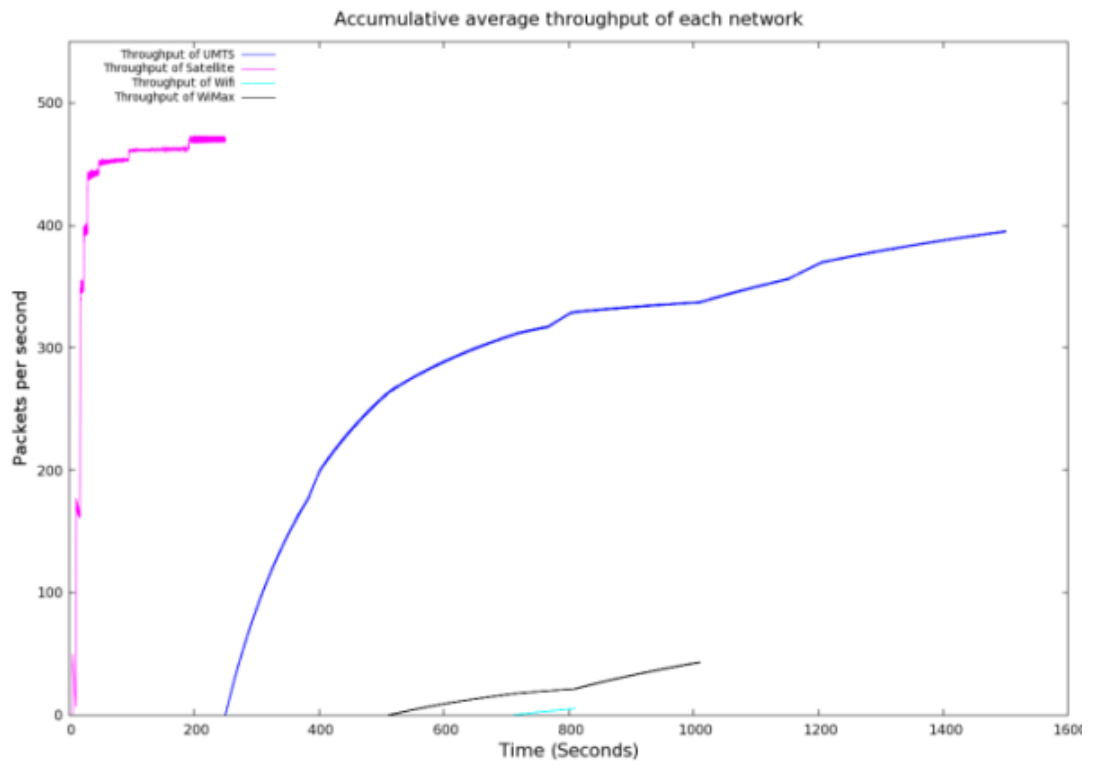
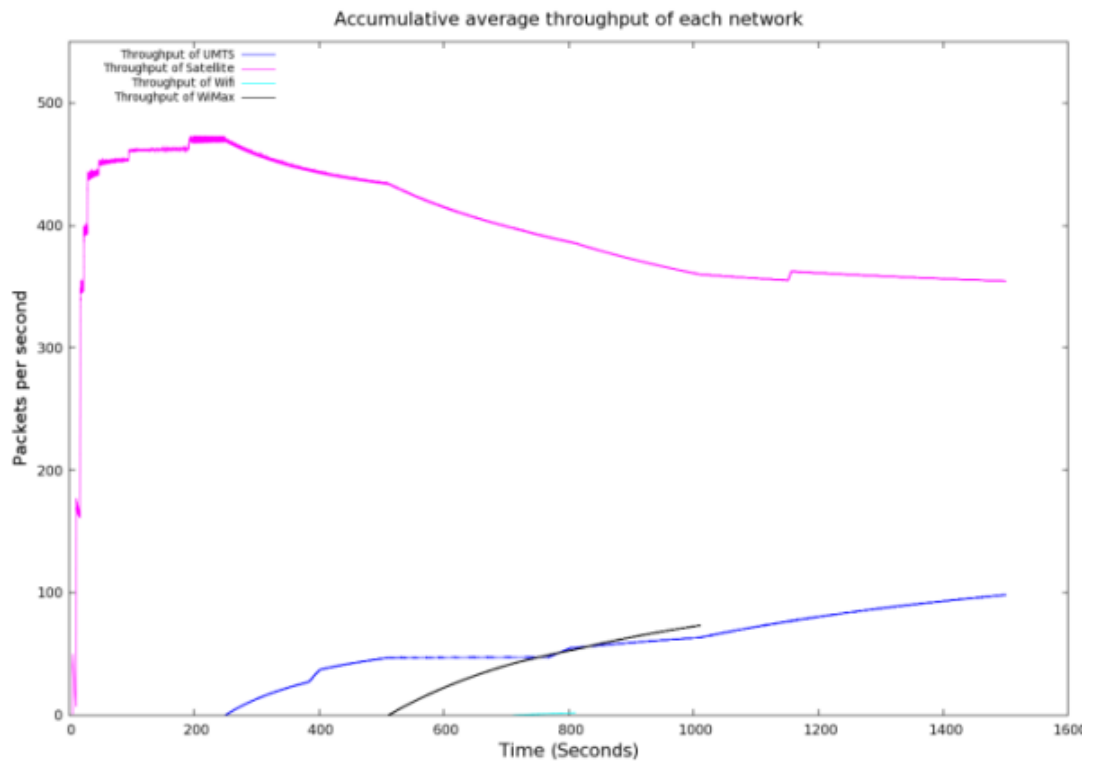Figure 6-34: Network throughput with no load balancing (speed = 2m/s)

Figure 6-35: Network throughput with baseline load balancing (speed = 2m/s)
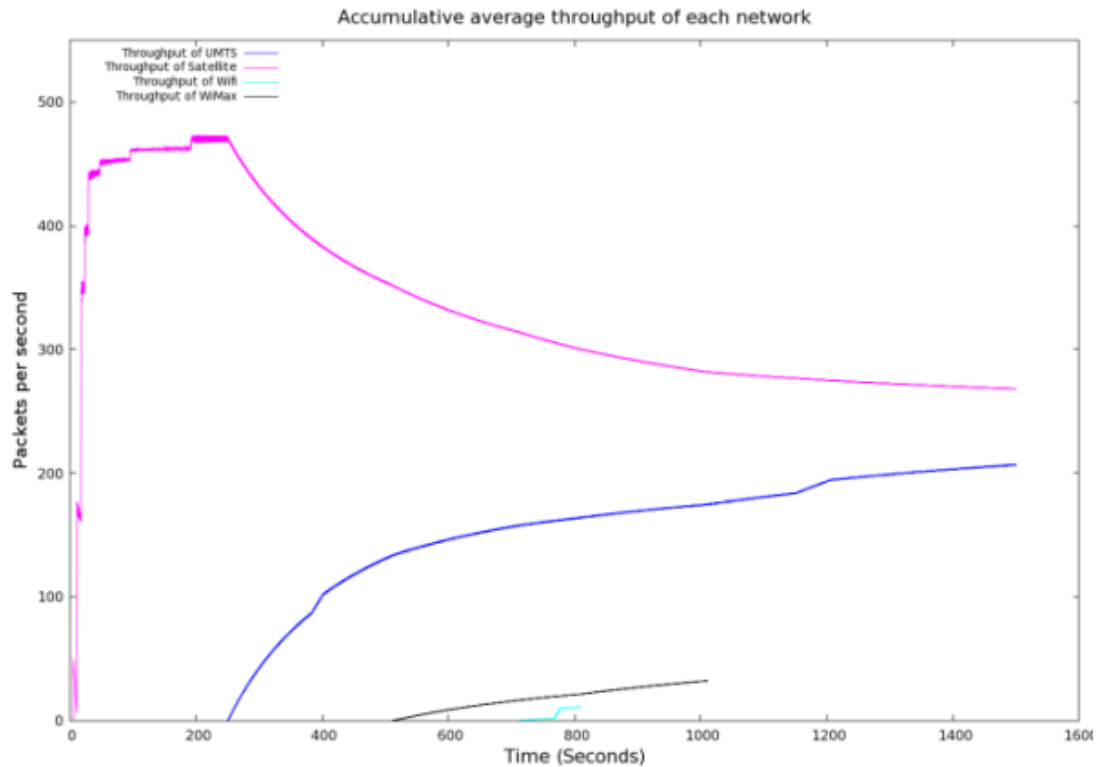
Figure 6-36: Network throughput with baseline load balancing with (speed = 2m/s)

Figure 6-34 shows the average throughput at all the networks using mobile node at velocity of 2m/s without load balancing algorithm. The comparison of Figure 6-34 and Figure 6-35 shows that in 2m/s scenario the average throughput for satellite is higher. This is because all the mobile nodes spend more time in the satellite network in the beginning due to low speed. As the mobile nodes enter the common coverage areas of different networks they handover to the best available networks.

Figure 6-35 shows the average throughput of all the networks when baseline load balancing is applied. This shows that average throughput of the satellite network is higher as all the mobile nodes spend more time in the satellite only coverage area at the beginning. When the mobile nodes enter the common coverage area of UMTS and WiMax they share the load to some extent but not uniformly.

Figure 6-36 shows the average network throughput for different networks using baseline load balancing with cost. This scenario assumes that 70 mobile nodes do not want to use satellite network if they can use the other terrestrial networks. For the first 250 seconds in this scenario all the mobile nodes stay in the satellite network as they cannot use any other network in that area. Once they see the UMTS coverage area 70 mobile nodes handover to the UMTS. Further ahead in the simulation when the mobile nodes move towards the common coverage area with other networks, it is allowed to move mobile nodes to other networks from satellite networks to balance the load but no mobile node is handover to the satellite if the mobile node prefer terrestrial network. This is the reason that average throughput of satellite network decreases in this scenario and that of UMTS increases as UMTS servers all the other mobile nodes which do not prefer the satellite in UMTS satellite coverage area.

6.2.3 Scenario 3 – Fuzzy based algorithm with 50 users

After applying the fuzzy based load balancing algorithm on 50 mobile nodes scenario with velocity of 25m/s and 2m/s, following results are obtained.

a) Network load

The graphs shown in Figure 6-37 to 6-42 represent the load in each network such as satellite, UMTS, WiMax and WLAN at the different position of travel trajectory for the 50 mobile nodes scenarios with the mobile nodes moving at 25m/s and 2m/s.

These graphs show the load for the simulation scenarios when no load balancing is applied, fuzzy load balancing is applied and when fuzzy load balancing algorithm is applied with cost preferences. The x-axis of the graphs

represents the selected time points where the load in each simulation scenario is monitored. The y-axis of the each load graph represents the total load in terms of number of users in that particular network.
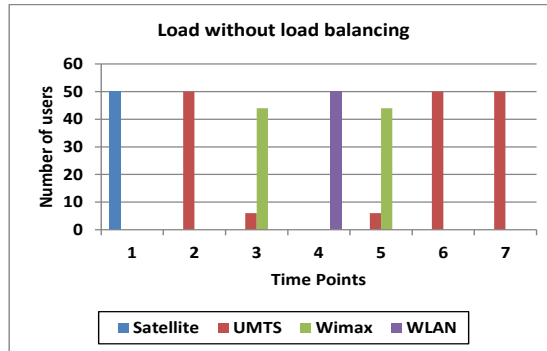

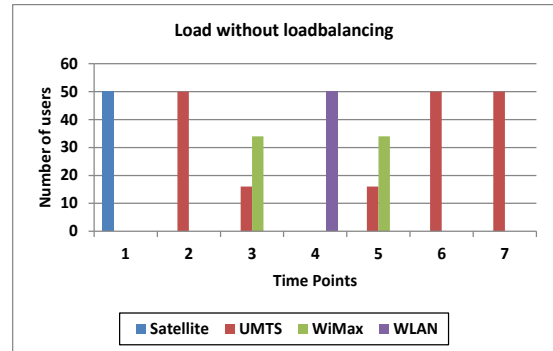Figure 6-37: Load distribution without load balancing in 25m/s

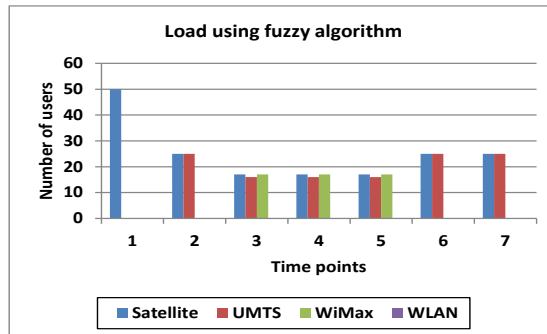Figure 6-38: Load distribution without load balancing in 2m/s

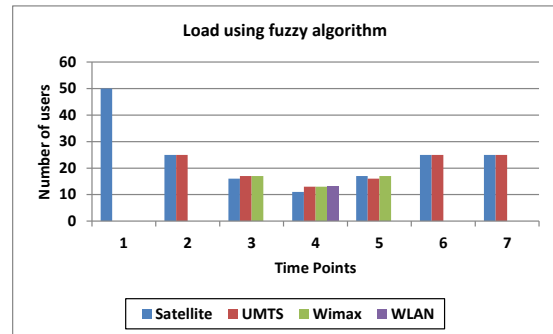Figure 6-39: Load distribution with fuzzy load balancing in 25m/s

Figure 6-40: Load distribution with fuzzy load balancing in 2m/s

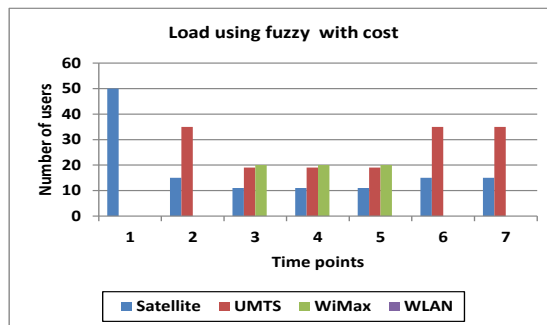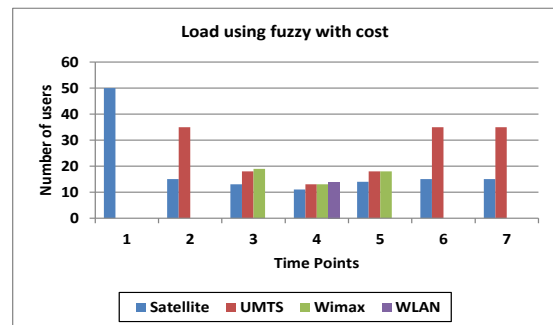Figure 6-41: Load distribution with fuzzy using cost in 25m/s

Figure 6-42: Load distribution with fuzzy using cost in 2m/s

Load on each network shown in the graphs from Figure 6-39 to Figure 6-42 reflect that the fuzzy based load balancing algorithm distributes the load efficiently. One unique difference in results obtained by fuzzy is that in 25m/s scenarios the fuzzy algorithm does not move mobile nodes to the WLAN

network and the load in WLAN remains zero throughout the simulation in 25m/s scenarios.

The reason for this is that, the fuzzy load balancing algorithm intelligently detects that velocity of mobile node is higher and the coverage area of the network is smaller, therefore it is not suitable to handover these mobile nodes to the WLAN. Hence the load in 25m/s scenarios is divided into satellite, UMTS and WiMax networks. On other hand in scenarios with 2m/s mobile nodes velocity, the mobile nodes could stay in WLAN for a considerable amount of time therefore fuzzy load balancing allows the mobile nodes to handover to WLAN.

It can be seen from these obtained load results that without load balancing most of the mobile nodes handover to the best available network in terms of cost and network latencies. For example in position 4 in Figure 6-37 and 6-38, we can see that all the users connected to WLAN. This however leaves the other networks under-loaded or underutilised. In the scenario where fuzzy load balancing is applied all the networks share the load where possible such as in the overlapped coverage areas. For example in the same position 4 in Figure 6-39 to 6-42, we can see that the users are distributed across the different networks.

Figure 6-41 and Figure 6-42 represent the load distribution in different networks using fuzzy load balancing with cost preferences from different mobile nodes. In these two figures there are two changes; one that with 25m/s the fuzzy algorithm does not allow the mobile nodes to handover to the WLAN and the other is that the satellite network has lower load. The

reason why satellite network is having lower load is that most mobile nodes in the terrestrial coverage area do not prefer satellite network. Therefore the load in the UMTS network goes higher.

b) Packet drops

The total packets drop rate in the 50 mobile nodes scenarios with mobile nodes velocity of 25m/s and 2m/s using the fuzzy, fuzzy with cost and no load balancing algorithms are shown in Figure 6-43 and Figure 6-44 below:
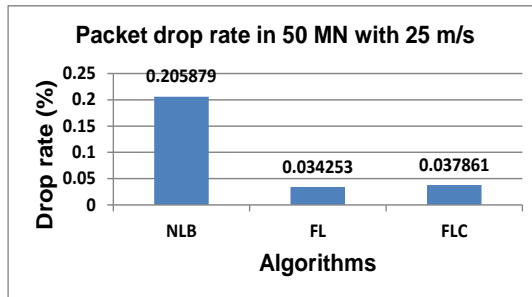

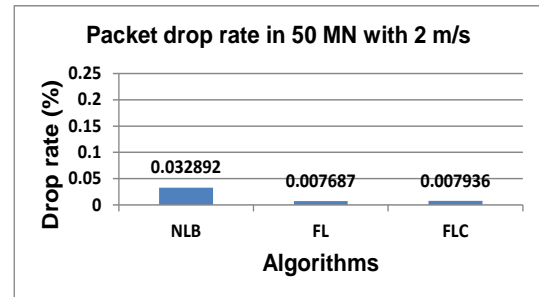Figure 6-43: Total packet drops with 25m/s

Figure 6-44: Total packet drops with 2m/s

The abbreviations used in these graphs are as follows: NLB stands for "No Load Balancing" FL stands for "Fuzzy Logic" and FLC stands for "Fuzzy Logic with Cost". The packet drops in fuzzy load balancing algorithm is considerably low as compared to the scenarios without load balancing and fuzzy load balancing with cost. As Figure 6-43 shows that packet drop rate using fuzzy load balancing algorithm is 0.034% and with fuzzy using cost is 0.037% both of which are fairly less than that the no load balancing scenario which is 0.2%. The cause for higher drops in scenario without load balancing is that, this algorithm moves all the mobile nodes to the newly detected better network by means of network latency, data rate and signal strength. This mounts the congestion on the network where all the mobile nodes handover

which results into high number of packet drops in scenarios using no load balancing. The scenarios using fuzzy load balancing with cost offload the satellite network due to mobile nodes preference for the terrestrial network. This redirects most of the mobile users' traffic to terrestrial networks which is the reason fuzzy algorithm with cost shows more drops as compared to the fuzzy load balancing algorithm.

Similarly Figure 6-44 shows the packets drop rate for the 50 mobile user scenarios using 2m/s mobile node speed. In this case no load balancing shows 0.03% packet drop, fuzzy load balancing shows 0.0076% packet drop and fuzzy with cost algorithm shows 0.0079% packets drop. The scenarios with mobile node velocity of 2m/s shows higher number of dropped packets as compared to the scenarios having 25m/s of mobile nodes velocity. This is because the scenarios with 2m/s mobile nodes velocity produce large traffic while covering the same distance as in case of scenarios with 25m/s.

c) Handover Latency

The total handover latencies observed by all the mobile nodes in this scenarios using no load balancing, fuzzy load balancing and fuzzy load balancing with cost are recorded with different mobile node velocities such as 25m/s and 2m/s. The graph shown in Figure 6-45 represents the total handover latencies observed by each mobile node in 50 mobile nodes scenario with mobile nodes velocity of 25m/s. The blue dots in Figure 6-45 represent the total handover latencies observed by each mobile node without load balancing. Pink dots represents the total handover latencies with fuzzy load balancing and cyan dots represent the total handover latencies with fuzzy load balancing using cost preferences from mobile users. The x-axis is

showing the individual mobile node and the y-axis represents the total

handover latency in seconds. Similarly the blue horizontal line represents the

mean of the total handover latencies observed by each mobile node using no

load balancing, pink horizontal line represents the mean value for the total

handover latencies observed by each mobile node using fuzzy load

balancing and cyan horizontal line represents the mean value of total

handover latencies observed by each mobile node using fuzzy load

balancing with cost preferences from mobile nodes.
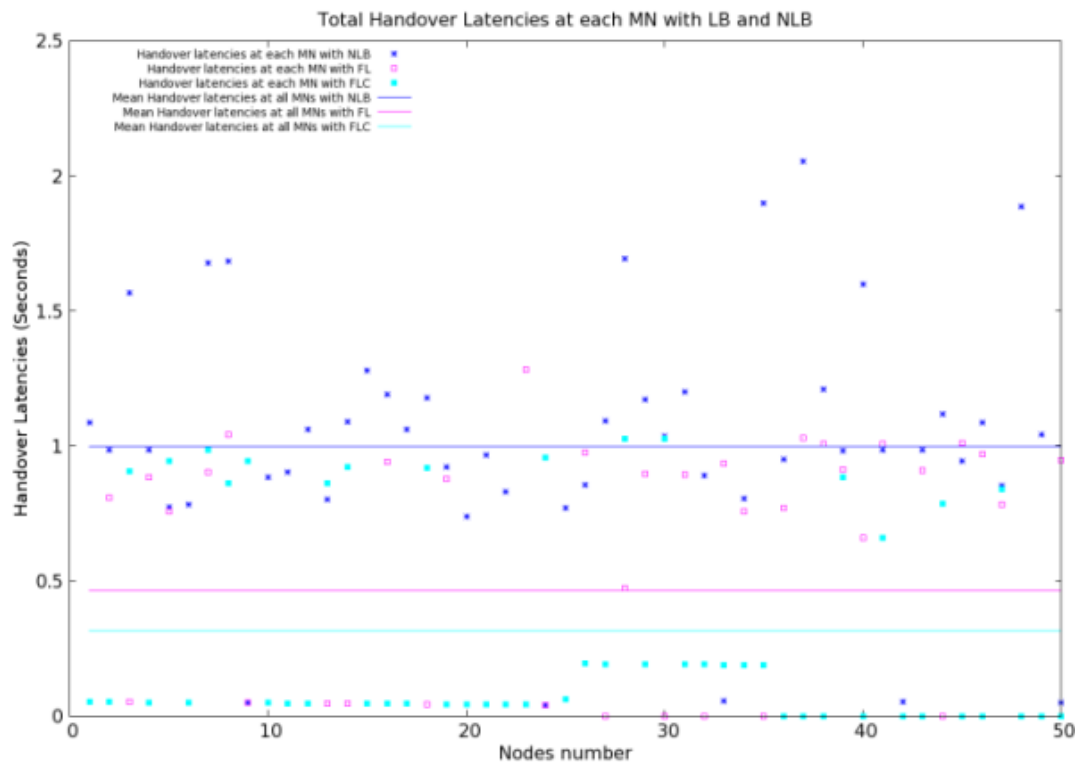
Figure 6-45: Handover Latency (speed - 25m/s)

It can be seen from Figure 6-45 that the average handover latency for no

load balancing and with fuzzy load balancing are at around 1 sec, as in this

case most of the mobile nodes handover to the best available network upon

entering the common coverage areas. Hence a large number of handovers

take place thereby resulting in this large overall delay. The mean values for

the fuzzy load balancing with and without cost preferences are much lower

than this at around 0.48 second and 0.35 second approximately. This is due to the fewer handovers that take place in these cases.
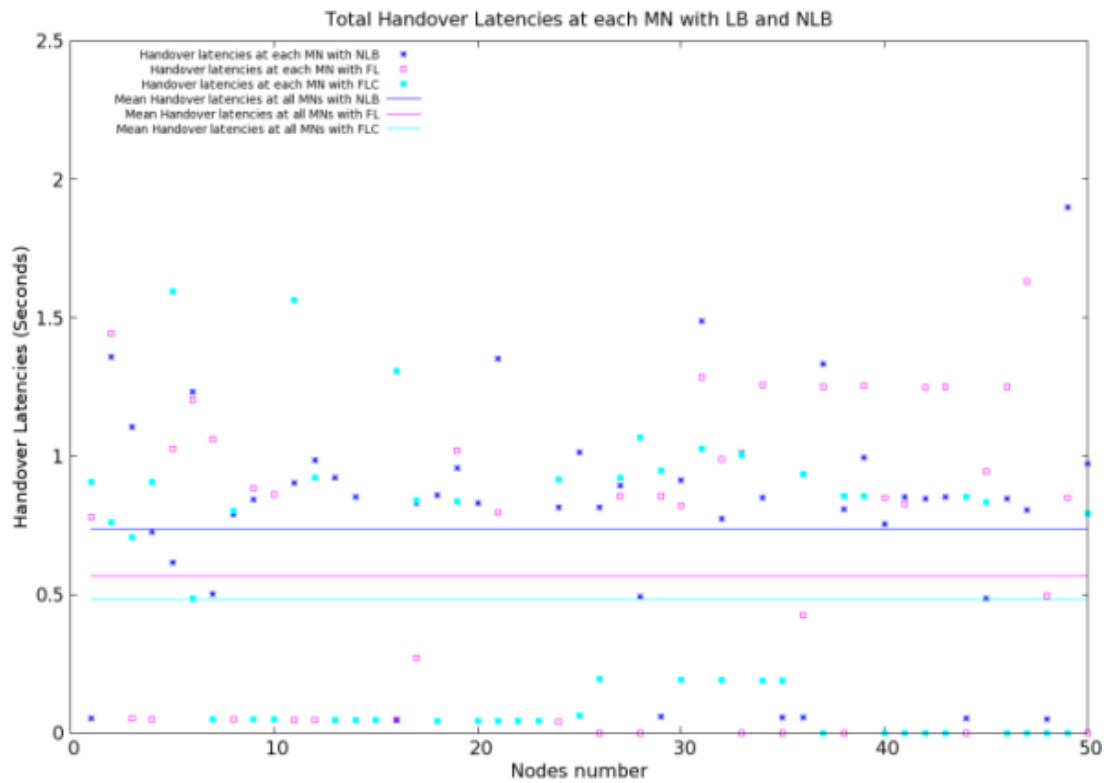

Figure 6-46: Handover Latency (speed - 2m/s)

Similarly Figure 6-46 shows the handover latencies when the mobile nodes are travelling at 2m/s. It can be seen from this graph that without load balancing the handover latencies are highest and the handover latencies for fuzzy and fuzzy with cost are lesser as compared to the no load balancing scenario.

It is noticed that the handover latencies observed by different mobile nodes in 2m/s scenarios are lower as compared to the handover over latencies observed in scenarios with mobile node velocity 25m/s. In case of no load balancing the mean value for the handover latencies is nearly 1 second, however in case of 2m/s the same scenario with no load balancing showed the mean value of handover latencies observed at each mobile node is approximately 0.5 second. In other scenarios such as fuzzy and fuzzy with

cost the differences are smaller but the delays are lesser and hence better than no load balancing.

d) Average throughput at mobile node

Figures 6-47 and 6-48 represent the average throughput observed by each mobile node in this scenario where the speed of mobile node is 25 m/s and 2 m/s respectively. The x-axis of the graphs in Figure 6-47 and Figure 6-48 represents the mobile nodes number and the y-axis represents the throughput in kilobits per second (kbps). The blue dots represents the values without load balancing, the pink dots represents the values with fuzzy load balancing algorithm and the cyan dots represents the fuzzy with cost case. The average throughput for all the mobile nodes is shown as horizontal lines in respective colours.
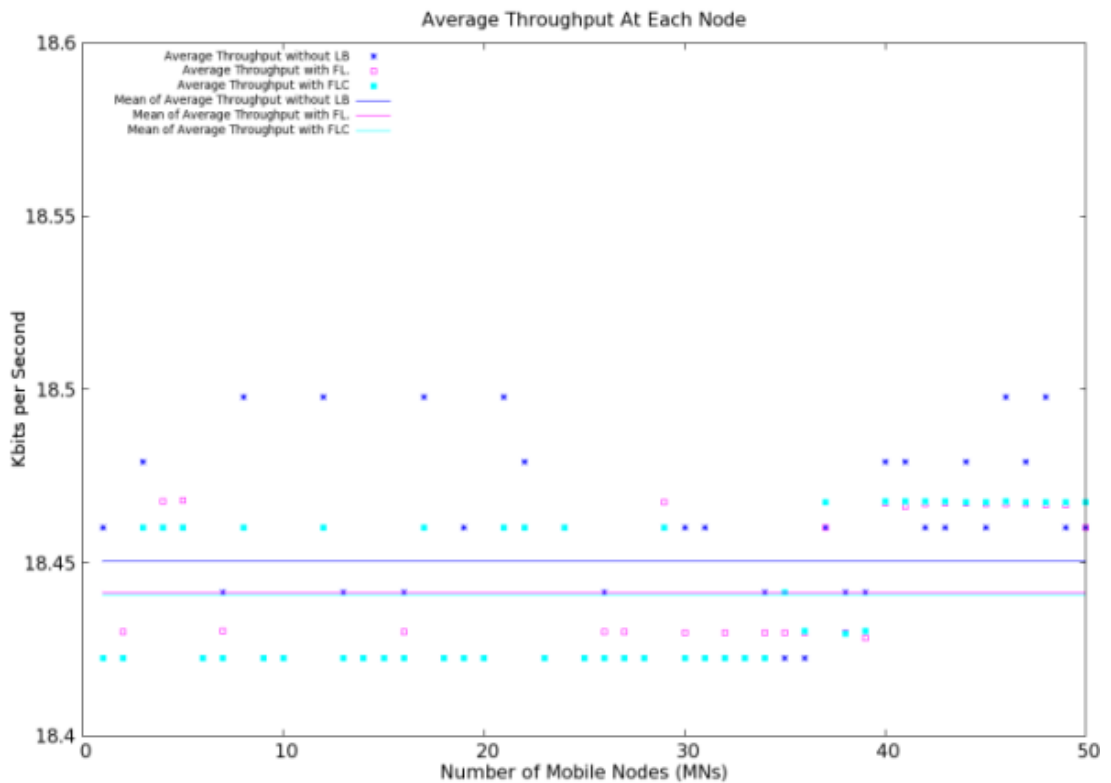

Figure 6-47: Average throughput (speed – 25 m/s)

It can be seen from Figure 6-47 that the average throughput is higher in case of no load balancing at 18.45 kbps as the mobile nodes select the best available network. On other hand the average throughput for all the mobile nodes in case of fuzzy and fuzzy with cost are almost similar at around 18.44 kbps and only slightly lower than that of no load balancing scenario. The reason for this is that load balancing tries to maintain the load equilibrium between the networks and this practice may result selection of network for some mobile node with high network latencies and lower data rate but only after making sure that the network can fulfil the required QoS of the mobile user.
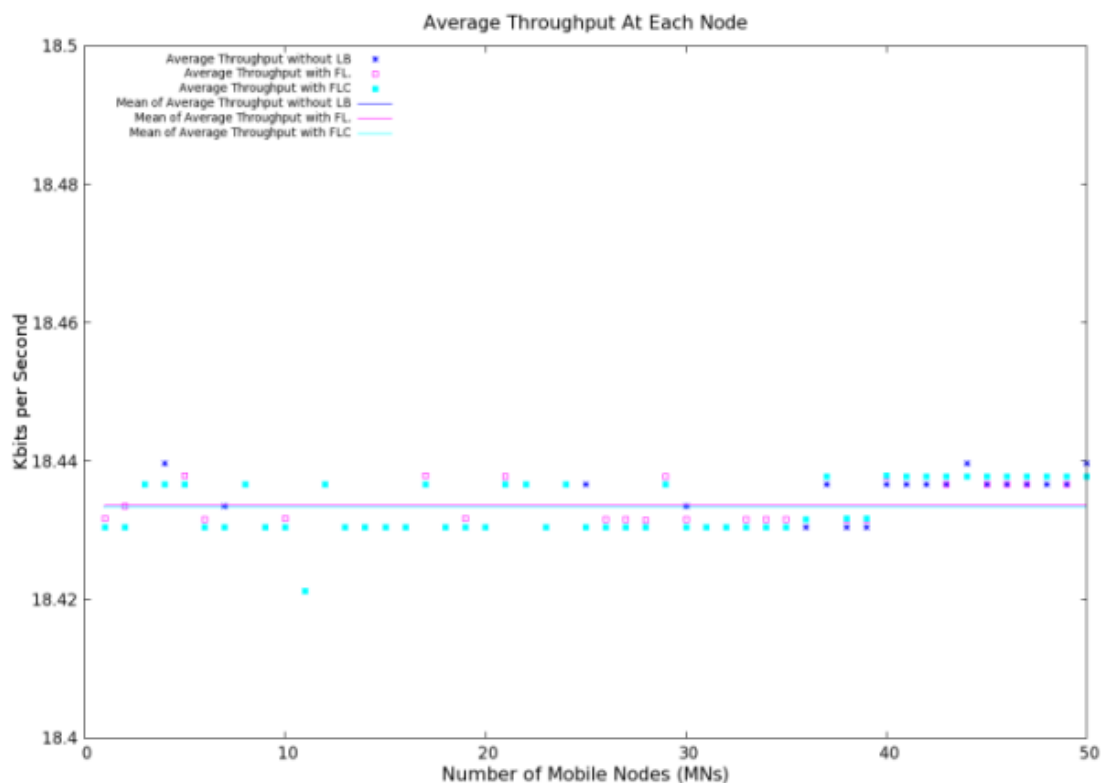

Figure 6-48: Average throughput (speed – 2 m/s)

Figure 6-48 shows that the average throughput is very close at approximately 18.432 Kbps in all the cases when moving at 2m/s. This shows that the use of load balancing does not really affect the throughput of the users. The main

reason for closer average values is this scenario is the high total time of simulation due to the slow moving users. The scenario with low speed takes longer to travel the trajectory and therefore generates a large amount of traffic which causes congestion on the networks. This results in a slightly reduced mean values of average throughput of all the mobile nodes for different algorithms.

e) Network Throughput

Figure 6-49 to Figure 6-54 show the throughput at each network for this scenario. These graphs show how the networks are being utilised at different times. Figure 6-49, 6-50 and 6-51 show the throughput of all the networks when the nodes are moving with the speed of 25m/s without load balancing, with fuzzy loaded load balancing and with fuzzy algorithm using cost preferences from users respectively. The x-axis of these graphs represents the time of simulation and the y-axis represents the throughput in terms of packets per second. The pink colour line represent the satellite throughput, blue colour line represents the UMTS throughput, black colour line represents the WiMax throughput and cyan colour line represents the throughput of WLAN.

The graph in Figure 6-49 shows that all the nodes handover from satellite to the UMTS at time approximately 30 seconds during simulation. At approximately 51 seconds the mobile nodes enter the WiMax coverage area and leave the WiMax coverage area at time 90 seconds. The traffic in the WLAN network starts at approximately 66 seconds and ends at approximately 74 seconds. As without load balancing the mobile nodes handover to the best available network therefore the average throughput on

179

each network shifts to the newly available better network, whenever the mobile nodes enter the network with low network latencies and high data rates. The graphs represent the number of TCP packets sent/received per unit time across the simulation.
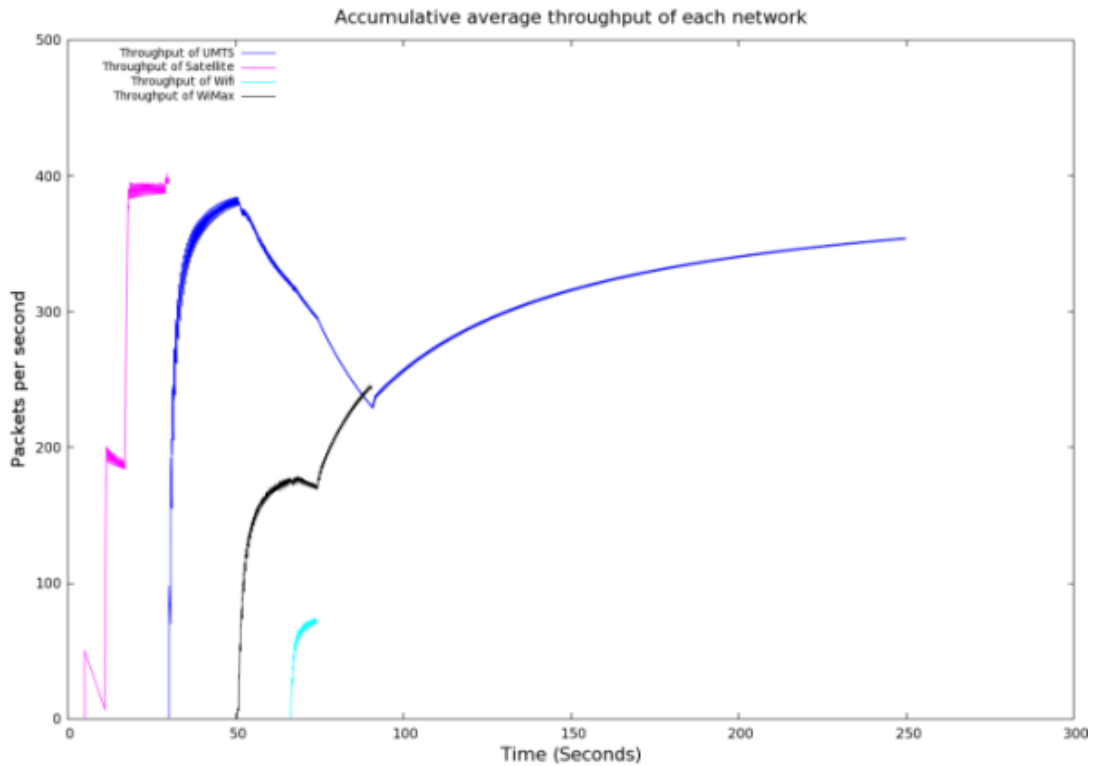

Figure 6-49: Network throughput with no load balancing (speed = 25m/s)

Figure 6-49 clearly shows that when no load balancing is used, in the beginning the throughput of the satellite network is 400 packets/second showing that, the satellite link is heavily utilised as all users are on this link. Similarly at 30 seconds when mobile nodes enter the common coverage area of satellite and UMTS networks, the traffic shifts to UMTS and the UMTS throughput is approximately 380 packets/second. When the users enter WiMax, we can see the network throughput for WiMax increases to around 180 packets/sec, while the network throughput of UMTS network decreases. This is because there are still some users in the UMTS network. When the users enter WLAN coverage area, we can see from Figure 6-49 that all the

users move into WLAN. However the network throughput is only at around 80 packets/second; due to the high congestion in the network and resulting packet drops.
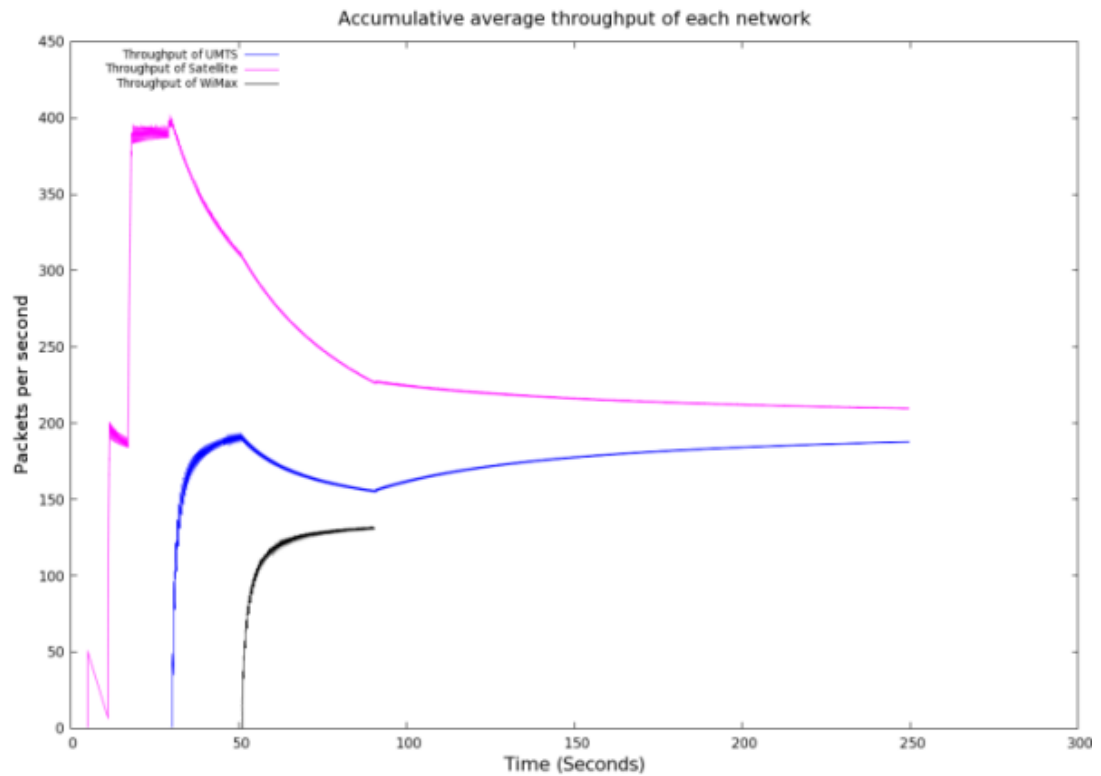

Figure 6-50: Network throughput with Fuzzy based load balancing (speed = 25m/s)

On the other hand Figure 6-50 shows the traffic on all networks when the fuzzy load balancing algorithm is applied. In this case until 30 seconds when all the mobile users are in satellite only coverage area, the satellite average throughput is approximately, 400 packets/second. At time 30 seconds when the mobile nodes enter the common coverage area of satellite and UMTS networks the traffic in satellite network decreases and the traffic in UMTS network increases as the load is now shared between satellite and UMTS networks.
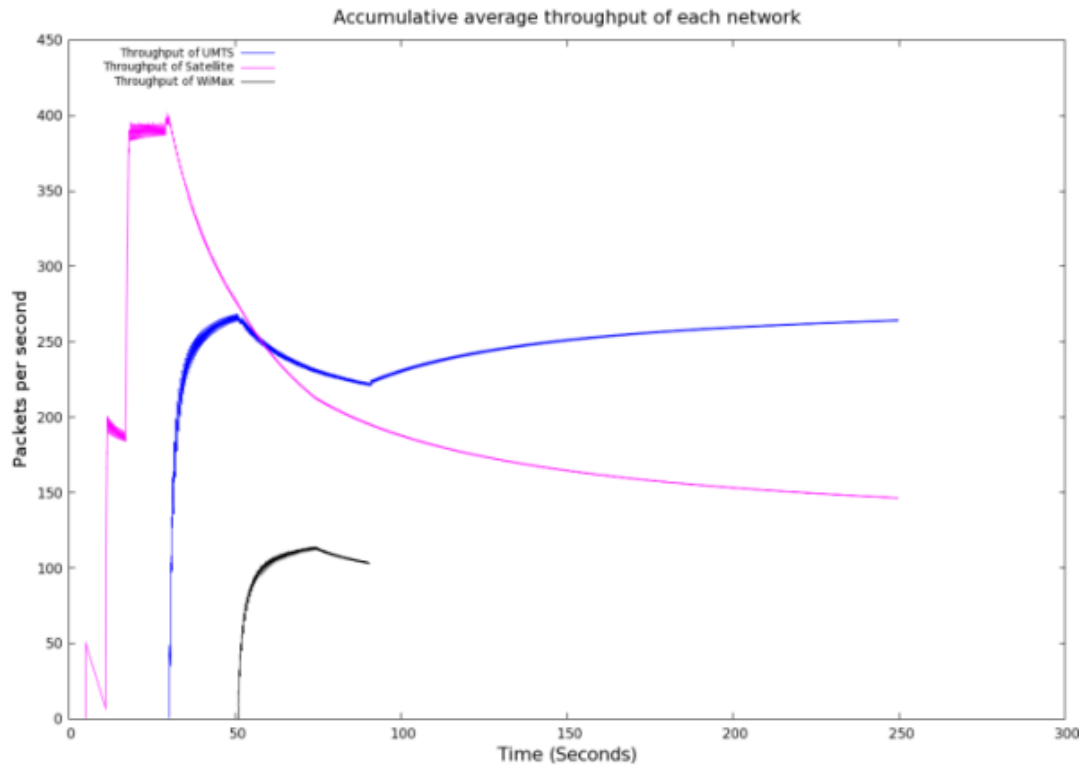
Figure 6-51: Network throughput with Fuzzy based load balancing with cost (speed = 25m/s)

Comparison of Figure 6-49 and Figure 6-50 shows that when load balancing is applied the traffic in other networks is lesser (e.g.: UMTS is around 200 packets/second which earlier was around 380 packets/sec) as the satellite network shares the load with terrestrial networks throughout the simulation time. This shows the benefit of load balancing algorithm for sharing the load between networks to avoid the congestion situation. One major change in this graph is that with fuzzy load balancing the WLAN does not get any user as the speed of mobile nodes is 25m/s which is high enough to pass the WLAN coverage area in a very short time. This is detected by the fuzzy algorithm intelligently and therefore it did not allow any mobile node to handover to the WLAN, while passing through WLAN coverage area.

Figure 6-51 represents the traffic in each network when fuzzy load balancing with cost is applied. Comparing Figure 6-50 with Figure 6-51 shows the

decreases in satellite traffic and increase in the other terrestrial networks traffic (e.g.: UMTS is around 270 packets/second which with no load balancing was around 380 packets/second but with fuzzy was around 200 packets/sec) due to the fact that most mobile nodes do not want to pay for satellite when they use terrestrial networks. This shows that the cost can degrade the efficiency of load balancing algorithms only slightly, but it is still far better than no load balancing.
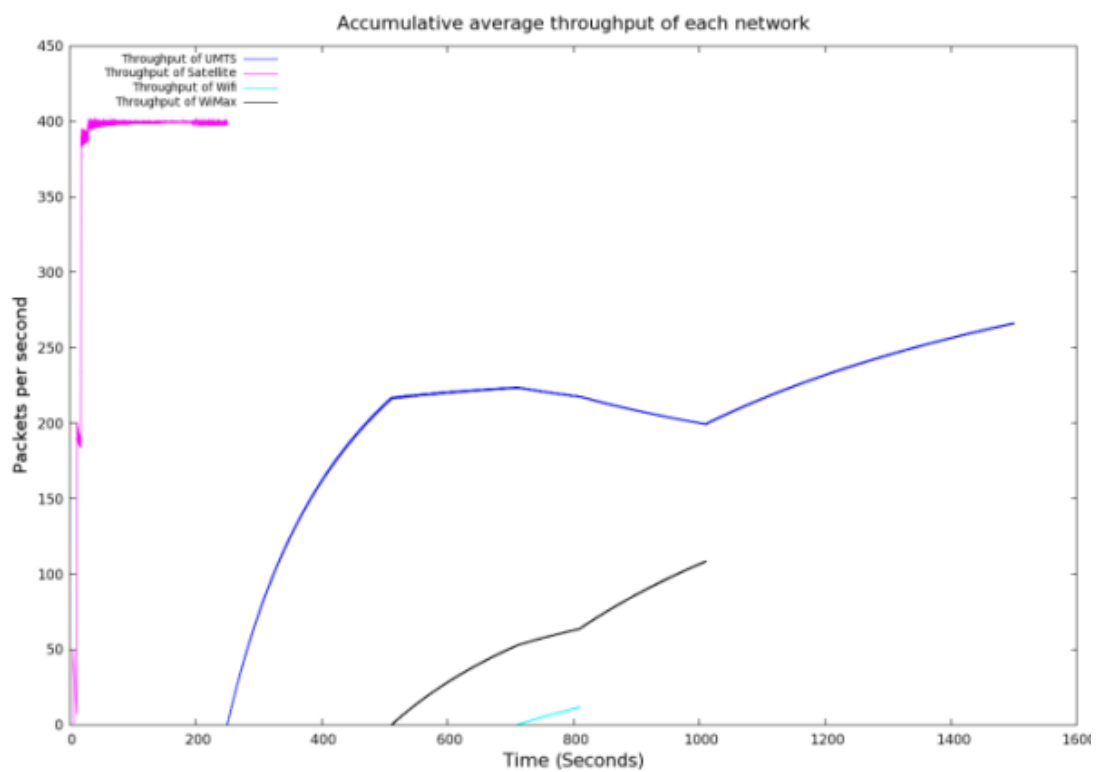
Figure 6-52: Network throughput with no load balancing (speed = 2m/s)

Similarly Figure 6-52, 6-53 and 6-54 represent the throughput in each network for the 2m/s scenario. Without load balancing the satellite network throughput is approximately 400 packets/second until 250 seconds. At 250 seconds the mobile nodes enter the common coverage area of satellite and UMTS and all the mobile nodes handover to the UMTS network making the throughput in UMTS to approximately 240 packets per second. At 511

183

seconds the mobile nodes enter the WiMax coverage area, therefore the traffic shifts from UMTS to WiMax making the WiMax throughput to approximately 70 packets/second and increasing. At 710 second the WLAN network appears in the trajectory of mobile nodes again the traffic shifts from WiMax to WLAN making the throughput at WLAN approximately 20 packets/second. At 810 seconds the mobile nodes leave WLAN coverage area shifting the traffic back to WiMax make the throughput on WiMax approximately 110 packets/second. At 1010 seconds mobile nodes leave the WiMax coverage area, shifting all the traffic to UMTS network making throughput at UMTS nearly 270 packets/second.
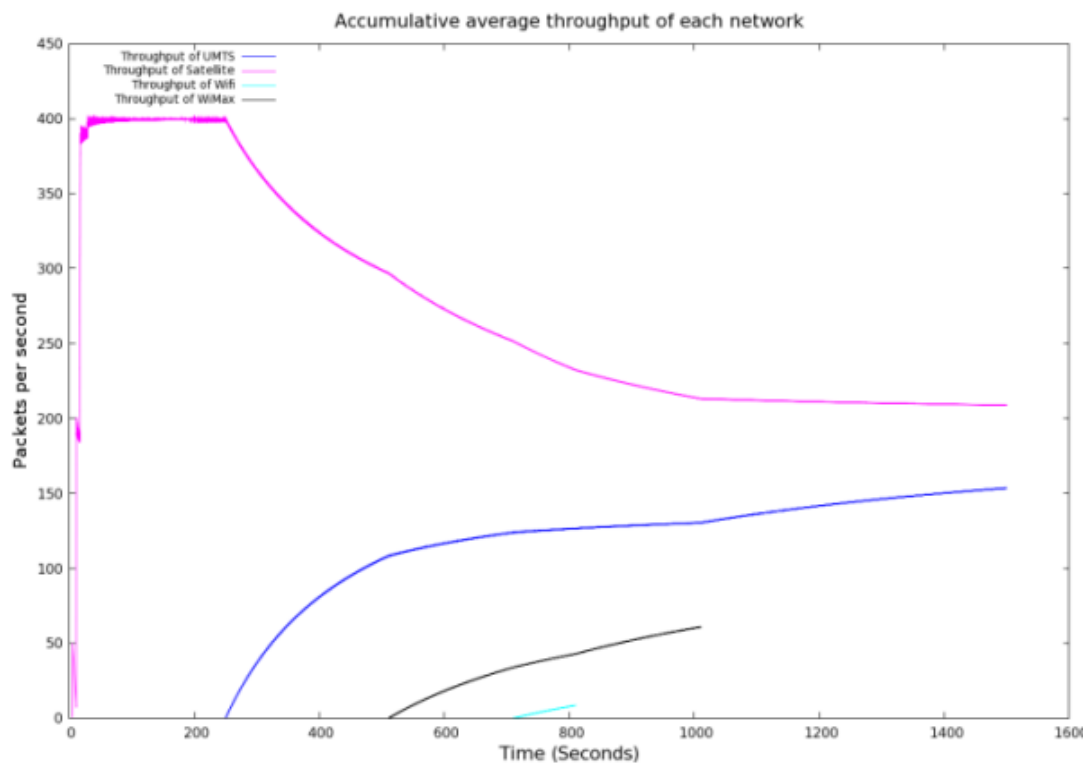

Figure 6-53: Network throughput with Fuzzy based load balancing (speed = 2m/s)

The scenario with mobile nodes velocity 2m/s show the similar behaviour as the fuzzy load balancing algorithm shares the load between co-located networks. However in case of scenario with 2m/s the fuzzy algorithm utilised

the WLAN when they pass through the coverage area of WLAN. In case of fuzzy load balancing the traffic is shared between co-located networks as shown in Figure 6-53. It shows that when load is shared between satellite and terrestrial networks the average throughput at satellite, UMTS and WiMax network reduces to 200 packets/second, 140 packets/second and 70 packets/second. This reduction of traffic in each network by sharing the load between different available networks minimizes the chances of congestion and hence improves the performance.
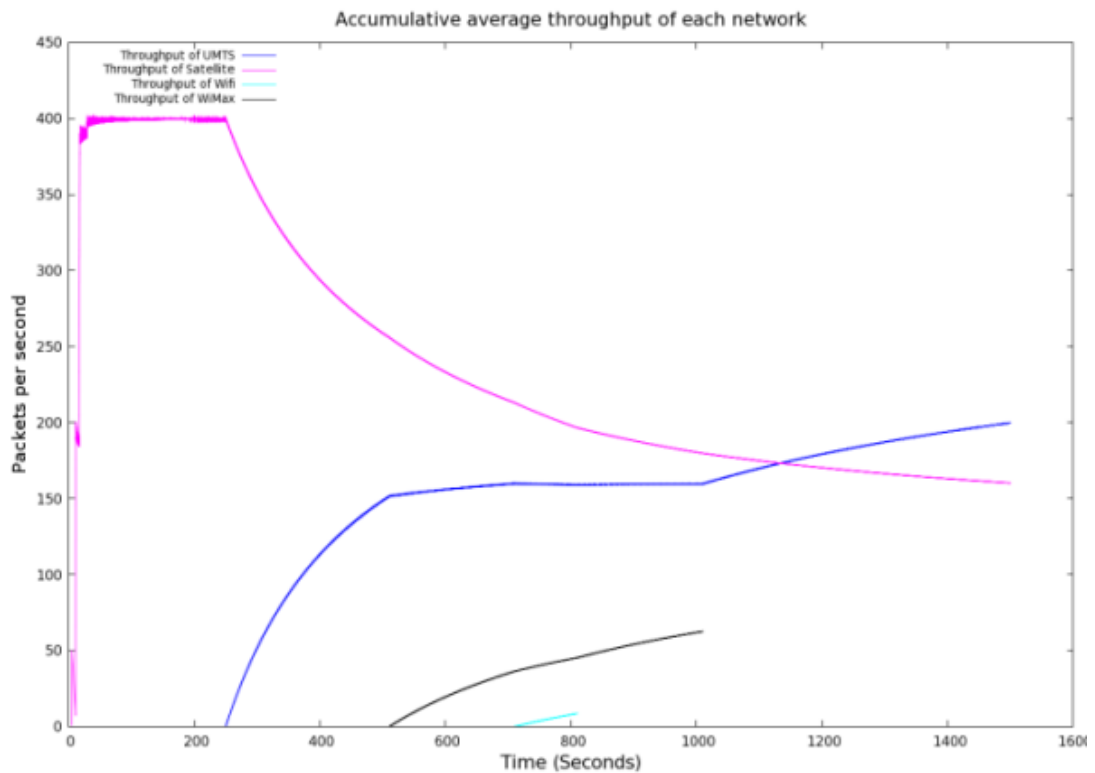

Figure 6-54: Network throughput with Fuzzy based load balancing with (speed = 2m/s)

Figure 6-54 represents the average throughput in each network when fuzzy load balancing is applied with cost. It shows the effects of cost on balancing the load in co-located wireless networks as the average throughput in satellite reduced to 160 packets/second and in UMTS and WiMax increased

210 packets and 60 packets approximately. This concludes that high network cost and users preferences towards using the inexpensive available networks may degrade the load balancing but the overall results with load balancing with cost are still improved as compared to the no load balancing. As the load balancing still share the traffic and avoids or minimizes the chances of congestion on the co-located networks using all available options.

6.2.4 Scenario 4 – Fuzzy based algorithm with 100 users

The 100 mobile nodes scenario is simulated with fuzzy load balancing algorithm using mobile nodes velocities of 25m/s and 2m/s. Then same 100 mobile nodes scenario is repeated using the cost preferences from mobile nodes. The results obtained from the 100 mobile nodes scenario applying fuzzy load balancing algorithms are presented below:

a) Network load

In 100 mobile nodes scenarios first set is simulated using high speed of 25m/s and the second set is simulated using low speed of 2m/s. The graphs shown in Figure 6-55 to 6-60 represent the load in each network such as satellite, UMTS, WiMax and WLAN at the different position of travel trajectory for the 100 mobile nodes scenarios with the mobile nodes moving at 25m/s and 2m/s.

These graphs show the load for the simulation scenarios when no load balancing is applied, fuzzy load balancing is applied and when fuzzy load balancing algorithm is applied with cost preferences. The x-axis of the graphs represents the selected time points where the load in each simulation scenario is monitored. The y-axis of the each load graph represents the total load in terms of number of users in that particular network.
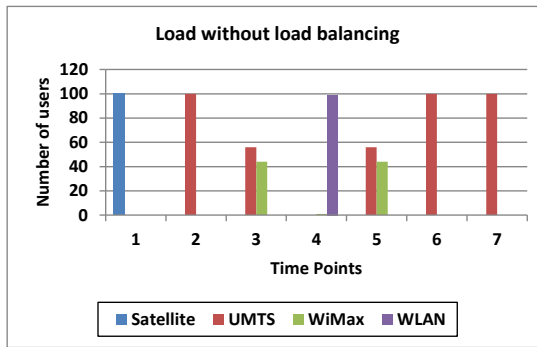
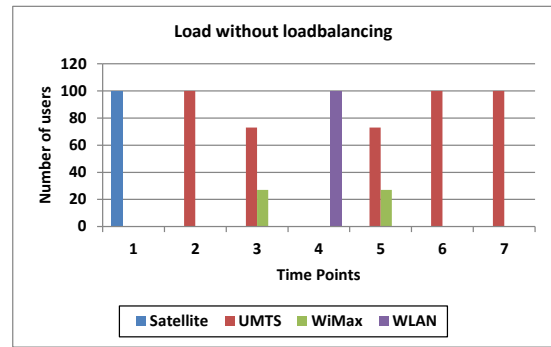Figure 6-55: Load distribution without load balancing in 25m/s

Figure 6-56: Load distribution without load balancing in 2m/s
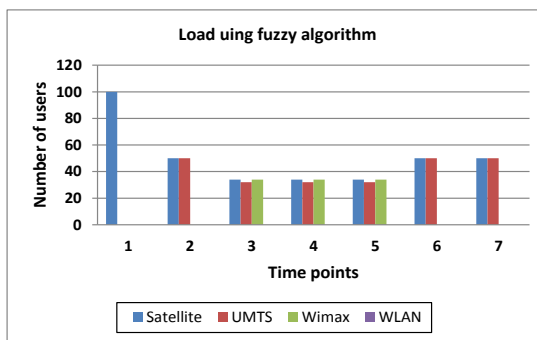
Figure 6-57: Load distribution with fuzzy load balancing in 25m/s
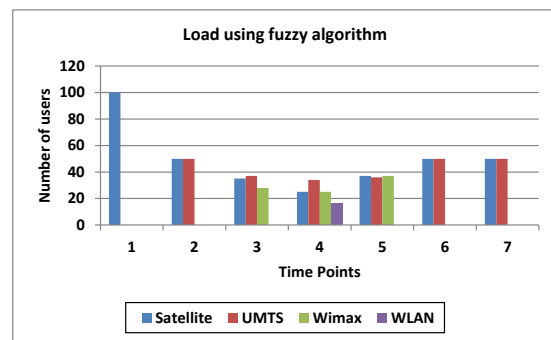
Figure 6-58: Load distribution with fuzzy load balancing in 2m/s
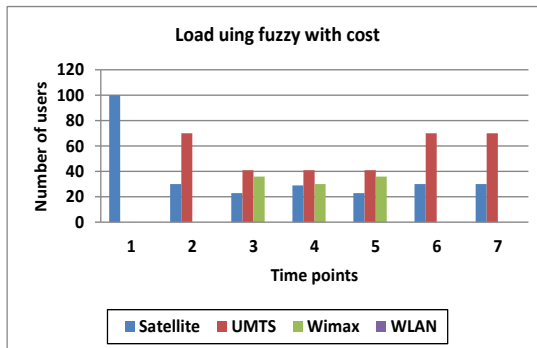
Figure 6-59: Load distribution with fuzzy using cost in 25m/s
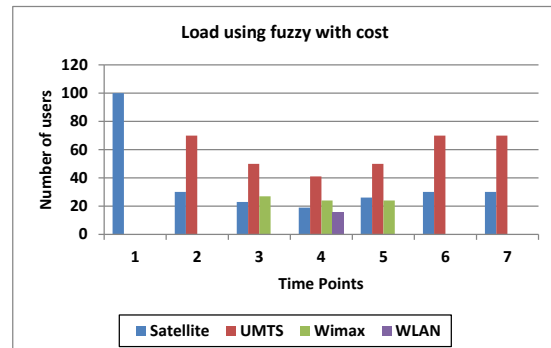
Figure 6-60: Load distribution with fuzzy using cost in 2m/s

It can be seen from these obtained load results that without load balancing most of the mobile nodes handover to the best available network in terms of cost and network latencies. For example in position 4 in Figure 6-55 and 6-56, we can see that all the users connected to WLAN. This however leaves the other networks under-loaded or underutilised. In the scenario where fuzzy load balancing is applied all the networks share the load where possible such

as in the overlapped coverage areas. For example in the same position 4 in Figure 6-57 and 6-59, we can see that the users are distributed across the different networks except the WLAN. The reason behind no user at WLAN is that the mobile nodes are moving with high speed of 25m/s which passing through the WLAN coverage area and the fuzzy algorithms has intelligently decided not to handover mobile nodes to WLAN as the mobile nodes would not spend considerable amount of time in WLAN.

Figure 6-59 and Figure 6-60 represent the load in different networks when fuzzy load balancing is applied with cost preferences. It can be seen that the network cost affects the load balancing in these scenarios. Looking at the points P2, P3 and P4 in Figure 6-57 and Figure 6-58 shows that load at these points is equally distributed among the available networks, however in Figure 6-59 and Figure 6-60 the same points show less load in satellite network and higher loads in other networks. This is due to the higher cost of satellite networks as compared to the other networks.

b) Packet drops

The packet drop rate in 100 mobile nodes scenarios are shown in Figure 6-61 and Figure 6-62 given as follows. These Figures represent that packet drops are higher in case where no load balancing is applied. The scenario with fuzzy load balancing suffers least packet drops and the scenario using fuzzy with cost possess packet drops higher than fuzzy and lesser than no load balancing scenario. The x-axis on graphs given in Figure 6-61 and Figure 6-62 represents the algorithms such as without load balancing, fuzzy load balancing and fuzzy load balancing with cost.

The packet drop rate represents the ratio of the number of packets dropped per total packets transmitted. It can be seen from Figure 6-61 that the packets drop rate when no load balancing is applied is the highest (2.1%). This is expected as most of the mobile nodes handover to the best available networks, which can cause congestion on that network thereby resulting in the large number of packet drops. The abbreviations used in the following graphs are as follows: NLB stands for "No Load Balancing" FL stands for "Fuzzy Logic" and FLC stands for "Fuzzy Logic with Cost".
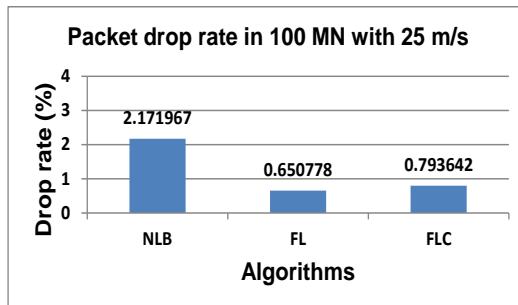
Figure 6-61: Total packet drops in 100 MNs with 25m/s
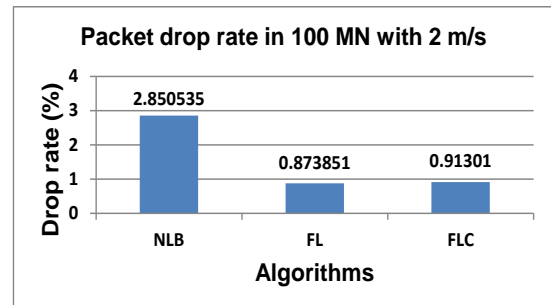
Figure 6-62: Total packet drops in 100 MNs with 2m/s

The packets drop rate is lower (0.65%) when the fuzzy load balancing algorithm is applied. This is because the networks are not overloading in this case. When cost preferences are also considered, the packet drops is (0.79%) which is slightly higher than the fuzzy algorithm with no cost as in this case the load in networks are not perfectly balanced due the varying network service cost and mobile node preferences. It can also be seen from the comparison of graphs shown in Figure 6-61 and Figure 6-62 that with 2mps in 100 mobile nodes scenario the packet drops are increased as the nodes stay longer in the networks and generate large amount of data causing congestion and ultimately resulting into higher drop rate.

c) Handover Latency

The graphs for the handover latencies observed by each mobile node in 100 MNs scenario with mobile nodes velocity of 25m/s and 2m/s are presented in this section. The x-axis in this graph represents the individual mobile nodes and y-axis represents the total handover latency observed by mobile nodes in seconds for their complete journey. This latency is the sum of all the delays for the different handovers any given user would be subjected to during its movement across the travel path. The blue, pink and cyan colour dots represent the total handover latencies of different mobile nodes using no load balancing, fuzzy load balancing and fuzzy load balancing with cost. Similarly the blue, pink and cyan lines show the mean value of the handover over latencies observed at all the mobile nodes using no load balancing, fuzzy load balancing and fuzzy load balancing with cost.
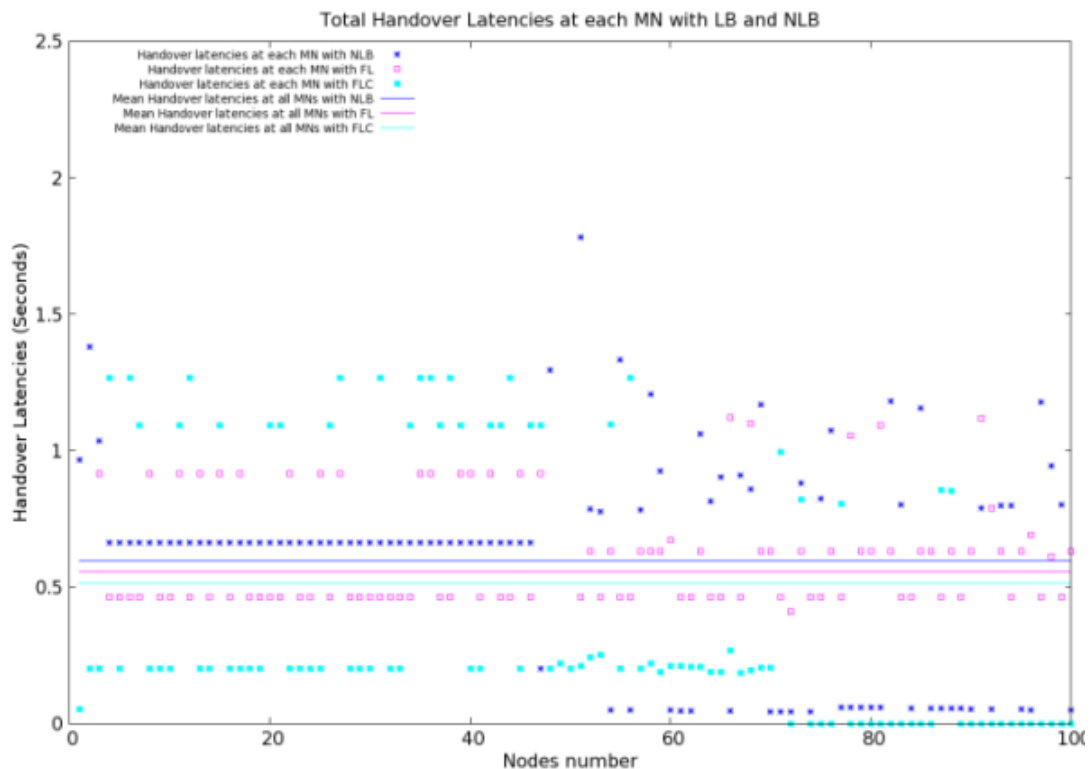
Figure 6-63: Handover Latency (speed - 25m/s)

Figure 6-63 represents the total handover latencies of each mobile node in scenario using mobile nodes velocity of 25m/s. It shown that the average handover latency for no load balancing is highest at around 0.6 second, as in this case most of the mobile nodes handover to the best available network upon entering the common coverage areas. Hence a large number of handovers take place thereby resulting in this large overall delay. The mean values for the fuzzy load balancing without and with cost preferences are much lower than this at around 0.55 second and 0.512 second. This is due to the fewer handovers that take place in these cases.
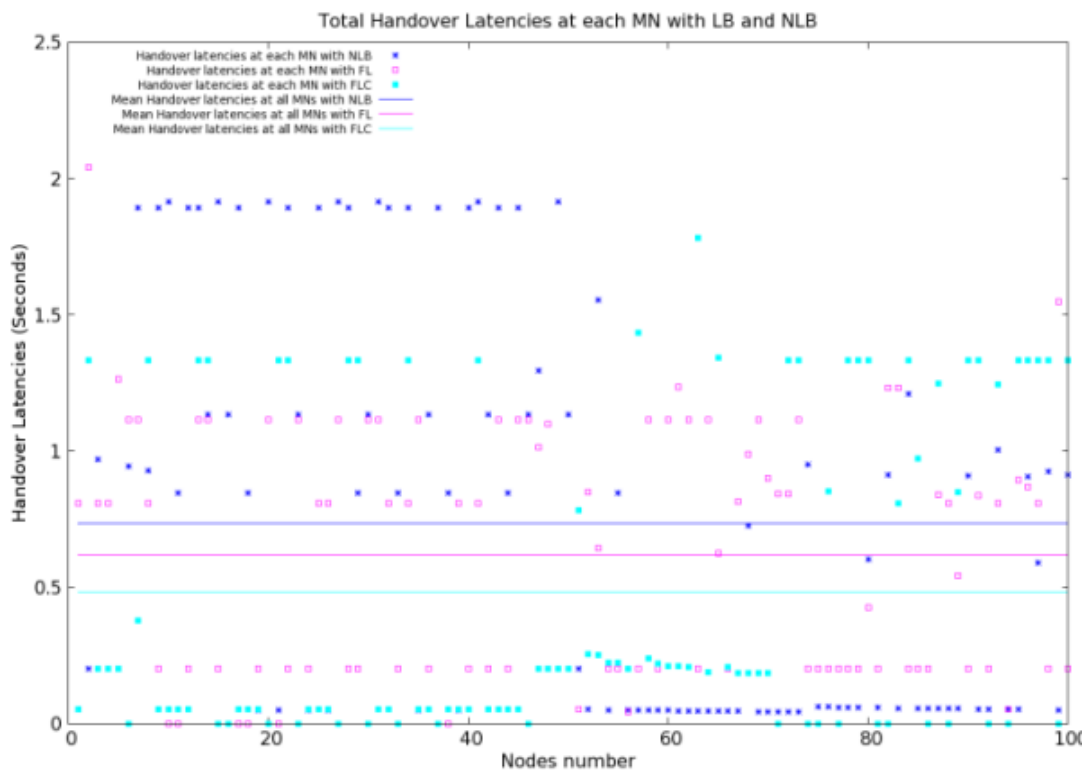
Figure 6-64: Handover Latency (speed - 2m/s)

Similarly Figure 6-64 shows the handover latencies when the mobile nodes are travelling at 2m/s. It can be seen from this graph that without load balancing the handover latencies are highest and the handover latencies for

fuzzy and fuzzy with cost are lesser as compared to the no load balancing scenario.

On comparing the graphs in Figure 6-63 and 6-64, we can also see that for the lower speed scenario the handover delays are lower. This is because as explained in the previous sub-section the packet drops are higher in the high speed scenario which also affects the handover procedure thereby requiring retransmissions of lost control messages during the handover process. The fuzzy with cost has least mean value for the total handover latencies observed by each mobile node. The reason for this is that in case of no load balancing all the mobile nodes perform handover whenever they detect better network or when they leave the coverage are of that network. However in case of fuzzy load balancing only a selected set of mobile users perform the handover in order to maintain the load equilibrium between the networks having common coverage areas or overlapping coverage areas. In case of fuzzy with cost the number of handover are further reduced as most of the mobile nodes do not want to go back to satellite network when the terrestrial network is available.

d) Average throughput at mobile node

The average throughput of each mobile node and their mean values for the 100 mobile nodes scenarios with mobile nodes velocity of 25m/s and 2m/s are shown in Figure 6-65 and Figure 6-66. The blue colour dot represents the average throughput of each mobile node using no load balancing. The pink colour dot represents the average throughput of each mobile node using fuzzy load balancing algorithm and the cyan dot represents the average throughput of each mobile node using fuzzy load balancing with cost. The

blue, pink and cyan colour horizontal lines represent the mean values of the average throughput observed by each mobile node using no load balancing, fuzzy load balancing and fuzzy load balancing with cost.
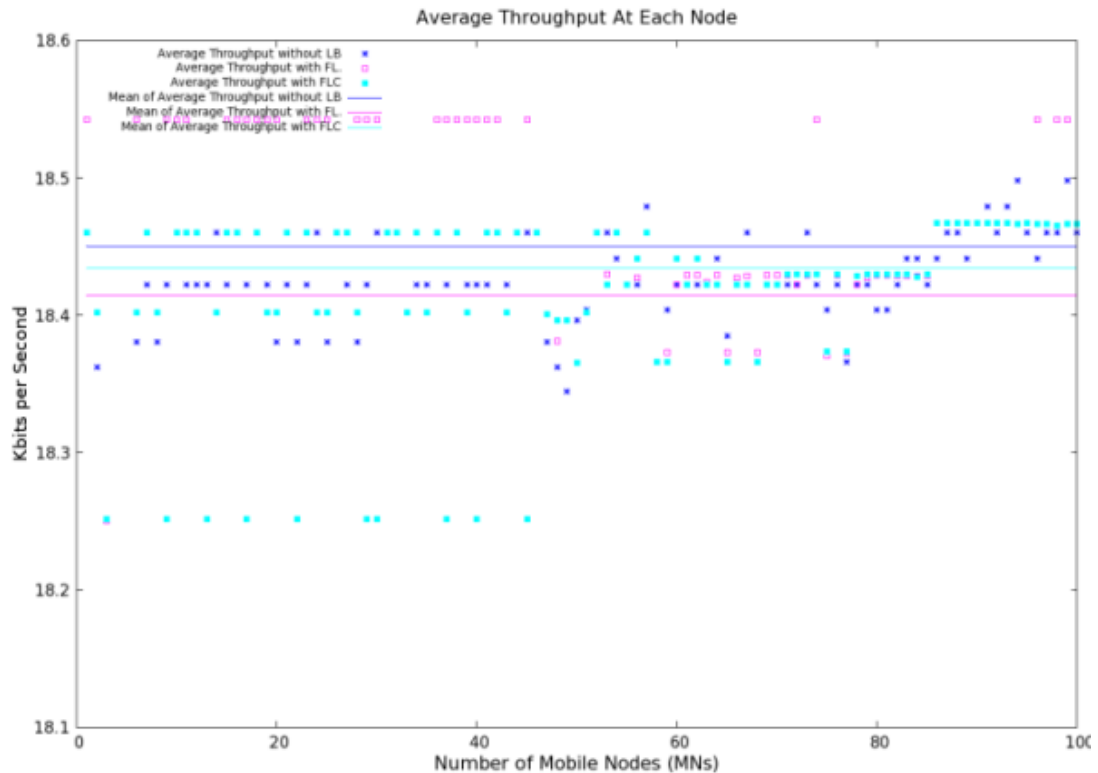

Figure 6-65: Average throughput (speed – 25 m/s)

It can be seen from Figure 6-65 that the average throughput is higher in case of no load balancing at 18.45 kbps as the mobile nodes select the best available network. On other hand the average throughput for all the mobile nodes in case of fuzzy and fuzzy with cost are almost similar at around 18.41 kbps and 18.43 kbps which is only slightly lower than that of no load balancing scenario. The reason for this is that load balancing tries to maintain the load equilibrium between the networks and this practice may result selection of network for some mobile node with high network latencies and lower data rate but only after making sure that the network can fulfil the required QoS of the mobile user.
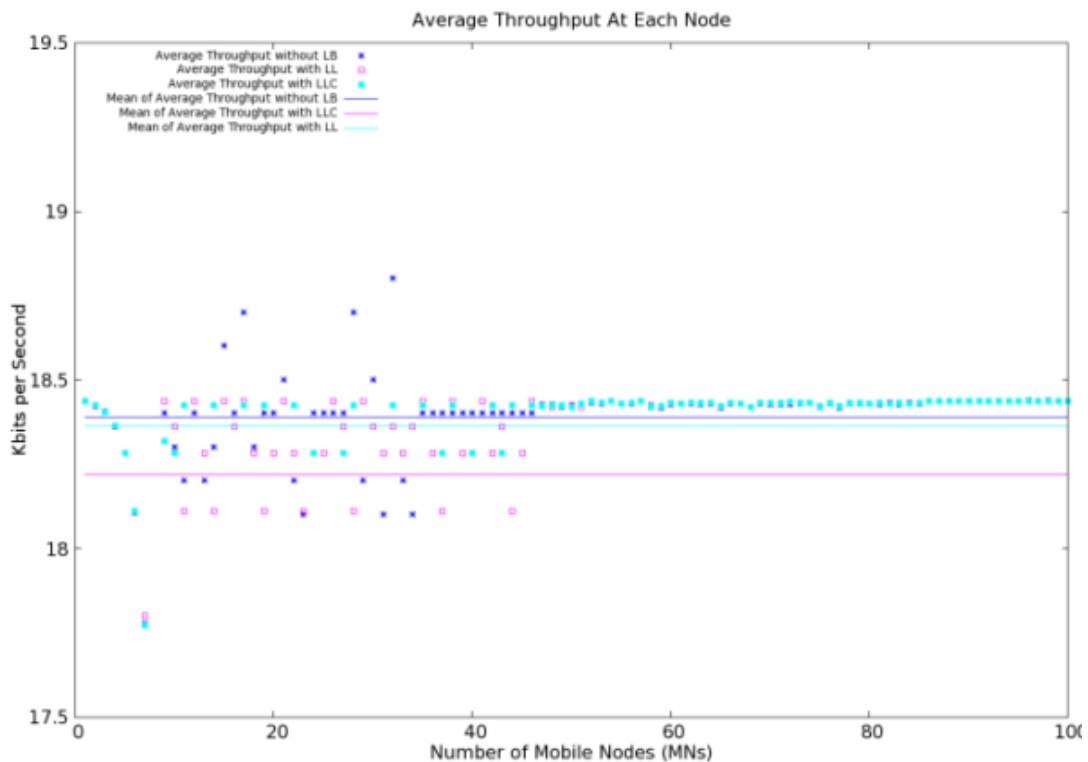
Figure 6-66: Average throughput (speed – 2 m/s)

Figure 6-66 shows that the average throughput is similar at (18.38 & 18.35 Kbps) in cases of no load balancing and fuzzy load balancing with cost and with fuzzy load balancing at (18.24 Kbps) when moving at 2m/s. This shows that the use of load balancing does not really affect the throughput of the users. The main reason for closer average values is this scenario is the high total time of simulation due to the slow moving users. The scenario with low speed takes longer to travel the trajectory and therefore generates a large amount of traffic which causes congestion on the networks. This results in a slightly reduced mean values of average throughput of all the mobile nodes for different algorithms.

It can also be seen on comparing the two graphs that the use of load balancing does not really affect the throughput of the users who can still access their services properly while at the same time the loads across the networks are more uniformly balanced.

e) Network Throughput

The throughput of all the networks such as satellite, UMTS, WiMax and WLAN is shown in Figure 6-67, Figure 6-68, and Figure 6-69 for the 100 mobile nodes scenarios with mobile modes velocity of 25m/s. The x-axis on these figures represents the simulation time and the y-axis represents the throughput such as packets per second. The Pink line represents the average throughput for the satellite network, blue line represents the average throughput for UMTS network, black line represents the average throughput for WiMax network and cyan line represents the average throughput for WLAN.
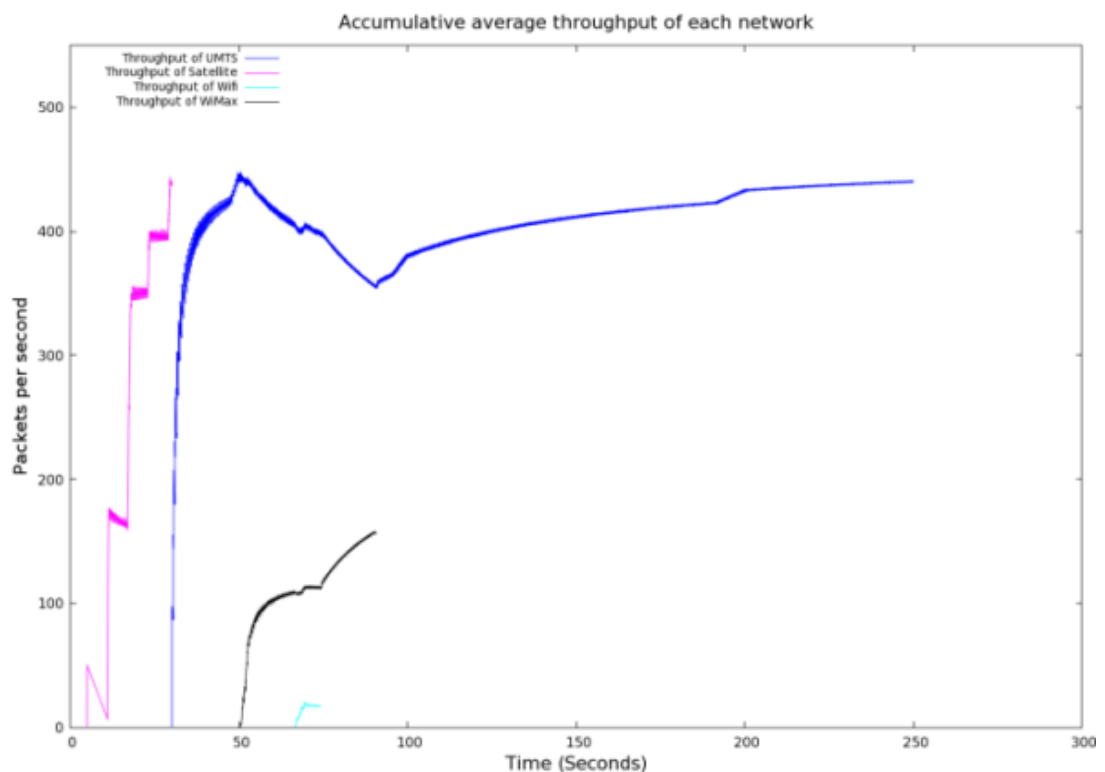

Figure 6-67: Network throughput with no load balancing (speed = 25m/s)

Figure 6-67 shows the average throughput of different networks for 100 mobile nodes scenarios using no load balancing with mobile nodes velocity of 25m/s. It shows that with no load balancing traffic is shifted to the best

available networks when the mobile nodes enter or leave coverage areas or detect a new network.

Figure 6-68: Network throughput with Fuzzy based load balancing (speed = 25m/s)

Figure 6-68 represents the average throughput of different networks when fuzzy load balancing is applied. In this case the traffic is partially shifted to the newly detected networks in order to maintain the load equilibrium between different networks having common coverage area or overlapped coverage area. In this way all the networks are being utilized on availability.

Figure 6-69 shows the average throughput of different networks when fuzzy load balancing with cost is applied. In this case the satellite network serves less number of mobile nodes when the mobile nodes are in the common coverage area of satellite and other terrestrial networks due to the cost. Therefore the load in UMTS network goes comparatively higher (at approximately 210 packets/second) when mobile nodes are in common coverage area of UMTS and satellite networks. This shows a considerable

growth in average throughput of UMTS and for satellite it shows lower average throughput as compare to the fuzzy load balancing scenario shown in Figure 6-68.
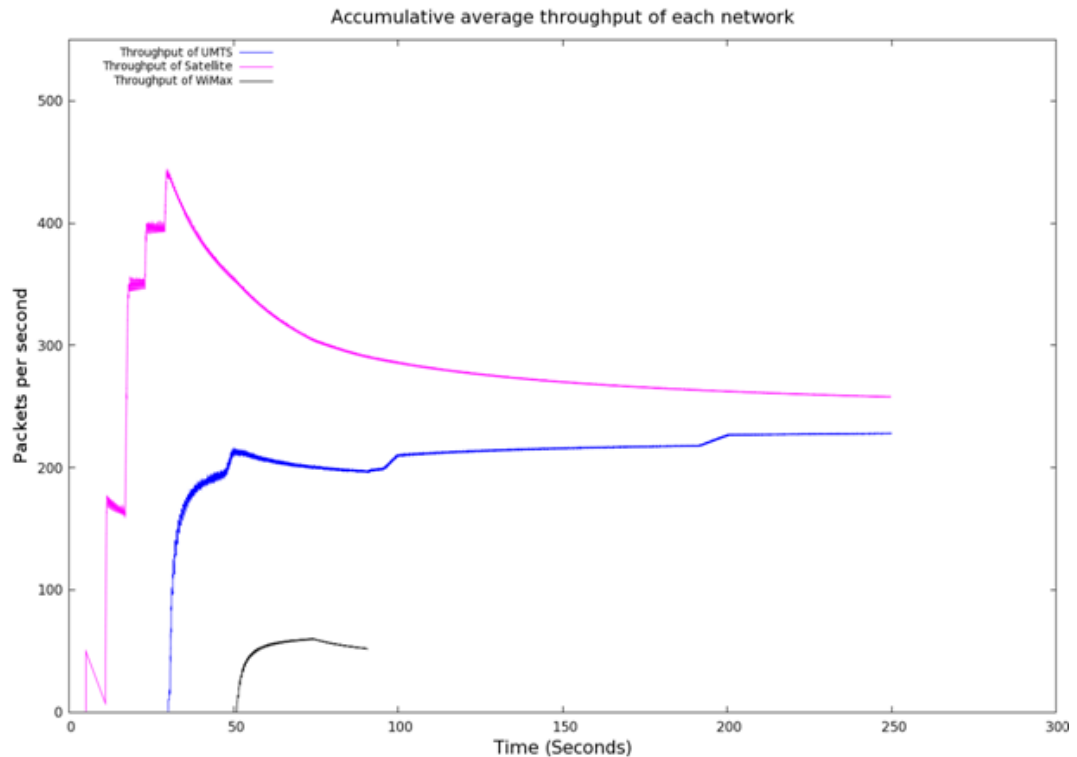
Figure 6-69: Network throughput with Fuzzy based load balancing with cost (speed = 25m/s)

For 100 mobile nodes scenario with mobile nodes velocity of 2m/s, the following Figures from Figure 6-70 to Figure 6-72 are showing the average throughput at all the networks. Figure 6-70 shows the average throughput at all the networks using mobile node at velocity of 2m/s without load balancing algorithm. It can be seen that in 2m/s scenario the average throughput for satellite and UMTS is higher. As the mobile nodes enter the common coverage areas of different networks they handover to the best available networks.

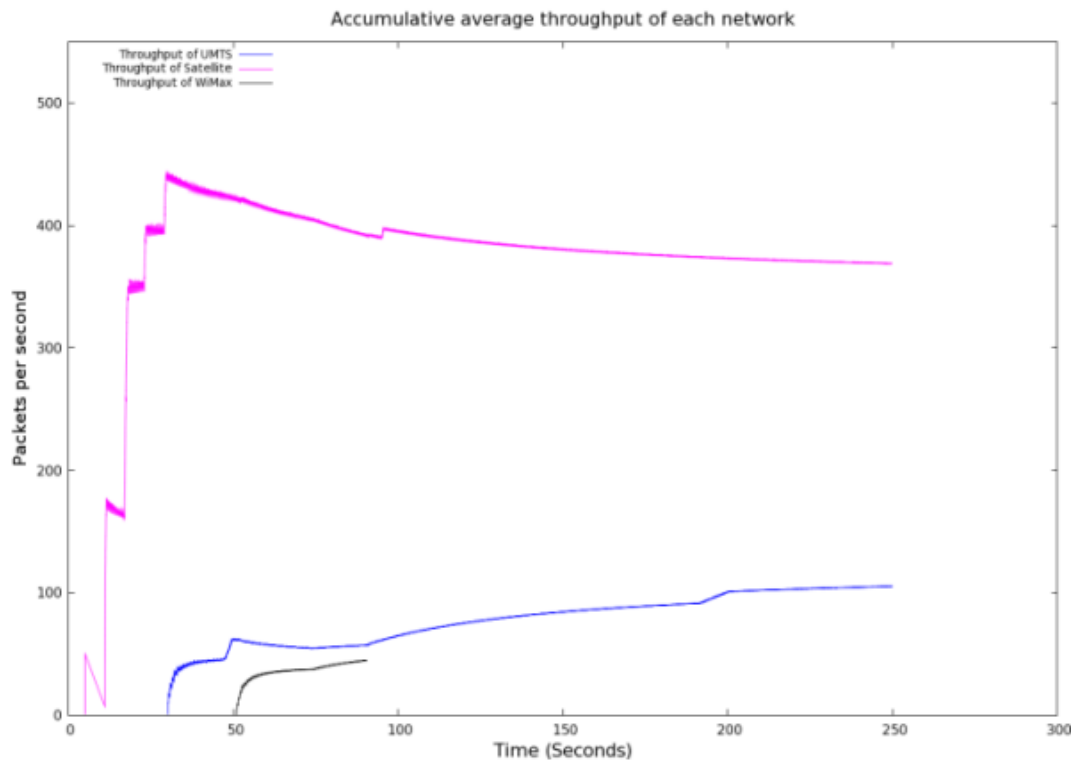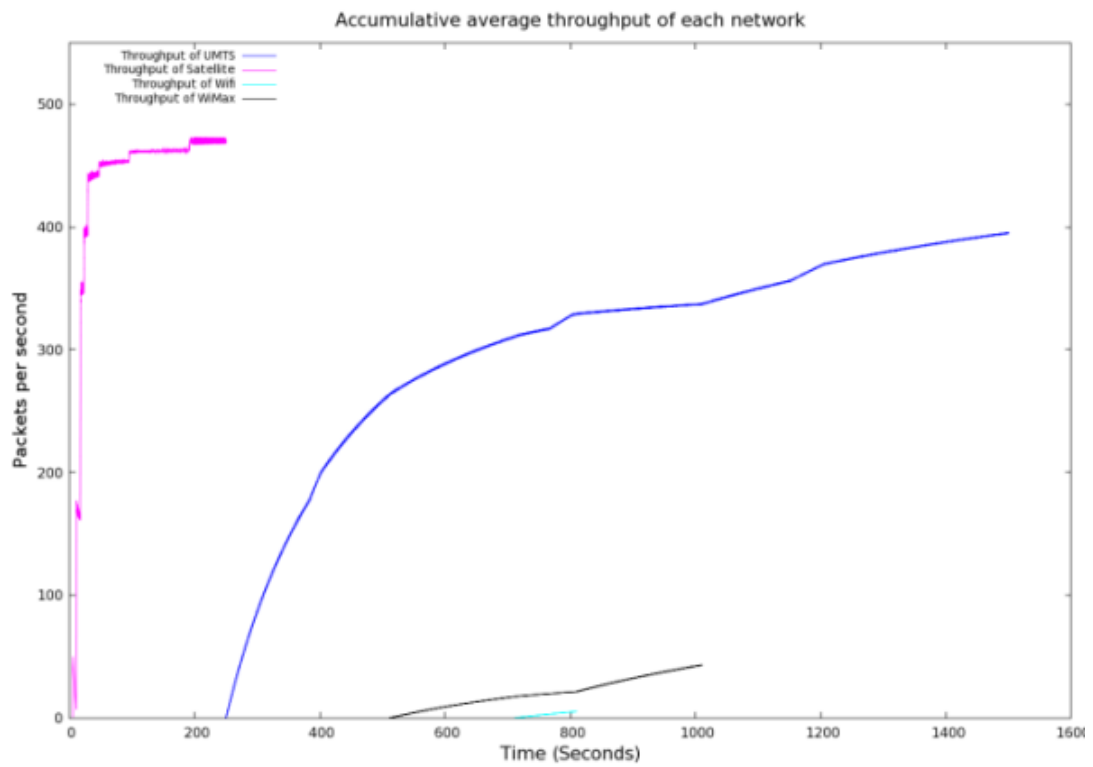Figure 6-70: Network throughput with no load balancing (speed = 2m/s)
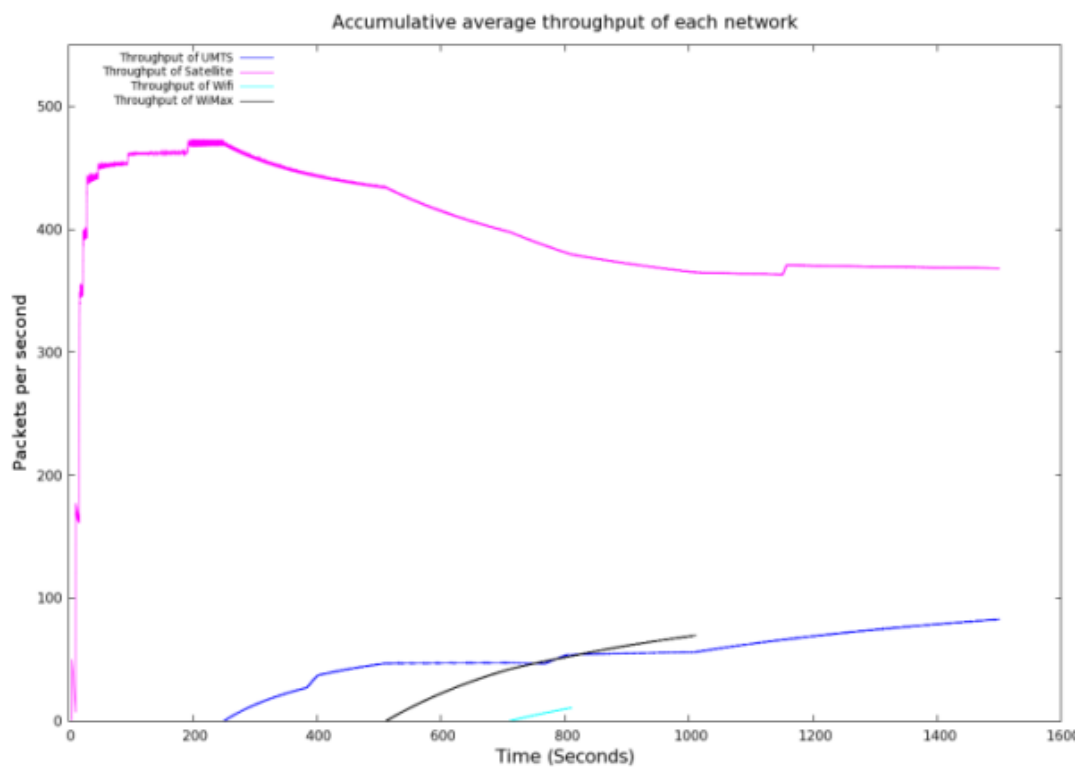
Figure 6-71: Network throughput with Fuzzy based load balancing (speed = 2m/s)

Figure 6-71 shows the average throughput of all the networks when fuzzy

load balancing is applied. This shows that average throughput of the satellite

network is higher (at around 470 packets/seconds) as all the mobile nodes are in the satellite only coverage area at the beginning. When the mobile nodes enter the common coverage area of satellite, UMTS, WiMax and WLAN; they share the load uniformly. Therefore the all the networks are being utilized throughout the simulation.
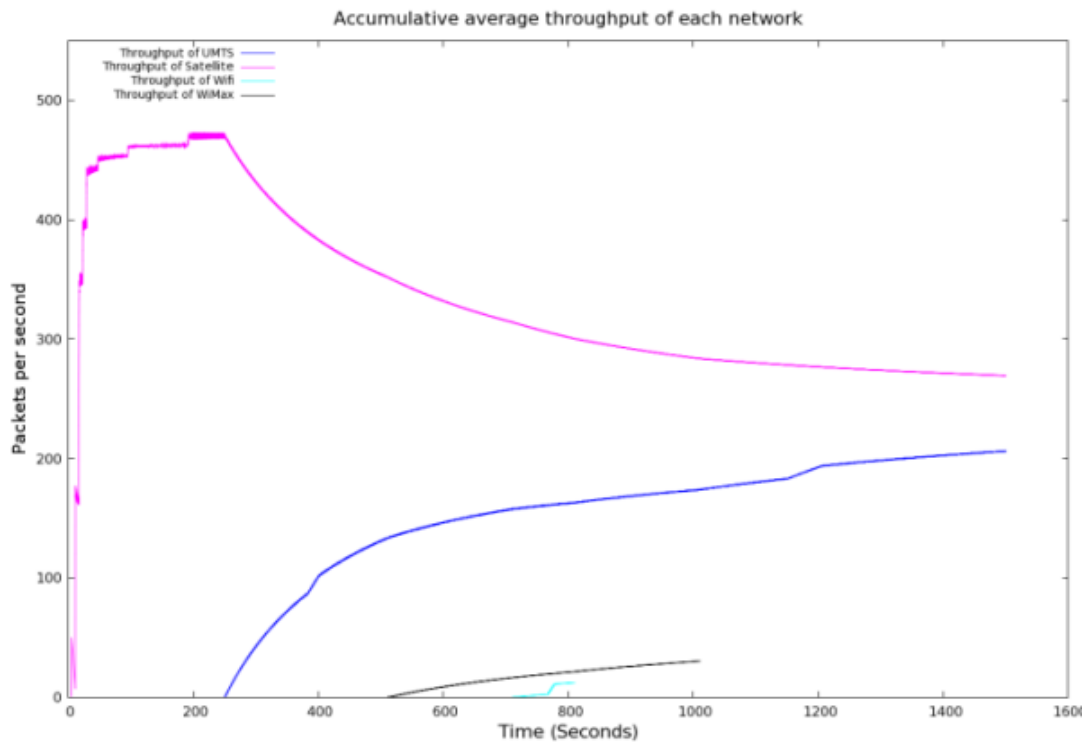
Figure 6-72: Network throughput with Fuzzy based load balancing with cost (speed = 2m/s)

Figure 6-72 shows the average network throughput for different networks using fuzzy load balancing with cost. This scenario assumes that 70 mobile nodes do not want to use satellite network if they can use the other terrestrial networks. For the first 250 seconds in this scenario all the mobile nodes stay in the satellite network as they cannot use any other network in that area this makes the throughput in satellite as 470 packets/second. Once they see the UMTS coverage area 70 mobile nodes handover to the UMTS making

throughput at UMTS 210 packets/second and satellite at 270 packets/second.

Further ahead in the simulation when the mobile nodes move towards the common coverage area with other networks, it is allowed to move mobile nodes to other networks from satellite networks to balance the load but no mobile node is handover to the satellite if the mobile node prefer terrestrial network. This is the reason that average throughput of satellite network decreases in this scenario and that of UMTS increases as UMTS servers all the other mobile nodes which do not prefer the satellite in UMTS satellite coverage area.

Comparing Figure 6-71 and Figure 6-72 shows the dramatic change in the throughput of satellite network and UMTS network and reason is cost of satellite network and most mobile nodes preference for the terrestrial network. Therefore when mobile nodes move to the satellite terrestrial common coverage area, most of them prefer UMTS over satellite network.

6.2.5 Scenario 5 – Neural-Fuzzy based algorithm with 50 users

After applying neural-fuzzy load balancing algorithm on 50 mobile nodes scenario with velocity of 25m/s and 2m/s, the following results were obtained.

a) Network load

The network load for the 50 mobile nodes scenario is represented in the following set of graphs shown from Figure 6-73 to Figure 6-78 when no load balancing is applied, when neural-fuzzy load balancing is applied and when neural-fuzzy load balancing with cost preferences from the mobile nodes is

applied. The x-axis of these graphs represents the selected time points in the simulation for both, 25m/s and 2m/s scenarios. These time points are already explained in the methodology section. The y-axis represents the load in each network in terms of users. The blue, brown, green and purple colour bars represent the load in satellite network, UMTS network, WiMax network and WLAN.

Similar to the fuzzy load balancing algorithm, the neural-fuzzy load balancing algorithm intelligently detects the velocity of the mobile nodes and does not allow the mobile nodes moving with high speed (25m/s) to handover to the WLAN. As the mobile nodes moving with high speed would not stay in the WLAN's coverage area for considerable amount of time.

It can be seen from these obtained load results that without load balancing most of the mobile nodes handover to the best available network in terms of cost and network latencies. For example in position 4 in Figure 6-73 and 6-74, we can see that all the users connected to WLAN. This however leaves the other networks under-loaded or underutilised. In the scenario where neural-fuzzy load balancing is applied all the networks share the load where possible such as in the overlapped coverage areas. For example in the same position 4 in Figure 6-75 to 6-78, we can see that the users are distributed across the different networks.
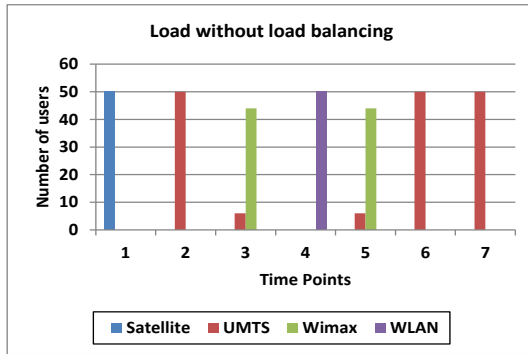
Figure 6-73: Load distribution without load balancing in 25m/s
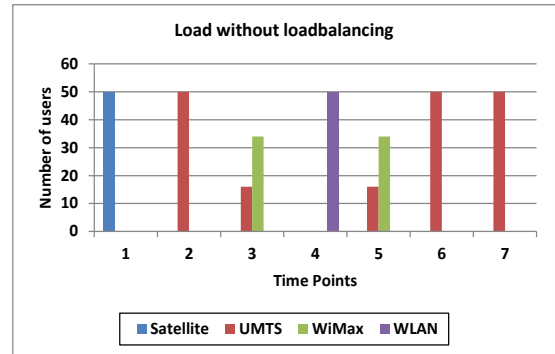
Figure 6-74: Load distribution without load balancing in 2m/s
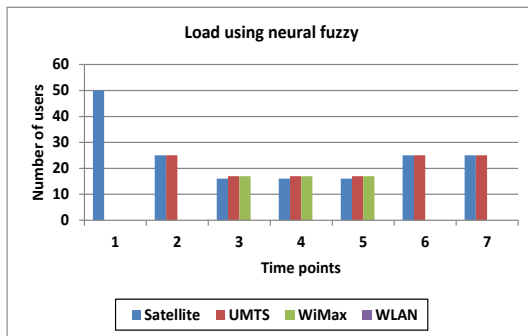
Figure 6-75: Load distribution with neural-fuzzy load balancing in 25m/s
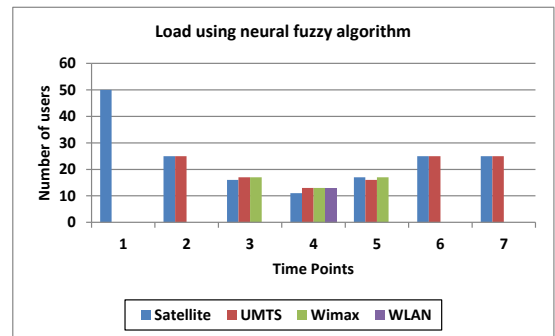
Figure 6-76: Load distribution with neural-fuzzy load balancing in 2m/s
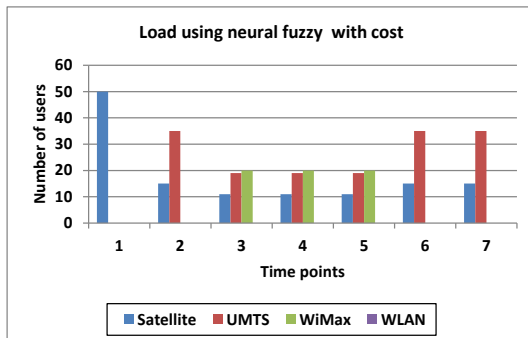
Figure 6-77: Load distribution with neural-fuzzy using cost in 25m/s
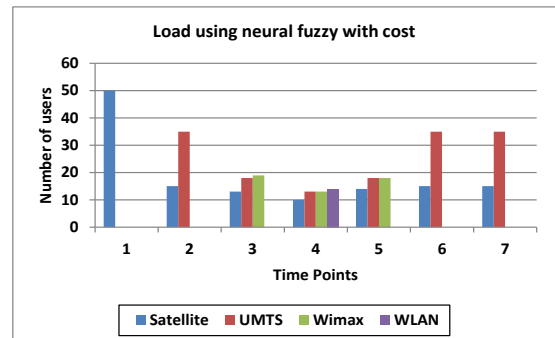
Figure 6-78: Load distribution with neural-fuzzy using cost in 2m/s

Figure 6-77 and Figure 6-78 represent the load distribution in different networks using neural-fuzzy load balancing with cost preferences from different mobile nodes. In these two figures there are two changes; one that with 25m/s the fuzzy algorithm does not allow the mobile nodes to handover to the WLAN and the other is that the satellite network has lower load. The reason why satellite network is having lower load is that most mobile nodes in the terrestrial coverage area do not prefer satellite network due to high

cost of satellite networks. Therefore the load in the UMTS network goes higher in satellite UMTS common coverage area.

b) Packet drops

The packet drop rate in 50 mobile nodes scenarios with mobile nodes velocities 25m/s and 2m/s using no load balancing, neural-fuzzy load balancing and neural-fuzzy load balancing with cost preference from mobile nodes is presented in the graphs below.

Figure 6-79 represents the packet drop rate in scenario with mobile nodes velocity 25m/s and Figure 6-80 represents the packet drop rate in the scenario with mobile nodes velocity 2m/s. The abbreviations used in the following graphs are as follows: NLB stands for "No Load Balancing" NFL stands for "Neural-Fuzzy Logic" and NFLC stands for "Neural-Fuzzy Logic with Cost".
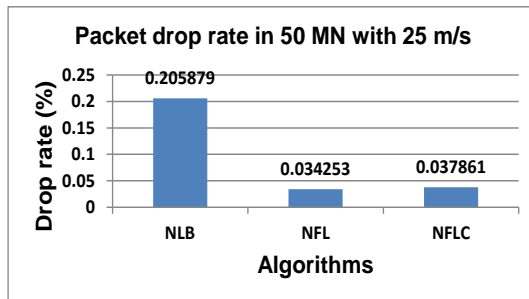
Figure 6-79: Total packet drops in 50 MNs with 25m/s
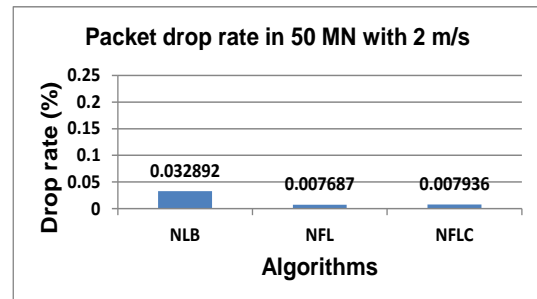
Figure 6-80: Total packet drops in 50 MNs with 2m/s

The packet drops in neural-fuzzy load balancing algorithm is considerably low as compared to the scenario without load balancing and neural-fuzzy load balancing with cost. As Figure 6-79 shows that packet drop rate using neural-fuzzy load balancing algorithm is 0.0342% and with fuzzy using cost is 0.0378% both of which are fairly less than that the no load balancing scenario which is 0.2%. The cause for higher drops in scenario without load balancing is that, this algorithm moves all the mobile nodes to the newly

detected better network by means of network latency, data rate and signal strength. This mounts the congestion on the network where all the mobile nodes handover which results into high number of packet drops in scenarios using no load balancing. The scenarios using neural-fuzzy load balancing with cost offload the satellite network due to mobile nodes preference for the terrestrial network. This redirects most of the mobile users' traffic to terrestrial networks which is the reason fuzzy algorithm with cost shows more drops as compared to the neural-fuzzy load balancing algorithm.

Similarly Figure 6-80 shows the packets drop rate for the 50 mobile user scenarios using 2m/s mobile node speed. In this case no load balancing shows 0.03% packet drop, neural-fuzzy load balancing shows 0.00768% packet drop and neural-fuzzy with cost algorithm shows 0.00798% packets drop. The scenarios with mobile node velocity of 2m/s shows higher number of dropped packets as compared to the scenarios having 25m/s of mobile nodes velocity. This is because the scenarios with 2m/s mobile nodes velocity produce large traffic while covering the same distance as in case of scenarios with 25m/s.

c) Handover Latency

The graphs shown in figures from Figure 6-81 to Figure 6-82 represent the total handover latencies observed by the average mobile nodes in 50 mobile nodes scenarios using no load balancing, neural-fuzzy load balancing and neural-fuzzy load balancing with cost preferences from mobile nodes. All these scenarios are repeated with mobile nodes velocity of 25m/s and 2m/s. The x-axis on these graphs represents the individual mobile nodes and y-axis represents the total handover latencies observed by each mobile node. The

blue, pink and cyan colour dots represent scenario without load balancing, with neural-fuzzy load balancing and with neural-fuzzy load balancing using cost preferences from mobile nodes.
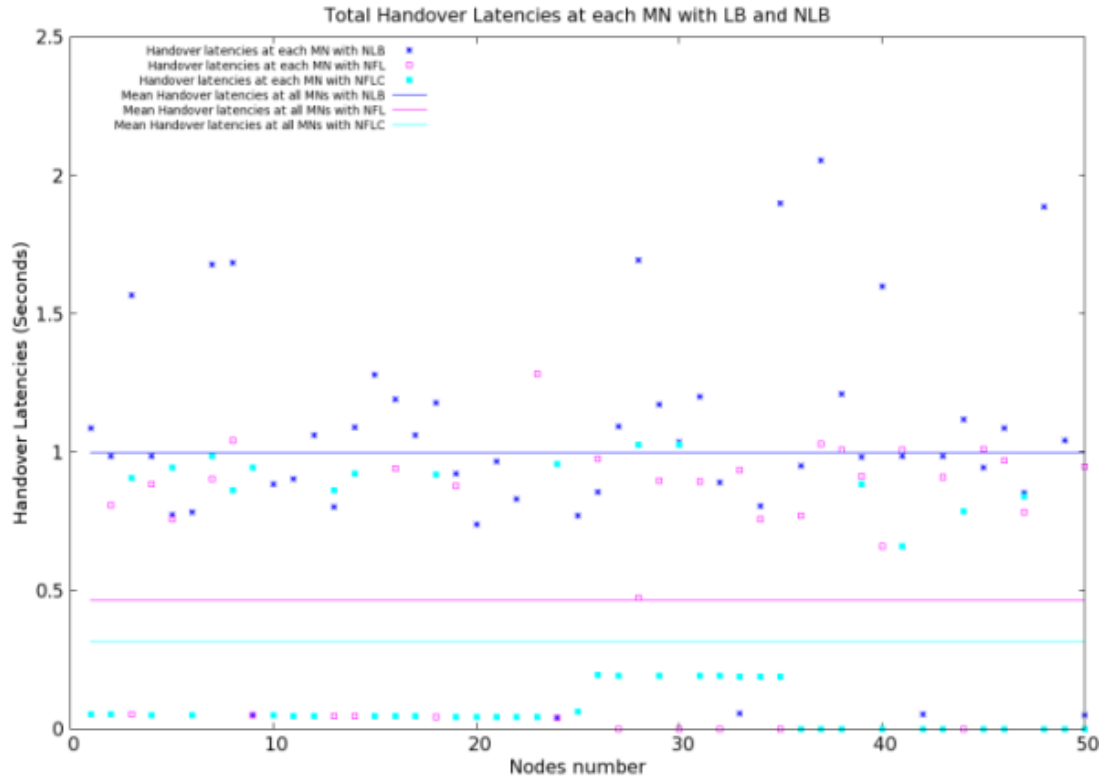
Figure 6-81: Handover Latency (speed - 25m/s)

Figure 6-81 shows that without load balancing the average handover latencies observed by each mobile node is approximately 1 second. Whereas the average handover latencies observed with neural-fuzzy and neural-fuzzy using cost are 0.465 second and 0.35 second respectively. For the scenario with 50 mobile nodes using 2m/s the handover latencies are lower i.e. 0.75, 0.55 and 0.49 for no load balancing, neural-fuzzy and neural-fuzzy with cost algorithms.
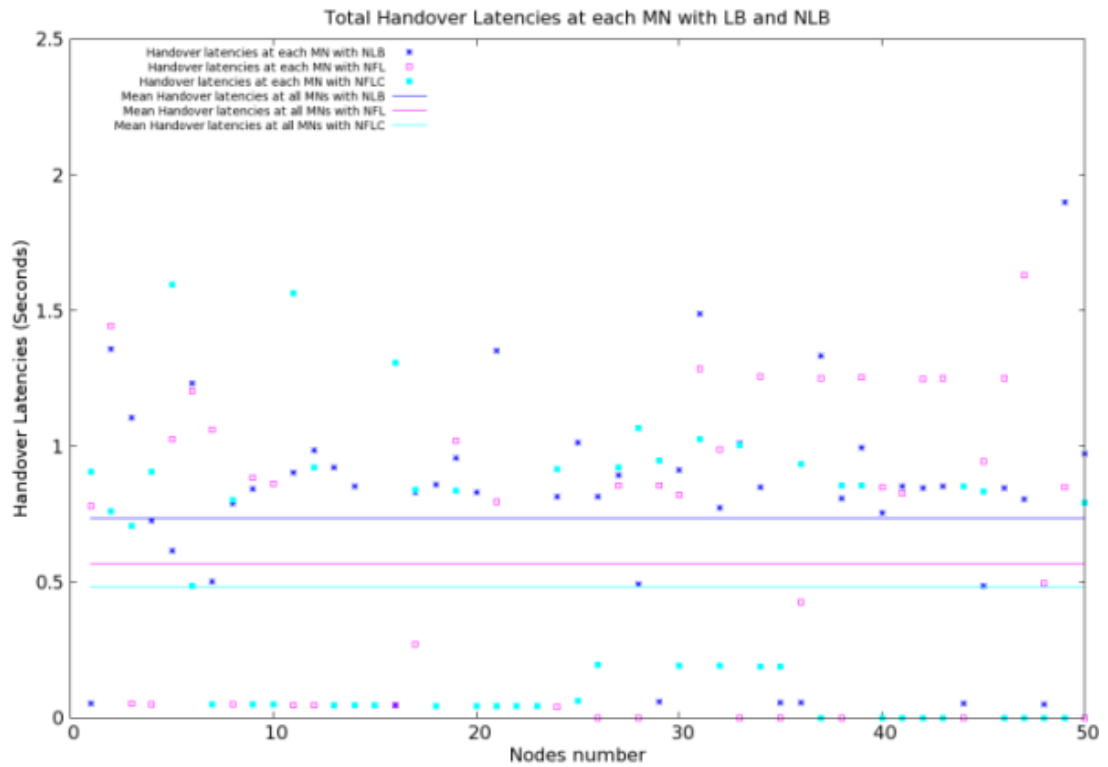
Figure 6-82: Handover Latency (speed - 2m/s)

Comparing the results shown in Figure 6-81 and Figure 6-82, shows that the means for the handover latencies observed at each mobile node throughout the simulation is lower in scenarios with low mobile node velocity which is 2m/s. The mean value for the handover latencies in case of neural-fuzzy with cost is lowest as this algorithm minimizes the number of handovers more than neural-fuzzy due to the cost constraint.

d) **Average throughput at mobile node**

The average throughput of each mobile node in 50 mobile nodes scenarios with 25m/s and 2m/s mobile nodes velocities using no load balancing and neural-fuzzy load balancing is presented in graphs from Figure 6-83 to Figure 6-84. The x-axis on these graphs represents the number of mobile nodes and y-axis represents the throughput in terms of packet per seconds.
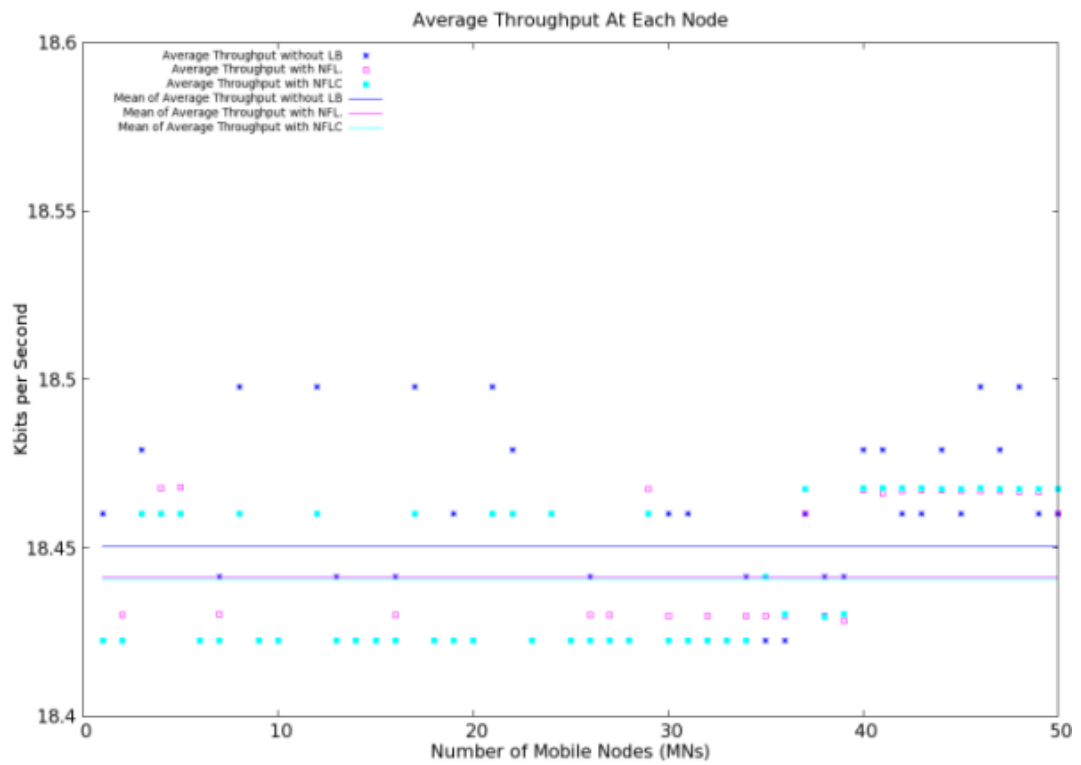
Figure 6-83: Average throughput (speed – 25 m/s)
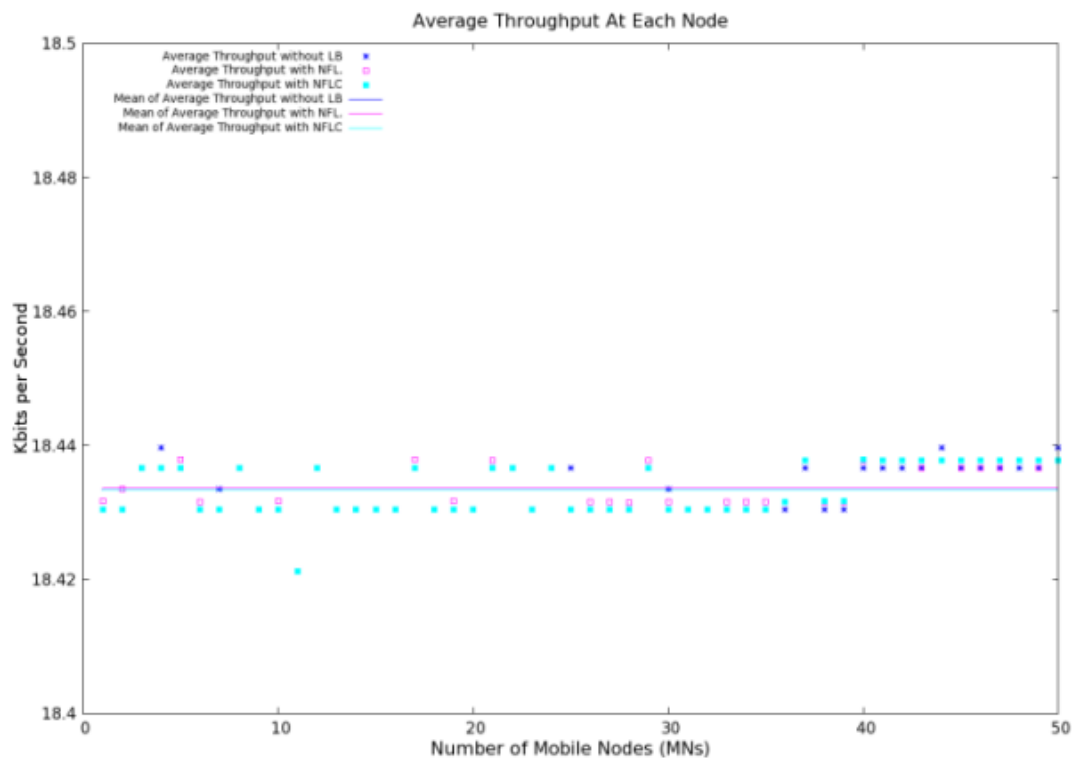
Figure 6-84: Average throughput (speed – 2 m/s)

It can be seen from Figure 6-83 that the average throughput is higher in case of no load balancing at 18.45 kbps as the mobile nodes select the best

available network. On other hand the average throughput for all the mobile nodes in case of neural-fuzzy and neural-fuzzy with cost are almost similar at around 18.44 kbps and only slightly lower than that of no load balancing scenario. The reason for this is that load balancing tries to maintain the load equilibrium between the networks and this practice may result selection of network for some mobile node with high network latencies and lower data rate but only after making sure that the network can fulfil the required QoS of the mobile user.

Comparison of both graphs shown in Figure 6-83 and Figure 6-84 shows that the average throughput at each mobile node using no load balancing is slightly higher as compare to the scenarios with load balancing in case of mobile nodes velocity of 25m/s. Whereas in scenarios with mobile node velocity 2m/s the average throughput of both no load balancing and load balancing algorithms is approximately same. This concludes that the load balancing does not affect the average throughput at each node to a considerable extent and if there is some degradation it is very minute and ignorable.

e) Network Throughput

Figure 6-85 to Figure 6-90 show the throughput at each network for this scenario. These graphs show how the networks are being utilised at different times. Figure 6-85, 6-86 and 6-87 show the throughput of all the networks when the nodes are moving with the speed of 25m/s without load balancing, with neural-fuzzy loaded load balancing and with neural-fuzzy algorithm using cost preferences from users respectively. The x-axis of these graphs represents the time of simulation and the y-axis represents the throughput in

terms of packets per second. The pink colour line represent the satellite throughput, blue colour line represents the UMTS throughput, black colour line represents the WiMax throughput and cyan colour line represents the throughput of WLAN.
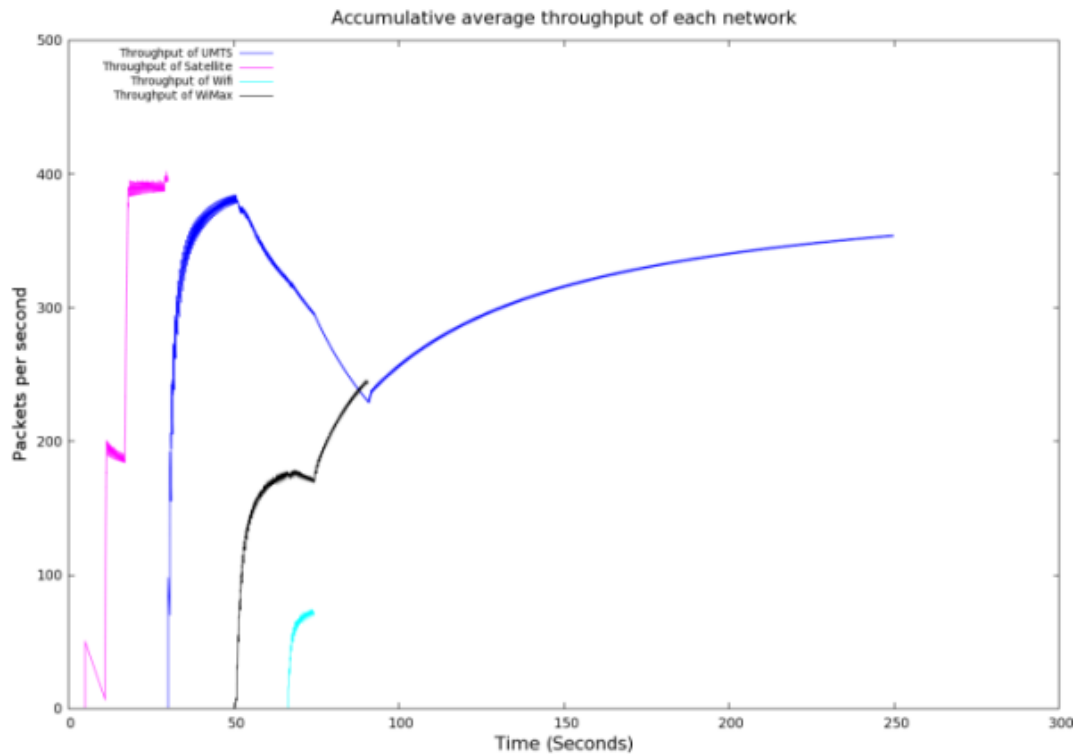
Figure 6-85: Network throughput with no load balancing (speed = 25m/s)

The graph in Figure 6-85 shows that all the nodes handover from satellite to the UMTS at time approximately 30 seconds during simulation. At approximately 51 seconds the mobile nodes enter the WiMax coverage area and leave the WiMax coverage area at time 90 seconds. The traffic in the WLAN network starts at approximately 66 seconds and ends at approximately 74 seconds. As without load balancing the mobile nodes handover to the best available network therefore the average throughput on each network shifts to the newly available better network, whenever the mobile nodes enter the network with low network latencies and high data

rates. The graphs represent the number of TCP packets sent/received per unit time across the simulation.

Figure 6-86: Network throughput with Neural-Fuzzy based load balancing (speed = 25m/s)

Figure 6-85 clearly shows that when no load balancing is used, in the beginning the throughput at the satellite network is, 400 packets/second showing that the satellite link is heavily utilised as all users are on this link. Similarly at 30 seconds when mobile nodes enter the common coverage area of satellite and UMTS networks, the traffic shifts to UMTS and the UMTS throughput is approximately, 380 packets/second. When the users enter WiMax, we can see the network throughput for WiMax increases to around 180 packets/sec, while the network throughput of UMTS network decreases. This is because there are still some users in the UMTS network. When the users enter WLAN coverage area, we can see from Figure 6-73 that all the

user's move into WLAN. However the network throughput at WLAN is only at around, 80 packets/second due to the high congestion in the network and resulting packet drops.
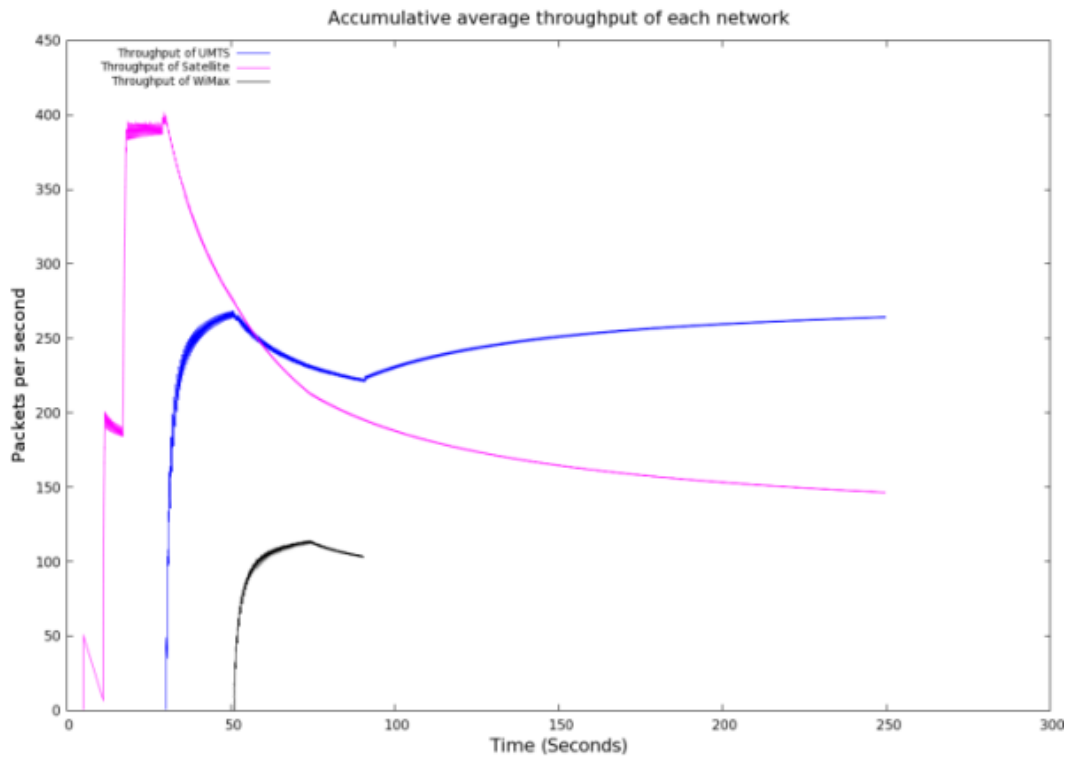
Figure 6-87: Network throughput with Neural-Fuzzy based load balancing with (speed = 25m/s)

On the other hand Figure 6-86 shows the traffic on all networks when the neural-fuzzy load balancing algorithm is applied. In this case until 30 seconds when all the mobile users are in satellite only coverage area, the satellite average throughput is approximately, 400 packets/second. At time 30 seconds when the mobile nodes enter the common coverage area of satellite and UMTS networks the traffic in satellite network decreases and the traffic in UMTS network increases (to approximately 200 packet/second) as the load is now shared between satellite and UMTS networks.

Comparison of Figure 6-85 and Figure 6-86 shows that when load balancing is applied the traffic in other networks is lesser (e.g.: UMTS is around 200

packets/second which earlier was around 380 packets/second) as the satellite network shares the load with terrestrial networks throughout the simulation time. This shows the benefit of load balancing algorithm for sharing the load between networks to avoid the congestion situation. One major change in this graph is that with neural-fuzzy load balancing the WLAN does not get any user as the speed of mobile nodes is 25m/s which is high enough to pass the WLAN coverage area in a very short time. This is detected by the neural-fuzzy algorithm intelligently and therefore it did not allow any mobile node to handover to the WLAN, while passing through WLAN coverage area.

Figure 6-87 represents the traffic in each network when neural-fuzzy load balancing with cost is applied. Comparing Figure 6-85 with Figure 6-87 shows the decreases in satellite traffic and increase in the other terrestrial networks traffic (e.g.: UMTS is around 270 packets/second which with no load balancing was around 380 packets/second but with neural-fuzzy was around 200 packets/sec) due to the fact that most mobile nodes do not want to pay for satellite when they use terrestrial networks. This shows that the cost can degrade the efficiency of load balancing algorithms only slightly, but it is still far better than no load balancing.

Similarly Figure 6-88, 6-89 and 6-90 represent the throughput in each network for the 2m/s scenario. Without load balancing the satellite network throughput is approximately 400 packets/second until 250 seconds. At 250 seconds the mobile nodes enter the common coverage area of satellite and UMTS and all the mobile nodes handover to the UMTS network making the throughput in UMTS to approximately 240 packets per second. At 511

212

seconds the mobile nodes enter the WiMax coverage area, therefore the traffic shifts from UMTS to WiMax making the WiMax throughput to approximately 70 packets/second and increasing.
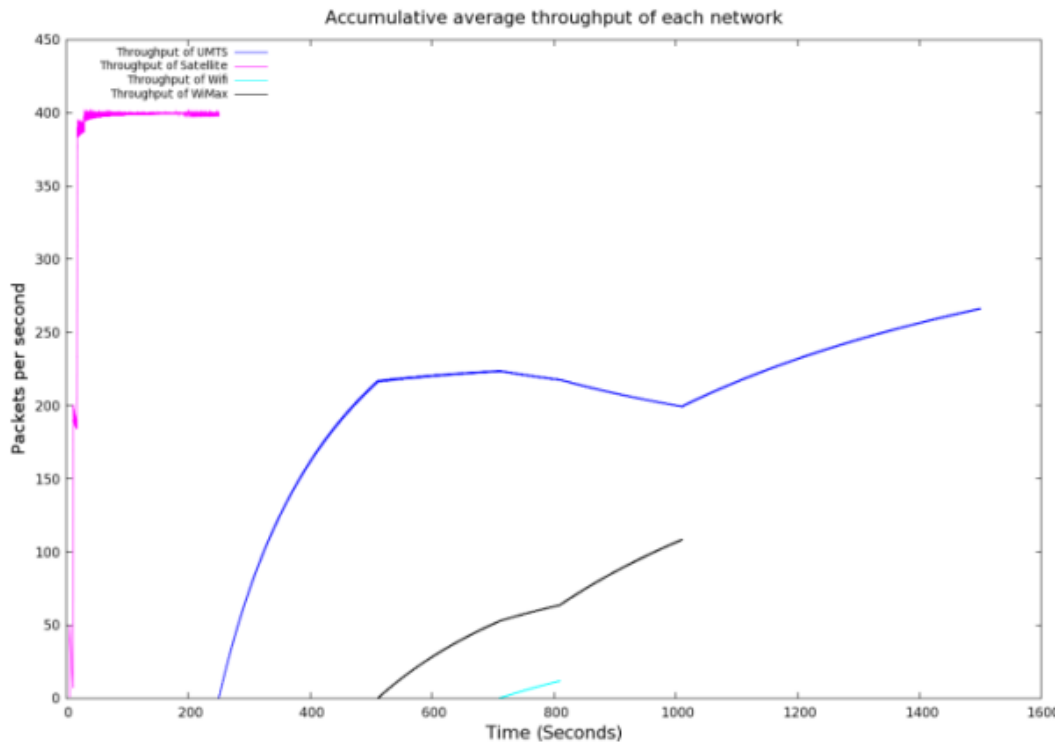
Figure 6-88: Network throughput with no load balancing (speed = 2m/s)

At 710 second the WLAN network appears in the trajectory of mobile nodes again the traffic shifts from WiMax to WLAN making the throughput at WLAN approximately 20 packets/second. At 810 seconds the mobile nodes leave WLAN coverage area shifting the traffic back to WiMax make the throughput on WiMax approximately 110 packets/second. At 1010 seconds mobile nodes leave the WiMax coverage area, shifting all the traffic to UMTS network making throughput at UMTS nearly 270 packets/second.
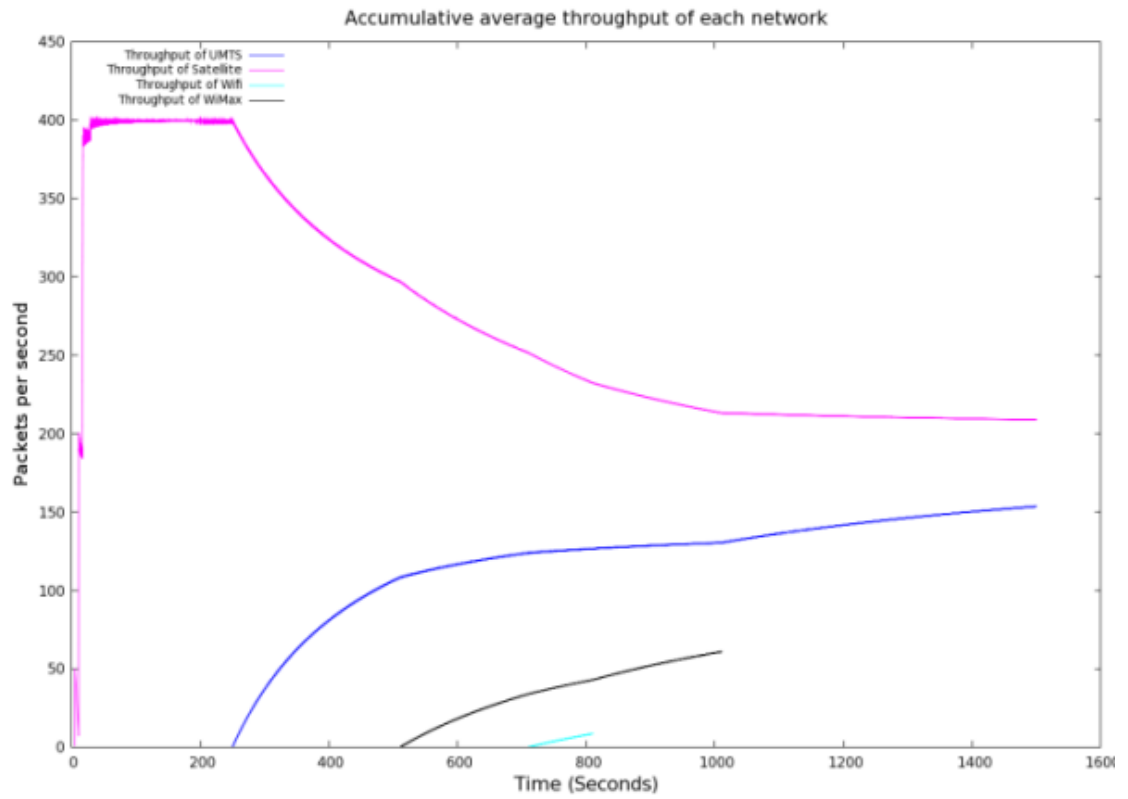
Figure 6-89: Network throughput with Neural-Fuzzy based load balancing (speed = 2m/s)
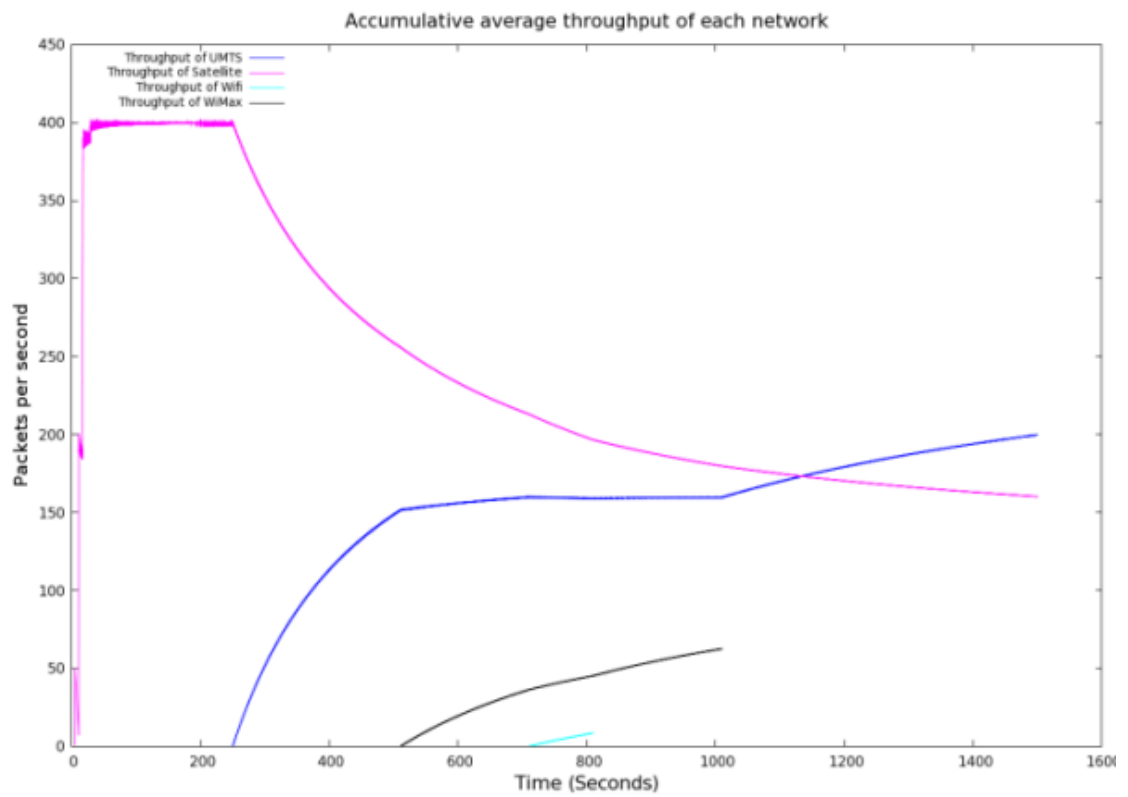
Figure 6-90: Network throughput with Neural-Fuzzy based load balancing with cost (speed = 2m/s)

The scenario with mobile nodes velocity 2m/s show the similar behaviour as the neural-fuzzy load balancing algorithm shares the load between co-located networks. However in case of scenario with 2m/s the neural-fuzzy algorithm utilised the WLAN when they pass through the coverage area of WLAN. In case of neural-fuzzy load balancing the traffic is shared between co-located networks as shown in Figure 6-89. It shows that when load is shared between satellite and terrestrial networks the average throughput at satellite, UMTS and WiMax network reduces to 210 packets/second, 155 packets/second and 70 packets/second approximately. This reduction of traffic in each network by sharing the load between different available networks minimizes the chances of congestion and hence improves the performance. Figure 6-90 represents the average throughput in each network when neural-fuzzy load balancing is applied with cost. It shows the effects of cost on balancing the load in co-located wireless networks as the average throughput in satellite reduced to 160 packets/second and in UMTS and WiMax increased 200 packets/second and 60 packets/second approximately. This concludes that high network cost and users preferences towards using the inexpensive available networks may degrade the load balancing but the overall results with load balancing with cost are still improved as compared to the no load balancing. As the load balancing still share the traffic and avoids or minimizes the chances of congestion on the co-located networks using all available options.

6.2.6 Scenario 6 – Neural-Fuzzy based algorithm with 100 users

The results for scenario 6 having 100 mobile nodes with mobile nodes velocities of 25m/s and 2m/s using no load balancing, neural-fuzzy load balancing and neural-fuzzy load balancing with cost preferences from mobile nodes are presented in this section. The same four parameters are monitored for analysis such as network load, packet drops, total handover latencies, average throughput of each mobile node and throughput of all the networks. The obtained results are explained as follows:

a) Network load

In 100 mobile nodes scenarios first set is simulated using high speed of 25m/s and the second set is simulated using low speed of 2m/s. The graphs shown in Figure 6-91 to 6-96 represent the load in each network such as satellite, UMTS, WiMax and WLAN at the different position of travel trajectory for the 100 mobile nodes scenarios with the mobile nodes moving at 25m/s and 2m/s.

These graphs show the load for the simulation scenarios when no load balancing is applied, neural-fuzzy load balancing is applied and when neural-fuzzy load balancing algorithm is applied with cost preferences. The x-axis of the graphs represents the selected time points where the load in each simulation scenario is monitored. The y-axis of the each load graph represents the total load in terms of number of users in that particular network.
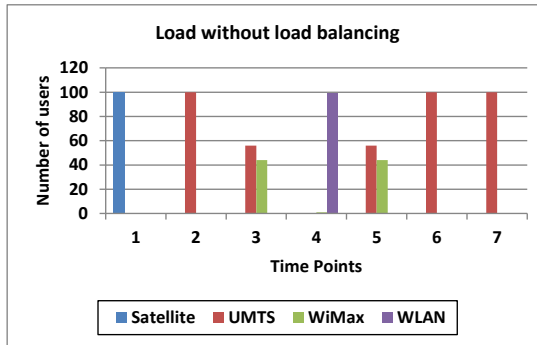
Figure 6-91: Load distribution without load balancing in 25m/s
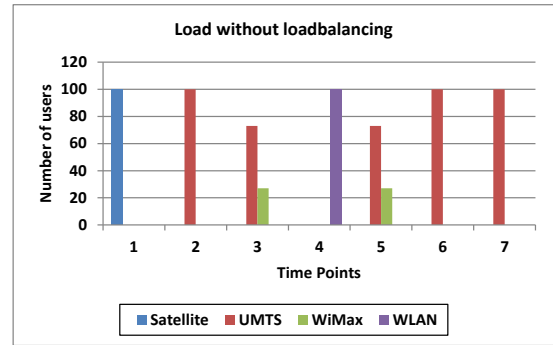
Figure 6-92: Load distribution without load balancing in 2m/s
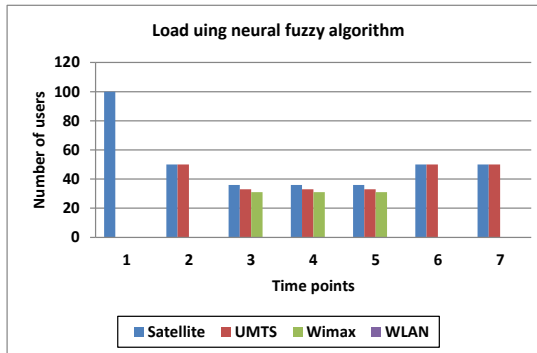
Figure 6-93: Load distribution with neural-fuzzy load balancing in 25m/s
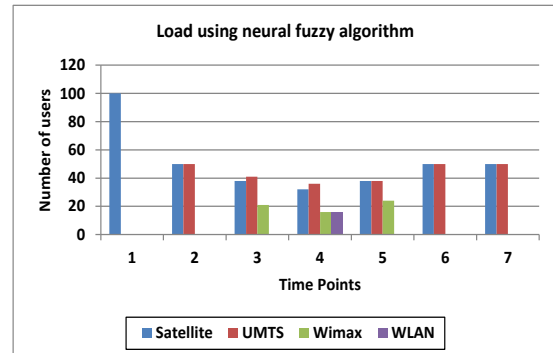
Figure 6-94: Load distribution with neural-fuzzy load balancing in 2m/s
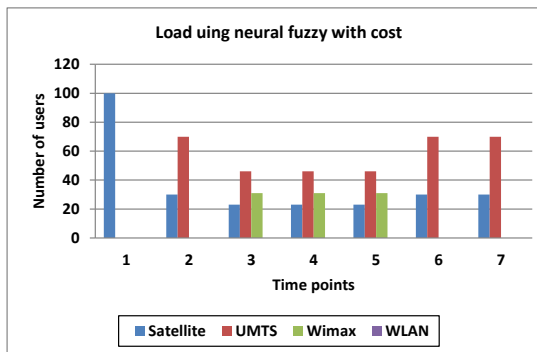
Figure 6-95: Load distribution with fuzzy using cost in 25m/s
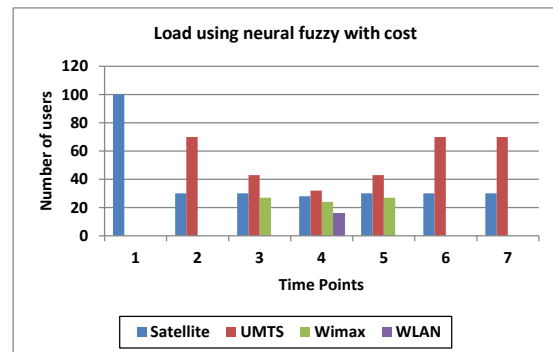
Figure 6-96: Load distribution with fuzzy using cost in 2m/s

It can be seen from these obtained load results that without load balancing most of the mobile nodes handover to the best available network in terms of cost and network latencies. For example in position 4 in Figure 6-91 and 6-92, we can see that all the users connected to WLAN. This however leaves the other networks under-loaded or underutilised. In the scenario where neural-fuzzy load balancing is applied all the networks share the load where possible such as in the overlapped coverage areas. For example in the same

position 4 in Figure 6-93 and 6-94, we can see that the users are distributed across the different networks except the WLAN in Figure 6-93. The reason behind no user at WLAN is that the mobile nodes are moving with high speed of 25m/s which passing through the WLAN coverage area and the neural-fuzzy algorithms has intelligently decided not to handover mobile nodes to WLAN as the mobile nodes would not spend considerable amount of time in WLAN.

Figure 6-95 and Figure 6-96 represent the load in different networks when neural-fuzzy load balancing is applied with cost preferences. It can be seen that the network cost affects the load balancing in these scenarios. Looking at the points P2, P3 and P4 in Figure 6-93 shows that load at these points is equally distributed among the available networks, however in Figure 6-95 the same points show less load in satellite network and higher loads in other networks. This is due to the higher cost of satellite networks as compared to the other networks.

b) Packet drops

The packet drop rate in 100 mobile nodes scenarios are shown in Figure 6-97 and Figure 6-98 given as follows. These Figures represent that packet drops are higher in case where no load balancing is applied. The abbreviations used in the following graphs are as follows: NLB stands for "No Load Balancing" NFL stands for "Neural-Fuzzy Logic" and NFLC stands for "Neural-Fuzzy Logic with Cost". The scenario with neural-fuzzy load balancing suffers least packet drops and the scenario using neural-fuzzy with cost possess packet drops higher than neural-fuzzy and lesser than no load balancing scenario. The x-axis on graphs given in Figure 6-97 and Figure 6-

98 represents the algorithms such as without load balancing, neural-fuzzy load balancing and neural-fuzzy load balancing with cost.
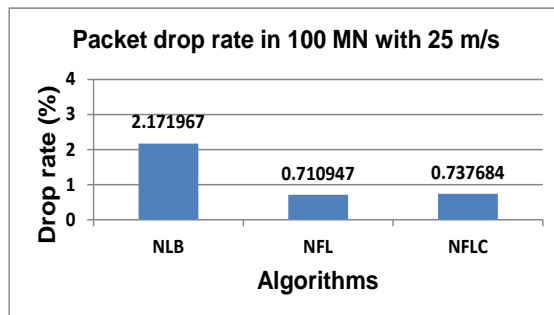
Figure 6-97: Total packet drops in 100 MNs with 25m/s
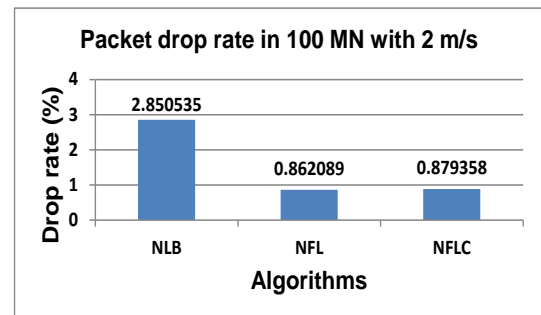
Figure 6-98: Total packet drops in 100 MNs with 2m/s

The packet drop rate represents the ratio of the number of packets dropped per total packets transmitted. It can be seen from Figure 6-97 that the packets drop rate when no load balancing is applied is the highest (2.1%). This is expected as most of the mobile nodes handover to the best available networks, which can cause congestion on that network thereby resulting in the large number of packet drops.

The packets drop rate is lower (0.71%) when the neural-fuzzy load balancing algorithm is applied. This is because the networks are not overloading in this case. When cost preferences are also considered, the packet drops is (0.73%) which is slightly higher than the neural-fuzzy algorithm with no cost as in this case the load in networks are not perfectly balanced due the varying network service cost and mobile node preferences. It can also be seen from the comparison of graphs shown in Figure 6-97 and Figure 6-98 that with 2mps in 100 mobile nodes scenario the packet drops are increased as the nodes stay longer in the networks and generate large amount of data causing congestion and ultimately resulting into higher drop rate.

c) Handover Latency

The graphs for the handover latencies observed by each mobile node in 100 MNs scenario with mobile nodes velocity of 25m/s and 2m/s are presented in this section. The x-axis in this graph represents the individual mobile nodes and y-axis represents the total handover latency observed by mobile nodes in seconds for their complete journey. This latency is the sum of all the delays for the different handovers any given user would be subjected to during its movement across the travel path.

Figure 6-99: Handover Latency (speed - 25m/s)

The blue, pink and cyan colour dots represent the total handover latencies of different mobile nodes using no load balancing, neural-fuzzy load balancing and neural-fuzzy load balancing with cost. Similarly the blue, pink and cyan lines show the mean value of the handover over latencies observed at all the mobile nodes using no load balancing, neural-fuzzy load balancing and

220

neural-fuzzy load balancing with cost. Figure 6-99 represents the total handover latencies of each mobile node in scenario using mobile nodes velocity of 25m/s. It is shown that the average handover latency for no load balancing is highest at around 0.6 second, as in this case most of the mobile nodes handover to the best available network upon entering the common coverage areas. Hence a large number of handovers take place thereby resulting in this large overall delay. The mean values for the neural-fuzzy load balancing without and with cost preferences are much lower than this at around 0.51 second and 0.42 second. This is due to the fewer handovers that take place in these cases.
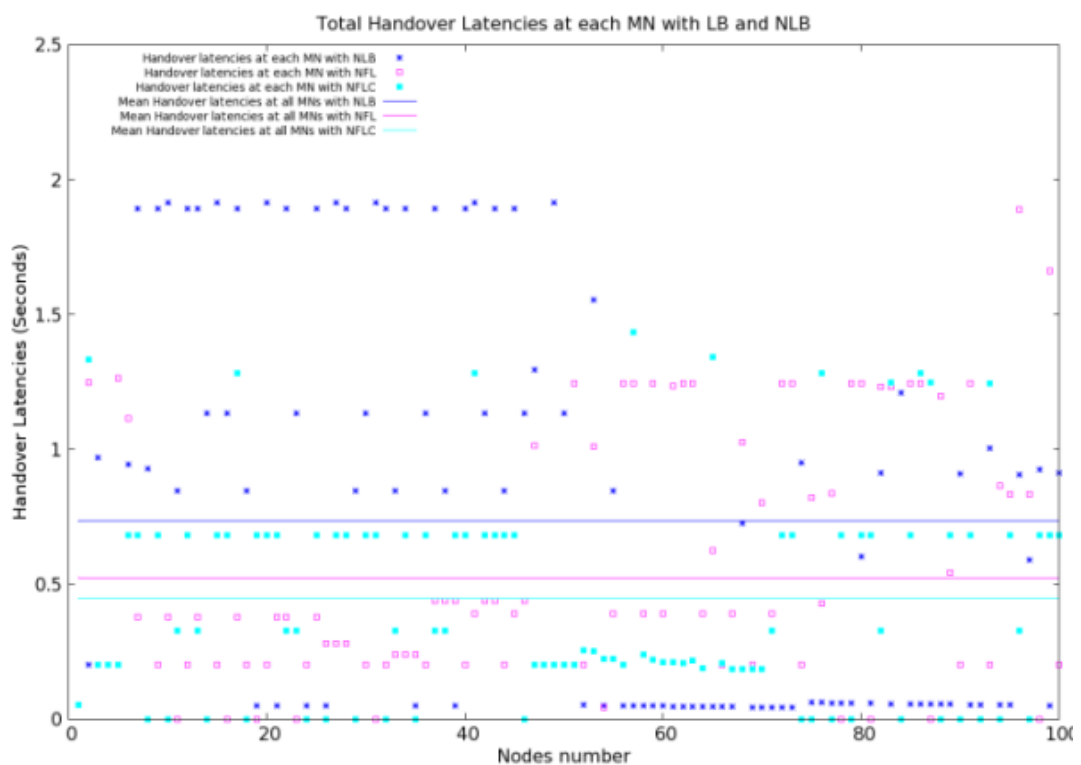
Figure 6-100: Handover Latency (speed - 2m/s)

Similarly Figure 6-100 shows the handover latencies when the mobile nodes are travelling at 2m/s. It can be seen from this graph that without load balancing the handover latencies are highest and the handover latencies for

neural-fuzzy and neural-fuzzy with cost are lesser as compared to the no load balancing scenario.

On comparing the graphs in Figure 6-99 and 6-100, we can also see that for the lower speed scenario the handover delays are lower. This is because as explained in the previous sub-section the packet drops are higher in the high speed scenario which also affects the handover procedure thereby requiring retransmissions of lost control messages during the handover process. The neural-fuzzy with cost has least mean value for the total handover latencies observed by each mobile node. The reason for this is that in case of no load balancing all the mobile nodes perform handover whenever they detect better network or when they leave the coverage are of that network. However in case of neural-fuzzy load balancing only a selected set of mobile users perform the handover in order to maintain the load equilibrium between the networks having common coverage areas or overlapping coverage areas. In case of neural-fuzzy with cost the number of handover are further reduced as most of the mobile nodes do not want to go back to satellite network when the terrestrial network is available.

d) Average throughput at mobile node

The average throughput at each mobile node using no load balancing, neural-fuzzy load balancing and neural-fuzzy load balancing with cost preferences from mobile nodes for the 100 mobile nodes scenarios are presented in Figure 6-101 and Figure 6-102. The 100 mobile nodes scenarios are repeated using mobile nodes velocity of 25m/s and 2m/s. The x-axis on these graphs represents the number of mobile nodes and the y-axis represents the throughput in terms of kilobit per second. Like other graphs

the blue, pink and cyan colours represent readings for no load balancing, neural-fuzzy load balancing and neural-fuzzy load balancing with cost preferences from mobile nodes.
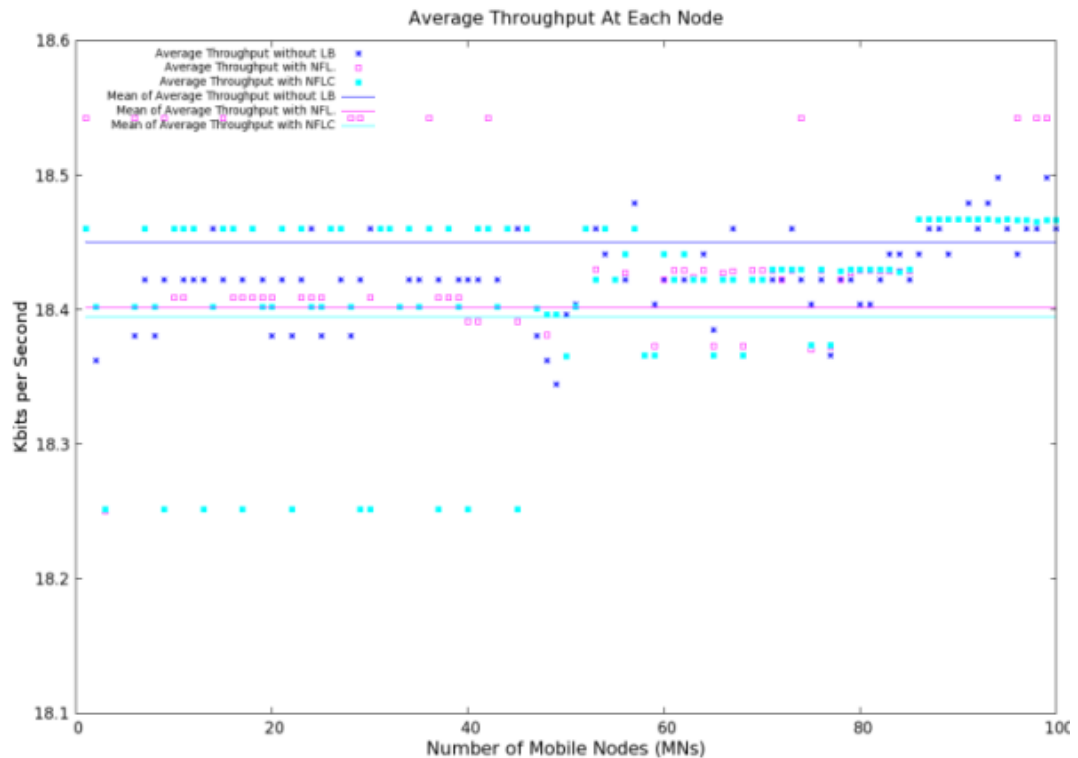
Figure 6-101: Average throughput (speed – 25 m/s)

It can be seen from Figure 6-101 that the average throughput is higher in case of no load balancing at 18.45 kbps as the mobile nodes select the best available network. On other hand the average throughput for all the mobile nodes in case of neural-fuzzy and neural-fuzzy with cost are almost similar at around 18.41 kbps and 18.39 kbps which is only slightly lower than that of no load balancing scenario. The reason for this is that load balancing tries to maintain the load equilibrium between the networks and this practice may result selection of network for some mobile node with high network latencies and lower data rate but only after making sure that the network can fulfil the required QoS of the mobile user. Figure 6-102 shows that the average

223

throughput is 18.45 kbps, 18.39 and 18.35 Kbps in cases of no load balancing, neural-fuzzy load balancing with cost and with neural-fuzzy load balancing when moving at 2m/s.
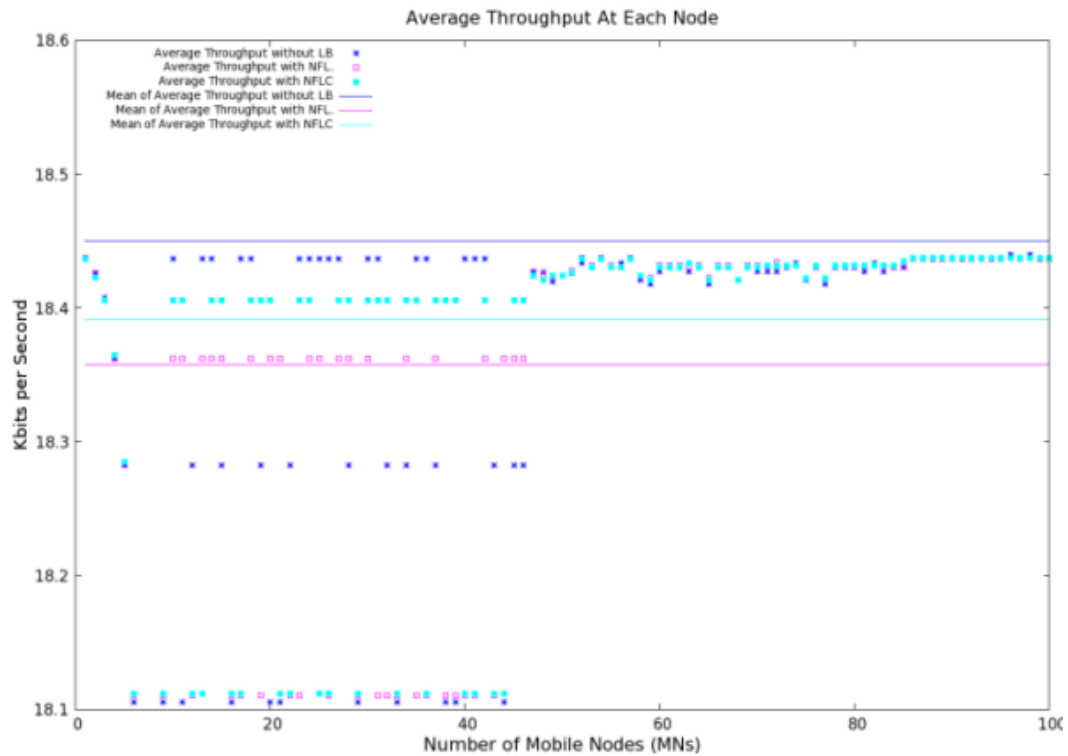
Figure 6-102: Average throughput (speed – 2 m/s)

The comparison of both graphs shown in Figure 6-101 and Figure 6-102 shows that the mean value of average throughput at all mobile nodes using no load balancing does not change when the speed of the mobile nodes is increased in the simulation scenarios. However the mean value of average throughput at all mobile nodes using neural-fuzzy load balancing decreases slightly when the speed of mobile node is reduced. This small reduction in mean value of average throughput at all the mobile nodes is very minute and can be easily ignored. It can also be seen on comparing the two graphs that the use of load balancing does not really affect the throughput of the users who can still access their services properly while at the same time the loads across the networks are more uniformly balanced.

e) Network Throughput

The throughput of all the networks such as satellite, UMTS, WiMax and WLAN is shown in Figure 6-103, Figure 6-104, and Figure 6-105 for the 100 mobile nodes scenarios with mobile modes velocity of 25m/s. The x-axis on these figures represents the simulation time and the y-axis represents the throughput such as packets per second. The Pink line represents the average throughput for the satellite network, blue line represents the average throughput for UMTS network, black line represents the average throughput for WiMax network and cyan line represents the average throughput for WLAN.
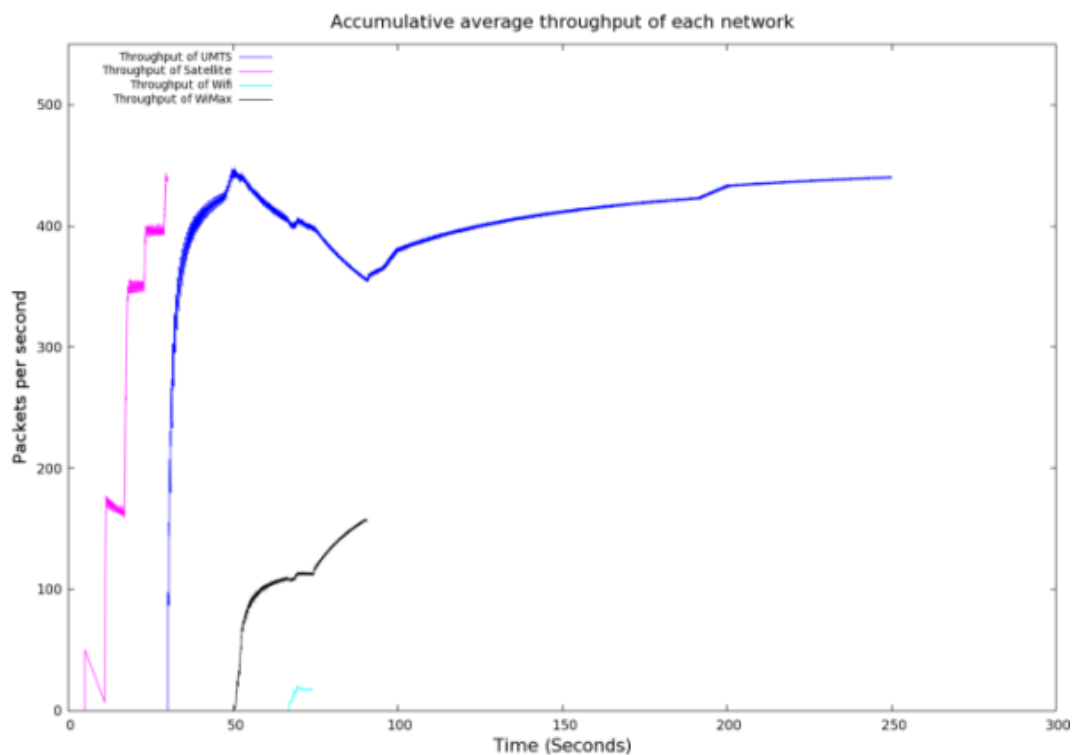
Figure 6-103: Network throughput with no load balancing (speed = 25m/s)

Figure 6-103 shows the average throughput of different networks for 100 mobile nodes scenarios using no load balancing with mobile nodes velocity of 25m/s. It shows that with no load balancing traffic is shifted to the best

available networks when the mobile nodes enter or leave coverage areas or detect a new network.

Figure 6-104: Network throughput with Neural-Fuzzy based load balancing (speed = 25m/s)

Figure 6-104 represents the average throughput of different networks when neural-fuzzy load balancing is applied. In this case the traffic is partially shifted to the newly detected networks in order to maintain the load equilibrium between different networks having common coverage area or overlapped coverage area. In this way all the networks are being utilized on availability. Figure 6-105 shows the average throughput of different networks when neural-fuzzy load balancing with cost is applied.
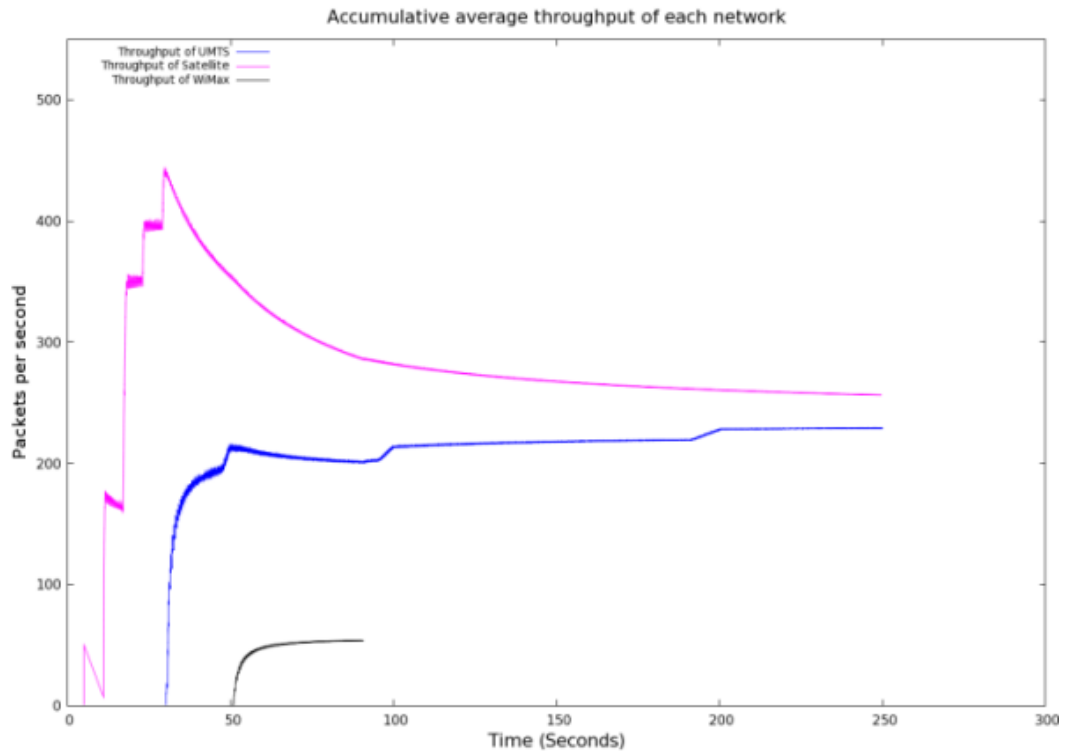
Figure 6-105: Network throughput with Neural-Fuzzy based load balancing with cost (speed = 25m/s)

In this case the satellite network serves less number of mobile nodes when the mobile nodes are in the common coverage area of satellite and other terrestrial networks due to the cost. Therefore the load in UMTS network goes comparatively higher (at approximately 210 packets/second) when mobile nodes are in common coverage area of UMTS and satellite networks. This shows a considerable growth in average throughput of UMTS and for satellite it shows lower average throughput as compare to the neural-fuzzy load balancing scenario shown in Figure 6-104.

For 100 mobile nodes scenario with mobile nodes velocity of 2m/s, the following Figures from Figure 6-106 to Figure 6-108 are showing the average throughput at all the networks. Figure 6-106 shows the average throughput at all the networks using mobile node at velocity of 2m/s without load balancing algorithm. It can be seen that in 2m/s scenario the average throughput for

satellite and UMTS is higher. As the mobile nodes enter the common coverage areas of different networks they handover to the best available networks.

Figure 6-107 shows the average throughput of all the networks when neural-fuzzy load balancing is applied. This shows that average throughput of the satellite network is higher (at around 470 packets/seconds) as all the mobile nodes are in the satellite only coverage area at the beginning. When the mobile nodes enter the common coverage area of satellite, UMTS, WiMax and WLAN; they share the load uniformly. Therefore the all the networks are being utilized throughout the simulation.

Figure 6-108 shows the average network throughput for different networks using fuzzy load balancing with cost. This scenario assumes that 70 mobile nodes do not want to use satellite network if they can use the other terrestrial networks. For the first 250 seconds in this scenario all the mobile nodes stay in the satellite network as they cannot use any other network in that area this makes the throughput in satellite as 470 packets/second. Once they see the UMTS coverage area 70 mobile nodes handover to the UMTS making throughput at UMTS 220 packets/second and satellite at 270 packets/second. Further ahead in the simulation when the mobile nodes move towards the common coverage area with other networks, it is allowed to move mobile nodes to other networks from satellite networks to balance the load but no mobile node is handover to the satellite if the mobile node prefer terrestrial network.
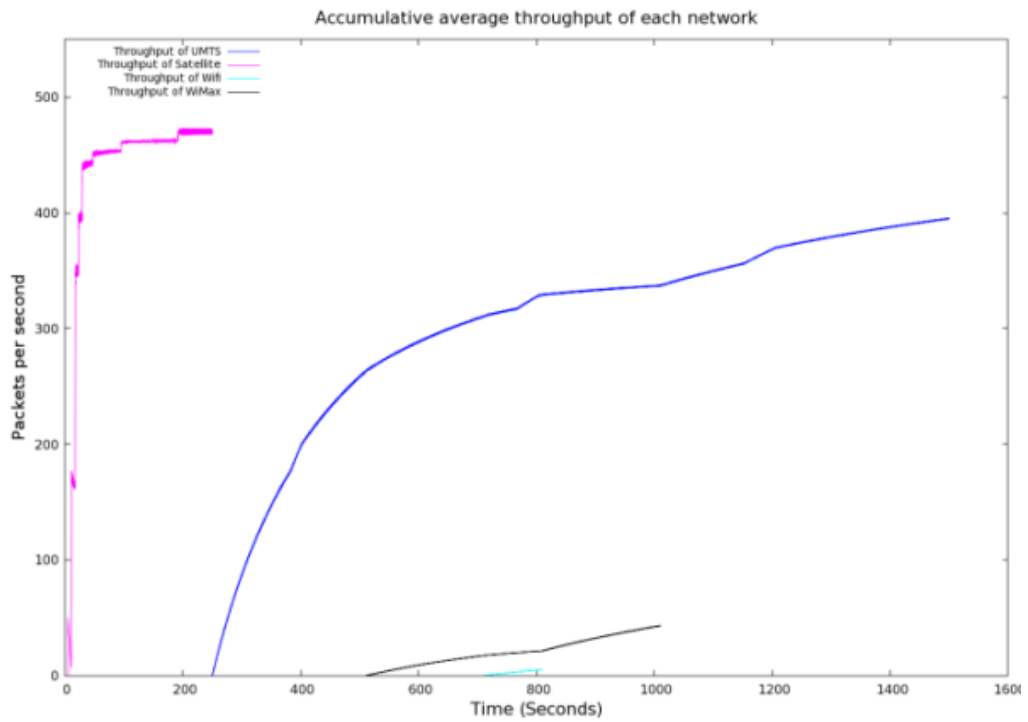
Figure 6-106: Network throughput with no load balancing (speed = 2m/s)
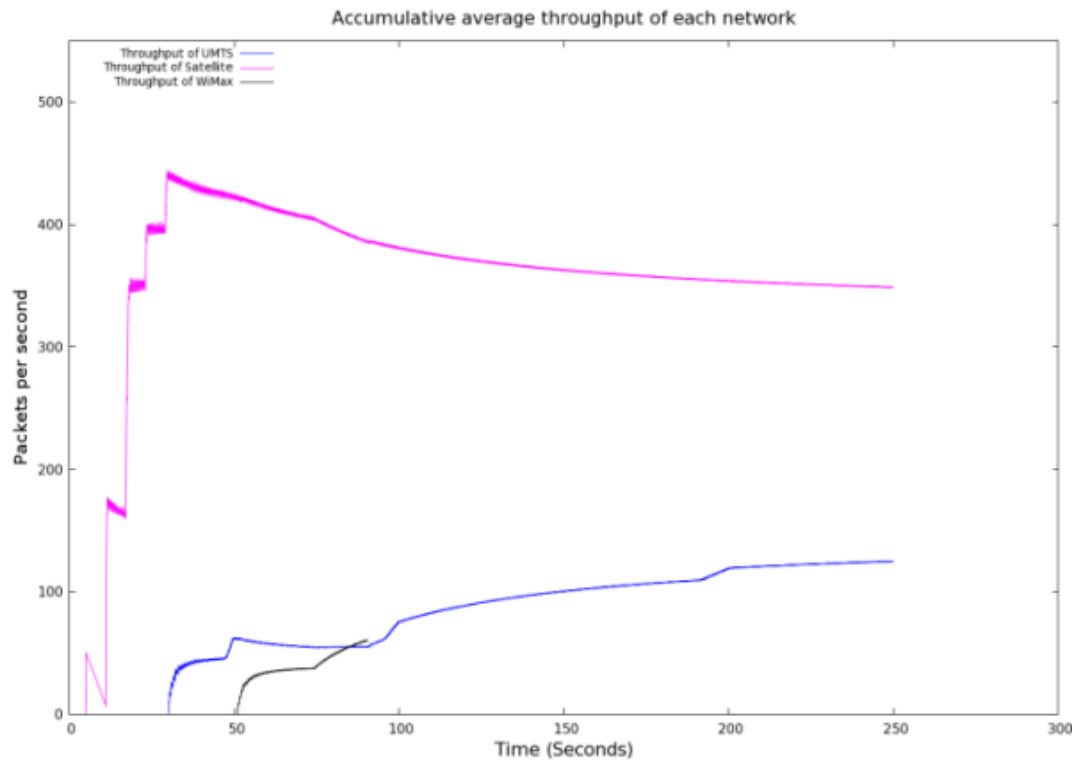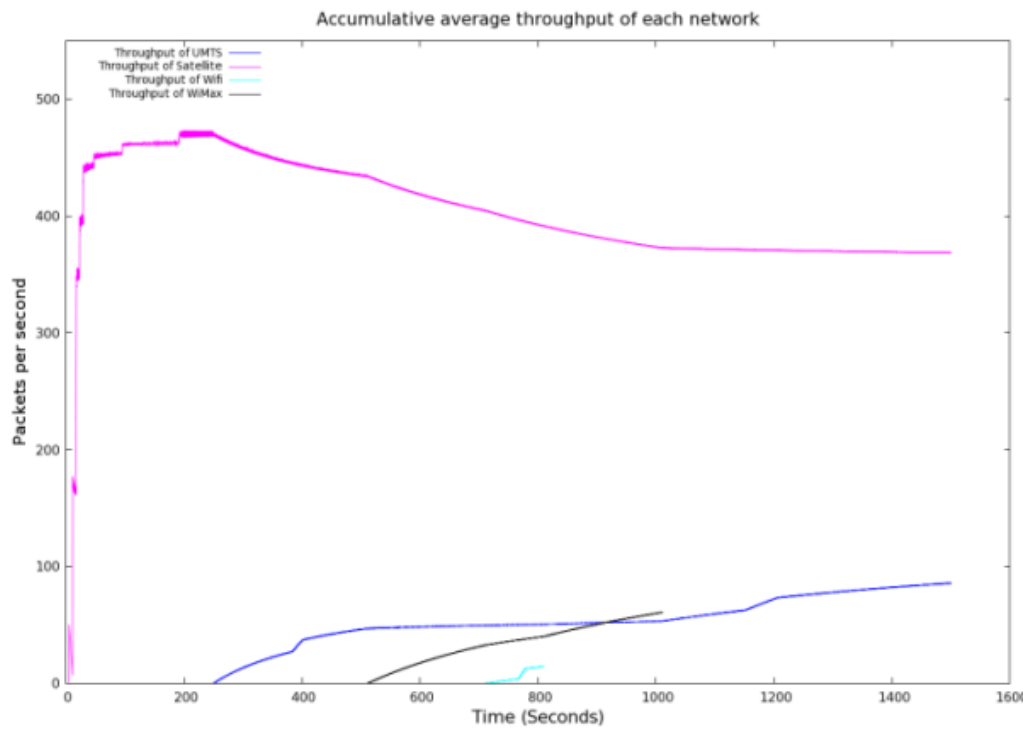

Figure 6-107: Network throughput with Neural-Fuzzy based load balancing
(speed = 2m/s)
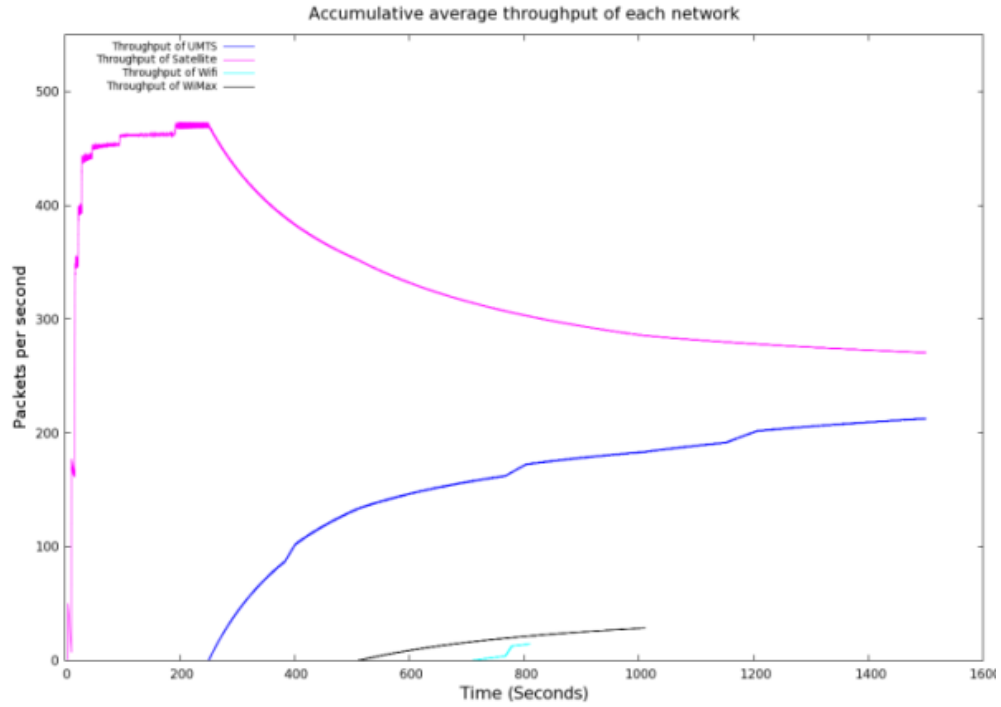
Figure 6-108: Network throughput with Neural-Fuzzy based load balancing with cost (speed = 25m/s)

This is the reason that average throughput of satellite network decreases in this scenario and that of UMTS increases as UMTS servers all the other mobile nodes which do not prefer the satellite in UMTS satellite coverage area.

Comparing Figure 6-107 and Figure 6-108 shows the dramatic change in the throughput of satellite network and UMTS network and reason is cost of satellite network and most mobile nodes preference for the terrestrial network. Therefore when mobile nodes move to the satellite terrestrial common coverage area, most of them prefer UMTS over satellite network.

6.3 Performance comparison of algorithms

This section presents a detailed comparison of the performance of the three proposed load balancing algorithms presented in the previous section. The results of all the load balancing algorithms are compared with results obtained from no load balancing algorithm to prove that proposed load

balancing algorithms are better than the technique using no load balancing for RAT selection and also compared with each other in order to find the most suitable load balancing algorithm.

a) Handover latencies comparison

Table 6-1 represents the mean value of the total handover latencies observed by each mobile node in all different scenarios using baseline, fuzzy and neural-fuzzy load balancing algorithms.

Table 6-1: Comparison of mean values for the total handover latencies at all node using different load balancing algorithms

Comparison Scenario	MN velocity	Number of MNs	Algorithm	HO latency (second)
A	25 m/s	50	Baseline	0.502643
			Fuzzy	0.485328
			Neural-fuzzy	0.465328
B	25 m/s	100	Baseline	0.684761
			Fuzzy	0.555694
			Neural-fuzzy	0.514697
C	2 m/s	50	Baseline	0.491362
			Fuzzy	0.568219
			Neural-fuzzy	0.558316
D	2 m/s	100	Baseline	0.639243
			Fuzzy	0.619963
			Neural-fuzzy	0.525619

In comparison scenario A, B and D, the neural-fuzzy load balancing algorithm has the least handover latency. However this is not true in comparison scenario C where baseline load balancing appeared better with least handover latencies. This means that the neural-fuzzy load balancing algorithm minimizes the total number of handovers and still manages to balance the load between different co-located networks.

b) Load comparison

The load in each network with 25m/s in 50 nodes scenario is shown as follows:
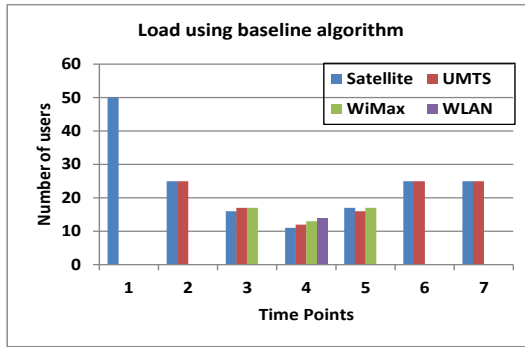
Figure 6-109: Load distribution with baseline load balancing in 25m/s and 50 MNs
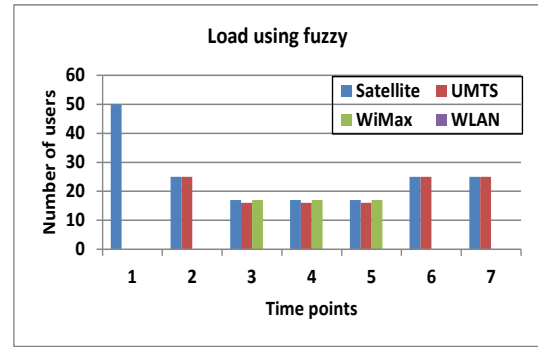
Figure 6-110: Load distribution with fuzzy load balancing in 25m/s and 50 MNs
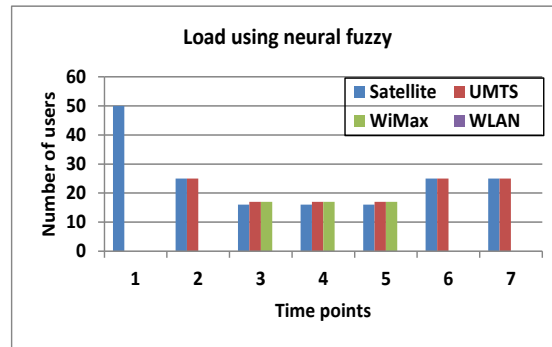
Figure 6-111: Load distribution with neural-fuzzy load balancing in 25m/s and 50 MNs

The graphs shown in figures from Figure 6-109 to Figure 6-111 show the load distribution between different wireless networks using different load balancing algorithm for scenario with 50 mobile nodes and with mobile nodes velocity of 25m/s. It shows that in case of fuzzy and neural-fuzzy the WLAN is not considered in load distribution, as the velocities of mobile nodes are high. The fuzzy and neural-fuzzy algorithms show a minor variation in the load distribution but for both the cases the load between the networks is distributed appropriately.

Similarly the load distribution in scenarios with 100 mobile nodes is shown in the following graphs from Figure 6-112 to Figure 6-114.
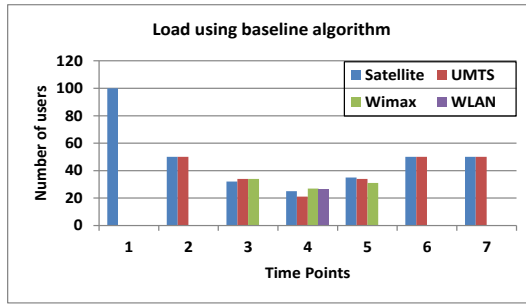
Figure 6-112: Load distribution with baseline load balancing in 25m/s and 100 MNs
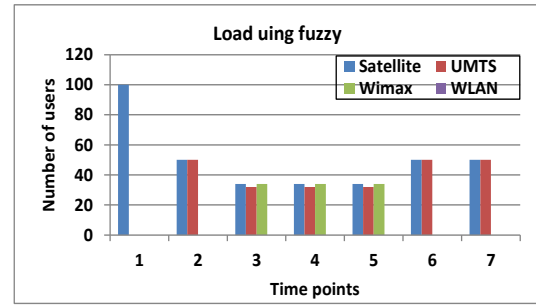
Figure 6-113: Load distribution with fuzzy load balancing in 25m/s and 100 MNs
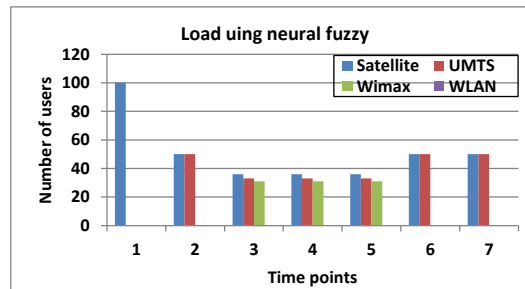
Figure 6-114: Load distribution with neural-fuzzy load balancing in 25m/s and 100 MNs

Like 50 mobile nodes scenario the load distribution in baseline load balancing algorithm is different from fuzzy and neural-fuzzy load balancing algorithms. The load distributions in fuzzy and neural-fuzzy algorithms show very minor variations but in neural-fuzzy this variation controls the total number of handovers and minimizes the total handover latencies. The performance of fuzzy and neural-fuzzy for load distribution is similar but the neural-fuzzy load balancing algorithm leads due to the limited number of handovers.

c) Packet drops comparison

The comparison of packet drop rate for all three proposed load balancing algorithms is shown in the graphs in Figure 6-115 to Figure 6-118. The abbreviations used in the following graphs are as follows: NLB stands for "No Load Balancing", LL stands for "Least Loaded", FL stands for "Fuzzy Logic" and NFL stands for "Neural-Fuzzy Logic".
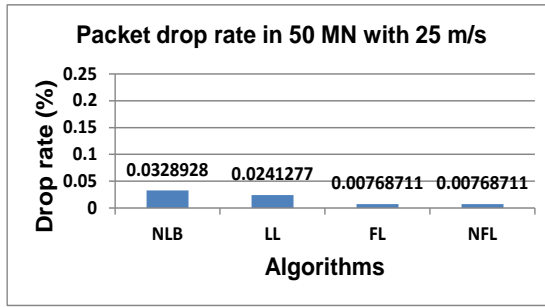
Figure 6-115: Packets drop rate with 50 MN and 25m/s
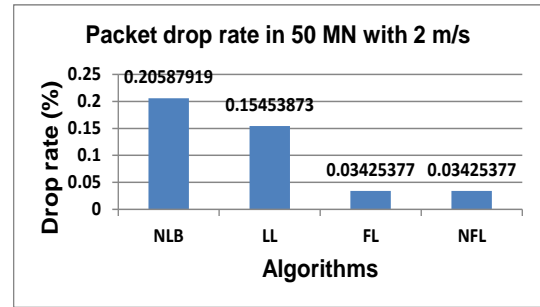
Figure 6-116: Packets drop rate with 50 MN and 2m/s
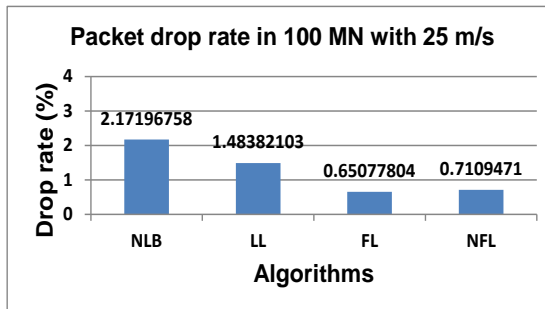
Figure 6-117: Packets drop rate with 100 MN and 25m/s
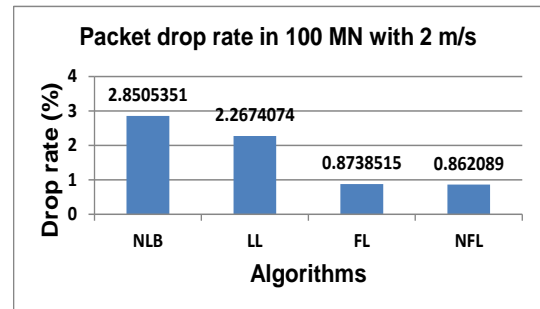
Figure 6-118: Packets drop rate with 100 MN and 2m/s

The comparison of the different approaches shown in the above graphs for packet drop rate, shows that fuzzy and neural-fuzzy have lowest drop rates. In 50 mobile nodes scenarios the packet drop rates show no difference in fuzzy and neural-fuzzy approaches, however in 100 mobile nodes scenario with mobile node speed 25m/s the performance of fuzzy algorithm is slightly better as it shows little less packet drop rate. In 100 mobile scenario with mobile node speed 2m/s the neural-fuzzy algorithm shows little less packet drop rate as compare to the fuzzy algorithm. There is one behaviour which is shown by the above graphs that is the scenarios with low speed i.e. 2m/s suffers from high drop rate as compared to the scenarios with high speed of mobile nodes i.e. 25m/s. The reason for this in these particular scenarios is that when the mobile nodes move slowly they remain in the same network for long time (particularly in WiMax and WLAN where bandwidth is shared between users) and the TCP window for each connection keeps on growing

which cause congestion and eventually results in to more packet drops. On other hand when mobile nodes move with high speed they pass on the coverage area of WiMax and WLAN quickly therefore in these scenarios the packet drop rate is lower. The neural-fuzzy approach has one more advantage that it encompasses lowest handover latencies for average mobile node in all scenarios as shown in Table 6-1. Therefore the neural-fuzzy load balancing algorithm is considered as the most dominant approach overall.

6.4 Summary

In this chapter the results of the target simulation scenarios are discussed with the help of graphs and statistical values in tabular format. In results all the proposed algorithms i.e. baseline load balancing, fuzzy load balancing and neural-fuzzy load balancing algorithms are analysed with the help of obtained results. Each of the proposed load balancing algorithm is simulated with different number of mobile nodes i.e. 50 and 100 and different velocities of mobile nodes i.e. 25m/s and 2m/s. Different parameters such as network load, packet drops, throughput and the handover latencies are monitored for each of the target scenario and results of different algorithms are also compared at the end to conclude which technique is better under different circumstances. The proposed algorithms are also simulated with different cost preferences from mobile nodes to analyse the effect of cost on load balancing.

Comparison of the proposed load balancing algorithms have also been made by considering different parameters such as load distribution in all the networks, packet drop rate and average handover latencies. The fuzzy and neural-fuzzy algorithms showed very close results. However the neural-fuzzy

proved itself better with a very minor margin in different scenarios. It is concluded with the help of results that load balancing improves the performance by avoiding the congestion and other problems which are caused by unbalanced utilisation of available wireless networks. The constraints implied by cost actually affects the load balancing strategy as it limits load balancing process to some extent but does not eliminate the benefits of load balancing.

Chapter 7: CONCLUSION AND FUTURE WORK

7.1 Conclusion

The main objective of this research work is to design a load balancing framework for satellite-terrestrial heterogeneous wireless networks which aims at satisfying the following general requirements:

- Reduce the congestion in the networks by sharing the load between co-located wireless networks.

- Minimize the number of handovers performed by the average mobile node.

- Reduce the total handover latencies observed by the average mobile nodes in the network.

- Efficient utilization of the available radio resource.

- Minimized drop ratio by avoiding the congestion.

- Generating revenue for the network operators by expanding their capacity using all available frequency bands in different wireless access technologies.

The different components of the proposed load balancing framework running on the mobile node and the network side, work together to efficiently balance the load between co-located heterogeneous wireless networks. The centralized CRRM server on the network side and distributed RAT selection algorithms on the mobile nodes and network entities such as RNC, BS and AP work in accordance to provide a constructive framework for load balancing in heterogeneous wireless networks. The extended IEEE 802.21 MIH is incorporated to take advantage of seamless vertical handovers. The

use of IEEE 802.21 MIH in the proposed load balancing framework has made it flexible enough to include any other wireless network in the future. Another advantage of this design is scalability, as the centralised server can keep updated information of all the networks in a uniform manner.

Three new RAT selection algorithms have been proposed, implemented and compared in this thesis for load balancing between heterogeneous wireless networks. The number of parameters used by each algorithm is also flexible as the framework provides these parameters as input to the algorithms and collects the decision form the algorithm that which network is suitable for handover based on load balancing. The proposed RAT selection algorithms are incorporated in the load balancing framework as an integrated module which enables the framework to use any new RAT selection algorithm in the future for comparison purpose. The overall conclusions of these algorithms are as follows:

- **Baseline least loaded algorithm:**

  This algorithm takes input from load balancing framework as a list of parameters like network cost, data rate, signal strength, user's required data rate, user's network preferences (depending upon the cost) and load on each network. It performs the simple if else operations using the input parameters and derives the decision whether to handover on any available network or not. The main advantage of using baseline least loaded algorithm is its simplicity and ease to implement. However it does not give precise output by considering all the input parameters equally. The load parameter is considered highly important in this algorithm and

therefore it may decide to handover to a least loaded network but with weaker signal strength, low coverage and high cost.

- **Fuzzy logic algorithm:**

The shortcomings of baseline least loaded algorithm are eliminated in fuzzy based load balancing algorithm as it considers all input parameters equally while deciding for the load-aware handover. While this algorithm is difficult to implement and its complexity increase with the number of input parameters considered, it was seen that this algorithm performed very well as compared to the baseline algorithm. It was seen that this algorithm could efficiently balance the load across the different networks while maintaining the user experience in terms of high throughput and low packet drops.

- **Neural-fuzzy algorithm:**

The proposed Neural-Fuzzy based algorithm combines fuzzy logic and neural network algorithms to take advantage of both these approaches. The neural-fuzzy load balancing algorithm is designed to get rid of limitations implied by fuzzy logic with the help of training the neural-fuzzy system using all combination of fuzzy rules as input-output training data. The training process turned out very useful for the target neural-fuzzy load balancing system in adjusting the weights for expected output for any set of input parameters. After training, the performance of the neural-fuzzy based load balancing algorithm is compared with the standard fuzzy based load balancing system using both MATLAB and NS2. The neural-fuzzy load balancing performs better as it shows reduced number of handover without considerably degrading the load balancing.

All of these three load balancing algorithms were analysed using different scenarios in NS2 such as with different number of users, variable speed of the mobile user and with variable network cost preferences from each mobile user. The results of each load balancing algorithm are compared with no load balancing scenarios and with other load balancing algorithms. The effects of cost were also monitored using each load balancing algorithm. Different parameters are monitored for analysis such as network load, packets drop rate, handover latencies, average throughput at each mobile node and throughput at different networks.

The comparison of derived results using different load balancing algorithms concluded that load balancing helps in efficient radio resource utilisation in heterogeneous wireless networks by better load distribution, less packet drops minimized congestion on the networks and low handover latencies without making considerable effects on the throughput at mobile nodes. The results obtained from simulation scenarios also conclude that variable network cost for different services and user preferences for inexpensive networks degrades the efficiency of load balancing algorithms, but the results with load balancing are still better as compared to the scenarios with no load balancing.

7.2 Future work

The research work presented in this thesis provides the foundation for future studies in load balancing for heterogeneous wireless networks incorporating the enhanced IEEE 802.21 MIH. This work may be extended and further research studies can be done to improve and enhance the scope of this

research. Some of the potential directions following this research are discussed as follows:

7.2.1 Enhancements in the architecture

The proposed framework targets the network architecture with single operator providing services over multiple access technologies, however there is potential to enhance the architecture design so that collaboration between multiple operators over multiple access technologies can be achieved using the service level agreement (SLA). This consideration can be supported by adapting the strategies discussed in IEEE 1900.4 [126]. The current architecture can also be used with some additional functionality to provide collaboration between multiple operators for load balancing.

7.2.2 New RAT selection algorithms

New RAT selection algorithms can be developed and inserted into the load balancing framework as the design of load balancing is flexible enough to easily integrate newly developed RAT selection algorithms. In the current framework three different algorithms have been employed such as least loaded, fuzzy and neural-fuzzy. However other approaches such as multiple objective decision making (MODM), fuzzy MODM, neural network (NN) based, utility function based and other hybrid strategies can also be implemented to examine if they can perform better using the load balancing framework for heterogeneous wireless networks.

7.2.3 Consideration of real life scenarios

The simulation scenarios which show more resemblance with real life scenarios will be considered in the future for example having multiple WiMax base stations and WLAN access points in the simulation topology. The current simulation model is fully capable of simulating such scenarios therefore it will not require adding new code or modifying the current source code in the simulation model.

REFERENCES

[1]   Panagiotis Dernestichas, G. Vivier, K. El-Khazen, and M. Theologou, "Evolution in Wireless Systems Management concepts: From Composite Radio Environments to Reconfigurability," IEEE Communications Magazine, May 2004, pp. 90 - 99.

[2]   IEEE P802.21/D10.0, "Draft Standard for Local and Metropolitan Area Networks: Media Independent Handover Services", April 2008.

[3]   R. Bekkers, B. Verspagen and J. Smits, " Telecommunications Policy Intellectual property rights and standardization: the case of GSM." 26(3-4): 171-188, 2002.

[4]   B. Ghribi and L. Logrippo, " Computer Networks: Understanding GPRS: the GSM packet radio service", 2002.

[5]   IEE802.11e, "IEEE Std. 802.11eT, Part 11: Wireless LAN Medium Access Control(MAC) and Physical Layer (PHY) specifications; Amendment 8: Medium Access Control(MAC) Quality of Service Enhancements, 2005.

[6]   IEEE802.11, "IEEE Draft STANDARD for Information Technology-- Telecommunications and information exchange between systems-- Local and metropolitan area networks--Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 7: Interworking with External Networks." [10] IEEE Unapproved Draft Std. P802.11u/D5.0, Feb 2009.

[7]   IEEE802.11, "IEEE Draft Amendment Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed and

Mobile Broadband Wireless Access Systems - Advanced Air Interface", IEEE P802.16m/D7 July 2010: 1-932.

[8]   IEEE802.11, "IEEE Draft Standard for Information Technology-Telecommunications and information exchange between systems-Local and metropolitan area networks-Specific requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications-Amendment 10: Mesh Networking", IEEE Std. P802.11s/D5.0 April 2010: 1-305.

[9]   IEEE 802.16, "IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1". IEEE Std. 802.16e-2006 and IEEE Std. 802.16-2004/Cor 1-2005 (Amendment and Corrigendum to IEEE Std. 802.16-2004): 0_1-822.

[10]   IEEE Standard for Local and metropolitan area networks—Part 16: Air Interface for Fixed Broadband Wireless Access Systems, May, 2009.

[11]   TS23.246, "Architecture and functional description (Release 6)", v.6.0.0, October 2003.

[12]   TS 24.007, "3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; Mobile radio interface signalling layer 3; General aspects", Release 8, V8.2.0 (2009-06).

[13]   Digital cellular telecommunications system (Phase2+); Universal Mobile Telecommunications System (UMTS0; AT command set for 3G User Equipment (UE), ETSI TS 127 007, March 2003.

[14]   A. Sayenko, O. Alanen, J. Karhula, and T. Hamalainen, "Ensuring the QoS requirements in 802.16 scheduling", Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems. Terromolinos, Spain, ACM: 108-117, 2006.

[15]   A. Ghosh, D.R. Wolter, J.G. Andrews, and R. Chen, "Broadband wireless access with WiMax/802.16: current performance benchmarks and future potential." Communications Magazine, IEEE 43(2): 129-136, 2005.

[16]   C. Eklund, R.B. Marks, K.L. Stanwood, and S. Wang, "IEEE standard 802.16: a technical overview of the Wireless MAN air interface for broadband wireless access", Communications Magazine, 2002.

[17]   N. Jain, S.R. Das, and A. Nasipuri, "A multichannel CSMA MAC protocol with receiver-based channel selection for multihop wireless networks". Tenth International Conference on Computer Communications and Networks, 2001. Proceedings.

[18]   M. Chatterjee, S. Sengupta and S. Ganguly, "Feedback-based real-time streaming over WiMax", IEEE Wireless Communications, 2007.

[19]   B. Rong, Y. Qian and K. Lu, "Integrated Downlink Resource Management for Multiservice WiMax Networks", IEEE Transactions on Mobile Computing.

[20]   D. Niyato and E. Hossain, "Integration of WiMax and WiFi: optimal pricing for bandwidth sharing". IEEE Communications Magazine, 2007.

[21]   W. Hung-Yu, S. Ganguly, R. Izmailov, and Z.J. Haas,"Interference-aware IEEE 802.16 WiMax mesh networks". Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st.

[22]   C. Hoymann, P. Dallas, A. Valkanas, J. Gosteau, D. Noguet, and R. Hoshyar, "Flexible Relay Wireless OFDM-based Networks", 2005.

[23]   N. Qiang, A. Vinel, X.Yang, A. Turlikov and J. Tao, "WIRELESS BROADBAND ACCESS: WIMAX AND BEYOND - Investigation of Bandwidth Request Mechanisms under Point-to-Multipoint Mode of WiMax Networks". IEEE Communications Magazine.

[24]   WiMax Forum, T32-001-R015v01, T32-002-R015v01, T32-003-R015v01, T32-004-R015v01, T32-005-5 R015v01, "Architecture Tenets, Reference Model and Reference Points" Part0 to Part3 and Abbreviations, 6 Release 1.5.

[25]   B. Celebi, B. Dericiogullari, and Y. Bitirim, "Performance Evaluation of IEEE 802.11b, IEEE 802.11g and GPRS/EDGE Based on Query Retrieval Time", Third International Conference on Wireless and Mobile Communications, 2007, ICWMC '07.

[26]   P. Mahasukhon, H. Sharif, M. Hempel, T. Zhou, W. Wang and H.H. Chen, "IEEE 802.11b based ad hoc networking and its performance in mobile channels", 2009.

[27]   H. Al-Khusaibi, F. Al-Wardi, S. Firas and X. Weidong, "Experiment and Analysis on the Comparison of the IEEE 802.11a and 802.11g

Wireless Local Area Networks". IEEE International Conference on Electro/information Technology, 2006.

[28]   W. Liu, G. Zehua, G. Feng, T. Di and Z. Ronghua,"Performance analysis of IEEE 802.11a in non-saturation conditions". IEEE International Conference on Network Infrastructure and Digital Content, 2009.

[29]   N.T. Dao and R.A. Malaney, "Throughput Performance of Saturated 802.11g Networks", The 2nd International Conference on Wireless Broadband and Ultra Wideband Communications, AusWireless 2007.

[30]   Inmarsat BGAN system descriptions from:

http://www.inmarsat.com/bgan.

[31]   Digital cellular telecommunications system (Phase2+); Universal Mobile Telecommunications System (UMTS0; AT command set for 3G User Equipment (UE), ETSI TS 127 007, March 2003.

[32]   Digital Video Broadcasting (DVB), DVB specification for data broadcasting, ETSI EN 301 192 V1.4.1, 2004.

[33]   Digital Video Broadcasting (DVB) User guidelines for the second generation system for Broadcasting, Interactive Services, News Gathering and other broadband satellite applications (DVB-S2). ETSI TR 102 376 V1.1.1.

[34]   Digital Video broadcasting (DVB) 2nd generation framing structure, channel coding & modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applications (DVB-S2). EN 302 307 V1.2.1.

[35] Digital Video Broadcasting (DVB), DVB-SH Implementation Guidelines, ETSI TS 102 584 V1.2.1, 2011.

[36] Digital Video Broadcasting (DVB), Framing Structure, channel coding and modulation for satellite Services to Handheld devices (SH) below 3 GHz DVB Document A111, May 2011.

[37] Digital Video Broadcasting (DVB), Interaction channel for satellite Distribution Systems, Guidelines for the use of EN 301 790 in mobile scenarios, ETSI TR 102 768 V1.1.1, 2009.

[38] Digital Video Broadcasting (DVB), Interaction channel for satellite distribution systems, ETSI EN 301 790 V1.5.1, 2009.

[39] D. Jones, T. Carroll, M. J. Knapp, "Technical Validation Report - BGAN as an EC2 Enhancement", 15, March, 2006.

[40] C. Makaya and S. Pierre, "An Interworking Architecture for Heterogeneous IP Wireless Networks", Third International Conference on Wireless and Mobile Communications ICWMC'07, 2007.

[41] R. Ferrus, O. Sallent and R. Agusti, "Interworking in heterogeneous wireless networks: Comprehensive framework and future trends", IEEE Wireless communications, April 2010.

[42] P. Vidales, J. Baliosion, J. Serrat, G. Mapp, F. Stejano, and A. Hopper, " Autonomic sytem for mobility support in 4G networks", IEEE Journal on Selcted Areas in Communications'05, volume 23, Dec 2005.

[43] A. Cuevas, J. I. Moreno, P. Vidales, and H. Einsiedler. The IMS platform: A solution for next generation network operators to be more

than bit pipes. In IEEE Communications Magazine, issue on Advances of Service Platform Technologies, Aug. 2006.

[44]   J.P. Singh, T. Alpcan, X. Zhu, and P. Agrawal, "Towards Heterogeneous Network Convergence: Policies and Middleware Architecture for Efficient Flow Assignment, Rate Allocation and Rate Control for Multimedia Applications", Workshop on Middleware for Next-generation Converged Networks and Applications (MNCNA 2007), ACM/IFIP/USENIX 8th International Middleware Conference, Nov. 2007.

[45]   A. Francini, A. Rajkumar, P. Feder and U.O. Hernandez, "21-06-0620-01-0000-New_section_5_5, IEEE Std. 802.21", 2006.

[46]   M. Kuroda, Y. Saito, K. Ishizu, and R. Komiya, "Clarification of MIH_NMS_SAP, DCN: 21-06-0786-00-0000" IEEE Std. 802.21, 2006.

[47]   L. Wu and K. Sandrasegaran, A Study on Radio Access Technology, Selection Algorithms, SpringerBriefs in Electrical and Computer Engineering, DOI: 10.1007/978-3-642-29399-3_2, 2012.

[48]   A.Tolli and Petteri Hakalin, "Adaptive load balancing between multiple cell layers", Proc on Vehicular Technology Conference, Vol.3, Fall 2002, pp. 1691 - 1695.

[49]   N. Passas, S. Paskalis, A. Kaloxylos, F. Bader, R. Narcisi, E. Tsontsis, A.S. Jahan, H. Aghvami, M. O'Droma, I. Ganchev, Enabling technologies for the always best connected concept.Wireless Commun Mob Comput 6, 523–540 (2006).

[50]   J. Pérez-Romero, O. Sallent, R. Agustí On Evaluating beyond 3G radio access networks:architectures, approaches and tools, in IEEE 61stVehicular TechnologyConference, Stockholm,Sweden, pp. 2964–2968, May-June 2005.

[51]   3GPP TR v5.0.0, Improvement of RRM across RNS and RNS/BSS (Release 5), 2001.

[52]   F. Casadevall, P. Karlsson, O. Sallent, H. Gonzalez, A. Barbaresi, M. Dohler, Overview of the EVEREST Project, 13th IST Mobile & Wireless Communications Summit (Lion, France, 2004).

[53]   Y. Cui, Y. Xue, H. Shang, X. Sha, Z. Ding, A novel scheme and access architecture for joint radio resource management in heterogeneous networks, in International Forum on Information Technology and Applications, 2009 (IFITA'09), Chengdu, China, pp. 24–27, May 2009.

[54]   P. Magnusson, J. Lundsjö, J. Sachs, P. Wallentin, Radio resource management distribution in a beyond 3G multi-radio access architecture, in IEEE Global Telecommunication Conference, pp. 3472–3477, 2004.

[55]   O.E.Falowo and H.A.Chana, "Joint call admission control algorithms: Requirements, approaches, and design considerations", Computer Communications, vol. 31, no. 6, April 2008, pp. 1200-1217.

[56]   A.AL-Sabbagh, R.Braun, M.Abolhasan, "A Comprehensive Survey on RAT Selection Algorithms for Heterogeneous Networks", World Academy of Science, Engineering & Technology, 2011.

[57]   K.H.Suleiman, H.A.Chan, M.E.Dlodlo, "Load balancing in the call admission control of heterogeneous wireless networks", International Conference on communications and mobile computing (IWCMC'06), Vancouver, Canada, July 3–6, 2006, pp. 245–250.

[58]   A.Pillekeit, F.Derakhshan, E.Jugl, A.Mitschele-Thiel, "Force-based load balancing in co-located UMTS/GSM networks", VTC 2004-Fall, IEEE 60th Vol. 6, September 26–29, 2004, pp. 4402–4406.

[59]   W.Zhang, "Performance of Real-time and Data Traffic in Heterogeneous Overlay Wireless Networks", 19th International Teletraffic Congress (ITC 19), Beijing, China, 2005.

[60]   W.Song, H.Jiang, W.Zhuang, X.Shen, "Resource management for QoS support in WLAN/cellular interworking", IEEE Network Magazine, 2005.

[61]   O.E.Falowo, H.A.Chan, "Joint Call Admission Control for Next Generation Wireless Networks", Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering, Ottawa, Canada, May 7–10, 2006.

[62]   A.Umbert, L.Budzisz, N.Vucevic and F.Bernardo, "An all-IP heterogeneous wireless testbed for RAT selection and e2e QoS evaluation", The 2007 International Conference on Next Generation Mobile Applications, Services and Technologies (NGMAST 2007), Cardiff, Wales, UK, September 12-14, 2007.

[63]   J.Perez-Romero, O.Sallent, R.Agusti, L Wang, H.Aghavmi, "Network-controlled cell-breathing for capacity improvement in heterogeneous CDMA/TDMA scenarios", IEEE Wireless Communications and

Networking Conference (WCNC'06), April 3–6, 2006, Las Vegas, NV, USA.

[64] R.B.Ali, S.Pierre, "An efficient predictive admission control policy for heterogeneous wireless bandwidth allocation in next generation mobile networks", International Conference on Communications and Mobile Computing (IWCMC'06), Vancouver, Canada, July 3–6, 2006.

[65] O.Ormond, J.Murphy, and G-M.Muntean, "Utility-Based Intelligent Network Selection in Beyond 3G Systems", 2006 IEEE International Conference on Communications (ICC 2006), Istanbul, Turkey, June 11– 15, 2006, pp. 1831–1836.

[66] J.Perez-Romero, O.Sallent, and R.Agusti, "A Novel Metric for Context-Aware RAT Selection in Wireless Multi-Access Systems", 2007 IEEE International Conference on Communications, Glasgow, Scotland, June 24–28, 2007, pp. 5622-5627.

[67] J.Perez-Romero, O.Salient, and R.Agusti, "A Generalized Framework for Multi-RAT Scenarios Characterisation", The 65th IEEE Vehicular Technology Conference (VTC2007-Spring), Dublin, Ireland, April 22- 25, 2007, pp. 980 – 984.

[68] J. Chen, K. Yu, Y. Ji, P. Zhang, Non-cooperative distributed network resource allocation in heterogeneous wireless data networks, in: Proceedings of IST Mobile Summit 06, Mykonos, Greece, June 5–7, 2006.

[69] A.L.Wilson, A.Lenaghan, R.Malyan, "Optimising Wireless Access Network Selection to Maintain QoS in Heterogeneous Wireless Environments", International Symposium on Wireless Personal

Multimedia Communications (WPMC 2005), Aalborg Denmark, September 18-22, 2005, pp. 1236-1240.

[70]   P.M.L.Chan, R.E.Sheriff, Y.F.Hu, P.Conforto, and C.Tocci, "Mobility Management Incorporating Fuzzy Logic for a Heterogeneous IP Environment", IEEE Communications Magazine, vol. 39, no.12, December 2001, pp. 42-51.

[71]   Q.Guo, X.Xu, J.Zhu, H.Zhang, "A QoS-guaranteed cell selection strategy for heterogeneous cellular systems", ETRI Journal 28 (1), 2006.

[72]   M.Ali, P.Pillai, Y.F.Hu and K.Xu, "Load aware radio access selection in heterogeneous terrestrial wireless networks", International Journal of Computer Networks & Communications (IJCNC) Vol.3, No.6, November 2011.

[73]   M.Ali, P.Pillai and Y.F.Hu, "Load-aware radio access selection in future generation wireless networks", 4th International ICST Conference on Personal Satellite Services, PSATS 2012.

[74]   S. Buljore, H. Harada, S. Filin, P. Houze, K. Tsagkaris, O. Holland, "Architecture and enablers for optimized radio resource usage in heterogeneous wireless access networks", The IEEE 1900.4 Working Group Communications Magazine, vol.47, no.1, pp.122–129, Jan. 2009.

[75]   B. Li, W. X. Shi, N. Li, "A Hierarchical Semi-centralized Architecture for Load Balancing of Heterogeneous Wireless etworks", 2010 Second International Conference on Networks Security Wireless

Communications and Trusted Computing NSWCTC), vol.2, pp.28–31, Apr. 2010.

[76]   Y. R. Lan, T. Yu, "A dynamic central scheduler load balancing mechanism", 1995 IEEE Fourteenth Annual International Phoenix Conference on Computers and Communications, pp.734–740, Mar. 1995.

[77]   G. Q. Ning, G. X. Zhu, "Load balancing based on traffic selection in heterogeneous overlapping cellular networks (in Chinese)", Journal of Chinese Computer Systems, vol.27, no.11, pp.2036–2041, Nov. 2006.

[78]   G. Q. Ning, G. X. Zhu, L. X. Peng, X. F. Lu, "Load balancing based on traffic selection in heterogeneous overlapping cellular networks", The First IEEE and IFIP International Conference in Central Asia on Internet, Sept. 2005.

[79]   G. Rumi, L. Kristina, "Structure of heterogeneous networks", 12th IEEE International Conference on Computational Science and Engineering, CSE 2009, vol.4, pp.98–105, Aug. 2009.

[80]   B. Venkata Ramana, Devesh Agrawal, C. Siva Ram Murthy, "Design and performance evaluation of meghadoot – A hybrid wireless network architecture", 2006 IEEE International Conference on Networks, vol.2, pp.605–610, Sept. 2006.

[81]   P. Arun, V. Rajesh, T. Rajeev, N. Kshirasagar, "Multiple mobile routers based seamless handover scheme for next generation heterogeneous networks", First International Conference on

Networks and Communications, NETCOM 2009, pp.72–77, Dec. 2009.

[82] M. X. Li, D. L. Xie, B. Hu, Y. Shi, S. Z. Chen, "A multihop routing mechanism based on fuzzy estimation for heterogeneous wireless networks," IEEE Vehicular Technology Conference, pp.1–5, Sept. 2009.

[83] K. Mokhtarian, M. Hefeeda, "Authentication of Scalable Video Streams with Low Communication Overhead," IEEE Transactions on Multimedia, 2011.

[84] K. Sharif, L. J. Cao, Y. Wang, T. Dahlberg, "A Hybrid Anycast Routing Protocol for Load Balancing in Heterogeneous Access Networks," International Conference on Computer Communications and Networks, ICCCN 2008, pp.99–104, Aug. 2008.

[85] A. Zouari, L. Suciu, J. M. Bonnin, K. Guillouard, "A Proactive and Distributed QoS Negotiation Approach for Heterogeneous Environments: An Evaluation of QoS Signalling Overhead," The 2nd International Conference on Next Generation Mobile Applications, Services, and Technologies, NGMAST 2008, pp.41–46, Sept. 2008.

[86] W. Shi, B. Li, N. Li, C. Xia, "A Network architecture for load balancing of heterogeneous wireless networks", Journal of networks, VOL. 6, NO. 4, April 2011.

[87] W.Song, W.Zhuang, and Y.Cheng, "Load balancing for cellular/WLAN integrated networks", IEEE network, vol. 21, issue 1, 2007.

[88] E.Hossain, "Heterogeneous wireless access networks: Architectures and protocols", Springer Science & Business media, LLC, 2008, ISBN: 978-0-387-09775-3, page 204.

[89] H.Jeounglak, K.JiYeon, K.Jin-Up and K.Sang-Ha, "Dynamic load balancing architecture in heterogeneous wireless network environment", 9th International Symposium on Communications and Information Technology, ISCIT 2009.

[90] L.Giupponi, Ramon Agusti, Jordi Perez-Romero and Oriol Sallent, "A Novel Approach for Joint Radio Resource Management Based on Fuzzy Neural Methodology", IEEE Tr.on Vehicular Technology, Vol. 57, No. 3, 2008, pp. 1789-1805.

[91] A.Pillekeit, Fariborz Derakhshan, Bruno Muller-Clostermann, "Modeling of common radio resource management scenarios", Proc. of spring simulation multiconference, Vol. 1, 2007, pp. 205 -210.

[92] O.Holland, et al., "Development of a Radio Enabler for Reconfiguration Management within the IEEE P1900.4 Working Group", Proc. on Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySpan), April 2007. pp. 233 - 239.

[93] E.F.Olabisi, H.Anthony Chan, "Load-Balancing in Heterogeneous Wireless Networks Implementing a User-Centric Joint Call Admission Control Algorithm," Proceedings of South Africa Telecommunication Networks and Applications Conference (SATNAC), Wild Coast Sun, South Africa, 6-10 September 2008.

[94]   Shiann-Tsong Sheu and Chih-Chiang Wu, "Dynamic Load Balance Algorithm (DLBA) for IEEE 802.11 Wireless LAN", Tamkang Journal of Science and Engineering, vol. 2, No. 1 pp. 45-52 (1999).

[95]   Ioannis Papanikos and Michael Logothetis, "A Study on Dynamic Load Balance for IEEE 802.11b Wireless LAN", 8th International Conference on Advances in Communications and Control, COMCON'01, June 2001.

[96]   Ted Scully, Kenneth Brown, "Wireless LAN load balancing with genetic algorithms", Knowledge Based Systems, (2009).

[97]   Mun-Suk Kim and SuKyoung Lee, "Load balancing and its performance evaluation for layer 3 and IEEE 802.21 frameworks in PMIPv6-based wireless networks", WIRELESS COMMUNICATIONS AND MOBILE COMPUTING, Wirel. Commun. Mob. Comput. (2009), DOI: 10.1002/wcm.832.

[98]   SuKyoung Lee, Kotikalapudi Sriram, Kyungsoo Kim, Yoon Hyuk Kim, and Nada Golmie, "Vertical Handoff Decision Algorithms for Providing Optimized Performance in Heterogeneous Wireless Networks", IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, JANUARY 2009.

[99]   Alexandros Kaloxylos, Ioannis Modeas, Fotos Georgiadis and Nikos Passas, "Network Selection Algorithm for Heterogeneous Wireless Networks: from Design to Implementation", Network Protocols and Algorithms, ISSN 1943-3581, 2009, Vol. 1, No. 2.

[100] G. Cybenko, "Dynamic load-balancing for distributed memory multi-computers", Journal of Parallel and Distributed Computing, October 1989.

[101] M. Willekk-Lemair and A.P. Reeves, "Strategies for dynamic load-balancing on highly parallel computers", IEEE Transaction on Parallel and Distributed Systems, September 1993, Pages 979-993.

[102] M. Wu and W. Sbu, "A load balancing algorithm for n-cube", Proceedings of the 1996 International Conference on Parallel Processing, IEEE Computer Society, 1996, Pages 148-155.

[103] Siu-Cheung Chuu, Ada Wui-Chee Fu, "Load Balancing between Computing Clusters", Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies, 2003. PDCAT 2003.

[104] Constantine K. Christakos, Antonio Izquierdo, Richard Rouil, Nada Golmie, "Using the Media Independent Information Service to Support Mobile Authentication in Fast Mobile IPv6", IEEE wireless communications and networking conference, April 2009.

[105] Pedro Neves, Jollo Soares, Susana Sargento, "Dynamic Media Independent Information Server", IEEE Symposium on Computers and Communications (ISCC), 2010.

[106] Pedro Neves, João Soares, Susana Sargento, Hugo Pires, Francisco Fontes, "Context-aware media independent information server for optimized seamless handover procedures", International Journal of Computer and Telecommunications Networking, 2011.

[107] Chi Ma, Enda Fallon, Yuansong Qiao, Brian Lee, "Operating Network Load Balancing with the Media Independent Information Service for Vehicular Based Systems", International Journal of Simulation Systems, Science & Technology, IJSSST, Vol. 11, No. 6, 2011.

[108] N.Efthymiou, Y.F.Hu, R.E.Sheriff, "Performance of Intersegment Handover Protocols in an Integrated Space/Terrestrial-UMTS Environment", IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, VOL. 47, NO. 4, NOVEMBER 1998.

[109] A.H.Aghvami, P. Smyth, "Forward or backward handover for W-CDMA", 3G Mobile Communication Technologies, First International Conference, 2002.

[110] C.Chrysostomou, A. Pitsillides, F. Pavlidou, "A Survey of Wireless ATM Handover Issues", International Symposium of 3G Infrastructure and Services, 2001.

[111] M.Ali, P.Pillai and Y.F.Hu, "Load-aware radio access selection in future generation Satellite-terrestrial wireless networks", International Journal of Wireless & Mobile Networks (IJWMN) 2012.

[112] Jairo J. Espinosa, Joos P.L. Vandewalle, Vincent Wertz, "Fuzzy Logic, Identification and Predictive Control", Springer verlag London limited, ISBN: 1-85233-828-8, 2005.

[113] Valluru B. Rao, "C++ Neural Networks and Fuzzy Logic", IDG Books Worldwide, Inc., ISBN: 1558515526, Jan 1995.

[114] J.W.Hines, "Fuzzy and Neural Approaches in Engineering MATLAB Supplement", John Wiley and Sons, New York, NY, Copyright 1997.

[115] Thomas A. Runkler, "Selection of Appropriate Defuzzification Methods Using Application Specific Properties", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 5, NO. 1, FEBRUARY 1997.

[116] N. Mogharreban, L. F. DiLalla, "Comparison of Defuzzification Techniques for Analysis of Non-interval Data", North American Fuzzy Information Processing Society, NAFIPS06, June 2006.

[117] M. Brown, K.M. Boosley, D.J. Mills, C.J. Harris, "High dimensional neurofuzzy system: Overcoming the curse of dimensionality", IEEE International Conference on Fuzzy Systems, 1995.

[118] Robert Fuller, "Introduction to Neuro-Fuzzy Systems", Physica-verlag Heidelberg, New York, ISBN: 3-7908-1256-0, 2000.

[119] Network simulator user manual: http://www.isi.edu/nsnam/ns/doc/index.html, visited May 30, 2012.

[120] T. Issariyakul, E. Hossain, "Introduction to Network Simulator NS2", ISBN: 978-0-387-71759-3, Springer science & business media, LLC, 2009.

[121] M. Mitchell, J. Oldham, and A. Samuel, "Advanced Linux Programming", Published by: New Riders Publishing, ISBN 0-7357-1043-0, First Edition, June 2001.

[122] D. Hanselman, B. Littlefield, "Mastering MATLAB", Published by: Prentice Hall, ISBN 13: 978-0-13-601330-3, 1997.

[123] P.Raines, J. Tranter, "TCL/TL in a nut shell", Published by: O'Reilly& Associates, ISBN 1-56592-433-9, March 1999.

[124]   N. Golmie, "Seamless and Secure Mobility", Department of Defense Annual Information Assurance Workshop, Philadelphia, February 7-10, 2005.

[125]   A. Izquierdo, N. Golmie, R. Rouil, "Optimizing Authentication in Media Independent Handovers using IEEE 802.21", IEEE ICC 2009, Dresden, August 06, 2009.

[126]   IEEE communications society, "IEEE Standard for Architectural Building BlocksEnabling Network-Device Distributed Decision Making for Optimized Radio Resource Usage in Heterogeneous Wireless Access Networks", IEEE Std 1900.4a™2011.